# LONG-TERM COMPOSED VIDEO RETRIEVAL: BENCHMARKING AND ADAPTIVE HISTORY REFOCUSING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

The rapid growth of video data has made Composed Video Retrieval (CVR) an emerging video retrieval paradigm. Existing CVR methods have achieved strong performance on short-term, i.e., single-turn, benchmarks, but they remain limited in real-world long-term retrieval. In practice, users often refine their requests iteratively based on intermediate retrieval results, rather than expressing the full intent in a single modification text. To systematically evaluate this challenge, we introduce LTCVR, a novel benchmark for Long-Term Composed Video Retrieval, featuring five representative history-dependency patterns: *progressive refinement*, *rollback*, *cross-turn combination*, *local correction*, and *preference change*. Our analysis on LTCVR shows that existing methods degrade markedly when the retrieval intent depends on earlier interaction context, and attention analysis further reveals that representative methods fail to focus on the key historical turns. To address this limitation, we propose InAHR, an Intent-Driven Adaptive History Refocusing framework. InAHR first infers the current retrieval intent and then selects a compact subset of historical visual evidence that is intent-relevant and representative. The selected evidence is re-injected into the generation process to produce retrieval-oriented target descriptions for video retrieval. Extensive experiments demonstrate that InAHR consistently outperforms strong baselines across diverse long-horizon CVR scenarios.

## 1 INTRODUCTION

Video content has become a dominant information carrier in real-world applications, from media search to content recommendation. In many scenarios, users often seek videos that are similar to a given example but differ in specific attributes or objects. Such demands are expressed through compositional queries, motivating composed video retrieval (CVR). CVR aims to retrieve videos that are visually related to a reference video while satisfying a textual modification instruction. Recent advances in vision-language models (VLMs), e.g., CLIP (Radford et al., 2021) and BLIP (Li et al., 2022), as well as multimodal large language models (MLLMs), including Qwen2.5-VL (Bai et al., 2025b) and GPT-4o (OpenAI, 2024), have substantially improved CVR (Ventura et al., 2024a; Hummel et al., 2024; Thawakar et al., 2024), making it an important testbed for compositional video understanding.

Despite this progress, most existing CVR benchmarks rely on a restrictive assumption: the user's retrieval intent can be fully specified by a single modification instruction. In real-world applications, however, this assumption rarely holds, as users typically refine their requests iteratively based on intermediate retrieval results. This limitation is particularly pronounced in video retrieval, where rich visual content and temporal dynamics make it difficult for a single instruction to fully specify what should be preserved, modified, or emphasized. For example, a user may begin with an underspecified request and later clarify the desired action or scene after inspecting retrieved results; alternatively, the final intent may depend on jointly satisfying constraints introduced across multiple turns (Figure 1). Such scenarios require models to reason over long-term interaction history and track how the user's retrieval intent evolves across turns, rather than treating each modification independently.

Recent work has begun exploring multi-turn composed retrieval. Early benchmarks such as MT-FashionIQ (Yuan & Lam, 2021) and MT-Shoes (Pal et al., 2023) primarily study short, sequential refinements in fashion image retrieval. More recently, FashionMT (Chen et al., 2025) introduces



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.

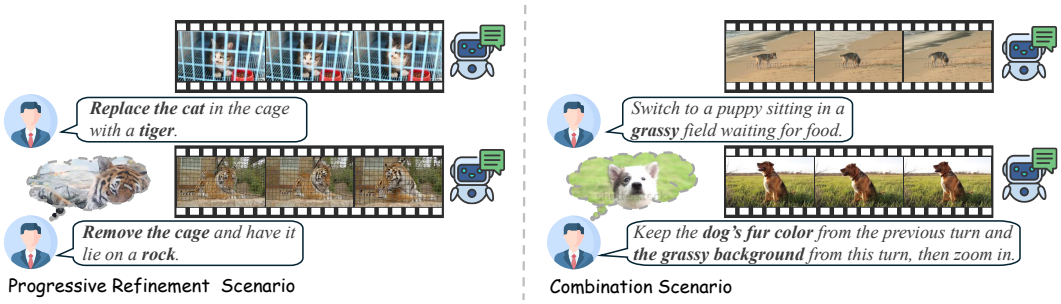



Figure 1: Examples of different history-dependency patterns in real-world applications.

retrospective interactions, while CoCo-IR (Cao et al., 2026) considers contextual relations across multi-turn image-text histories. However, these benchmarks remain image-centric and cover limited history dependencies, making them insufficient for long-term video retrieval with dynamic visual content. Therefore, we need a benchmark that jointly captures long-term interactions, diverse history dependencies, and the temporal dynamics of video content.

To fill this gap, we introduce **LTCVR**, the first benchmark designed for long-term composed video retrieval with diverse history dependencies. LTCVR includes five representative history-dependency patterns that capture how retrieval intent evolves across interactions: (i) *progressive refinement*, where the user iteratively adds constraints based on the latest retrieved result; (ii) *rollback*, where the user returns to an earlier result and continues the search from that state; (iii) *cross-turn combination*, where attributes from different turns must be jointly satisfied; (iv) *local correction*, where a specific aspect of the latest result is revised; and (v) *preference change*, where the target shifts and obsolete history should be ignored. Together, these scenarios enable systematic evaluation of how retrieval models accumulate, revisit, combine, revise, or discard historical information as user intent evolves.

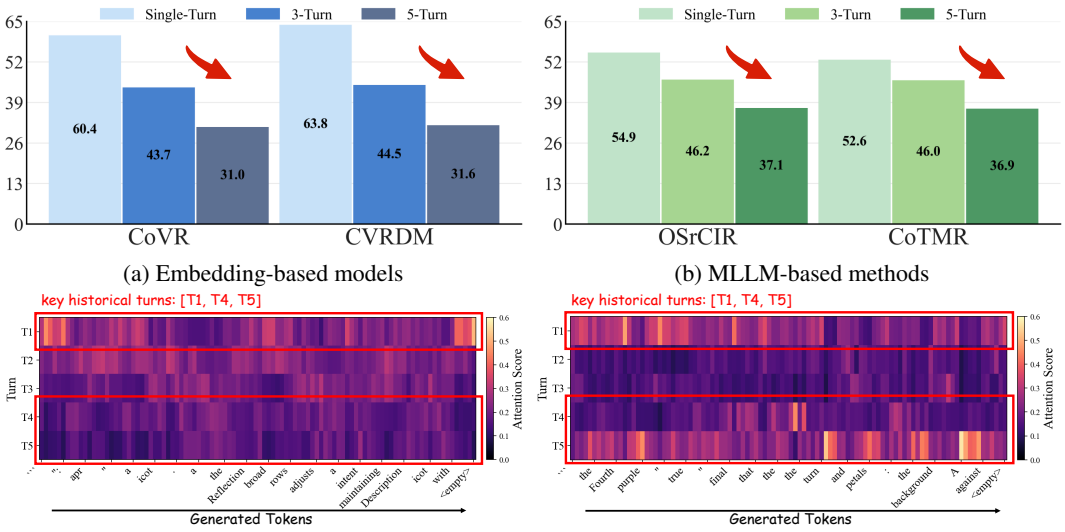
Evaluation on LTCVR exposes a shared weakness of existing methods: retrieval performance degrades as interaction history grows. Embedding-based models, such as CoVR (Ventura et al., 2024a) and CVRDM (Thawakar et al., 2025), degrade substantially as more interaction turns are introduced (Figure 2(a)). MLLM-based methods, such as OSrCIR (Tang et al., 2025) and CoTMR (Sun et al., 2025), benefit from stronger long-context reasoning, yet their performance still declines as interaction histories grow longer and history dependencies become more complex (Figure 2(b)). Attention analysis further shows that these models often distribute attention broadly across the interaction history rather than focusing on the turns most relevant to the current intent (Figure 2(c)). These results suggest that longer context alone is insufficient; effective long-term CVR requires selectively identifying and leveraging intent-relevant history.

Inspired by human cognitive reasoning, where people first form a coarse understanding and then selectively revisit relevant prior context (Sabo et al., 2024; Moccia et al., 2025), we propose **InAHR**, an Intent-Driven Adaptive History Refocusing framework for long-term CVR. Specifically, given a sparse interaction overview, InAHR first infers the current retrieval intent and identifies the relevant historical turns. It then adaptively refocuses on historical evidence from the turn level to the frame level, using the predicted turns as a soft prior and further balancing intent relevance with representative coverage. Finally, the refocused evidence is combined with the inferred intent to reconstruct the target from multiple complementary semantic facets, producing retrieval-oriented descriptions that preserve fine-grained historical cues. Through this coarse-to-fine process, InAHR selectively exploits useful history while reducing the influence of redundant or obsolete context (Figure 2(d)).

In summary, our contributions are threefold:

• We introduce **LTCVR**, a long-term composed video retrieval benchmark covering five representative history-dependency patterns, designed to systematically evaluate how models accumulate, revisit, combine, revise, or discard historical information as user intent evolves.

• To address the challenge of effectively utilizing historical evidence revealed by LTCVR, we introduce **InAHR**, an Intent-Driven Adaptive History Refocusing framework. Inspired by human cognition, InAHR first forms a global understanding of the evolving user intent and then adaptively refocuses on intent-relevant historical evidence to generate retrieval-oriented descriptions.




(a) Embedding-based models

(b) MLLM-based methods

(c) Attention Visualization(OSrCIR)

(d) Attention Visualization(Ours)

Figure 2: Motivation for long-term history-aware CVR. (a)–(b) Existing methods degrade as the number of turns increases. (c)–(d) Compared with OSrCIR's diffuse attention over history, InAHR better concentrates on the key turns needed to infer the final retrieval intent.

• Extensive experiments show that InAHR consistently improves performance across diverse long-term CVR scenarios, particularly under complex history dependencies, demonstrating improved robustness to obsolete context while preserving target-relevant evidence. The dataset and code are available at: https://anonymous.4open.science/r/InAHR-1FC1.

## 2 Related Work

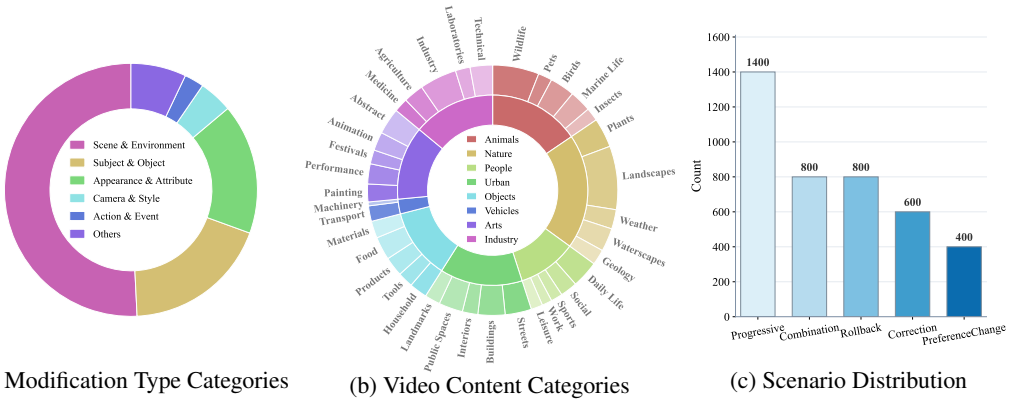
**Composed Visual Retrieval.** Composed Image Retrieval (CIR) aims to retrieve a target image given a multimodal query comprising a reference image and a textual modification instruction (Vo et al., 2019; Saito et al., 2023; Gu et al., 2024). Existing CIR methods have developed increasingly effective composition mechanisms, including mapping images into textual query tokens (Saito et al., 2023), zero-shot retrieval with large pretrained VLMs (Gu et al., 2024; Psomas et al., 2025), and MLLM-based approaches that reason over the reference image and modification text to generate retrieval-oriented target descriptions (Sun et al., 2025; Tang et al., 2025). Building on this formulation, recent works extend composed retrieval from images to videos and introduce composed video retrieval (CVR) (Ventura et al., 2024a; Hummel et al., 2024). Compared with CIR, CVR additionally requires modeling temporal dynamics and evolving actions in video content. Specifically, CoVR (Ventura et al., 2024a) learns multimodal alignment and fusion between the reference video and the modification text for retrieval. EgoCVR (Hummel et al., 2024) extends composed retrieval to egocentric videos and improves performance with a training-free re-ranking strategy. Subsequent methods further strengthen retrieval by enriching contextual information (Thawakar et al., 2024), modeling denser modifications (Thawakar et al., 2025), or improving fine-grained video-text alignment through textual token disentanglement (Wu et al., 2025; Gupta et al., 2025).

Despite these advances, most existing CIR and CVR methods remain limited to single-turn retrieval, making it difficult to capture evolving user intent and dependencies on prior interactions. This limitation has motivated recent studies on multi-turn composed retrieval.

**Multi-turn and Long-term Retrieval.** Recent studies have extended composed retrieval to multi-turn settings. Early works such as MT-FashionIQ (Yuan & Lam, 2021) and FashionNTM (Pal et al., 2023) mainly focus on successive refinements in fashion image retrieval. FashionMT (Chen et al., 2025) further introduces retrospective dependencies across turns, while CoCo-IR (Cao et al., 2026) considers richer contextual dependencies over multi-turn image-text histories. However, existing benchmarks remain image-centric and cover limited forms of long-term history dependency, leaving long-term CVR insufficiently explored. As shown in Table 1, LTCVR fills this gap by jointly

Table 1: Comparison of representative multimodal retrieval benchmarks. **Multi-scenario:** whether the benchmark covers diverse patterns of history dependency. **Intent Eval.:** whether the benchmark explicitly assesses intent understanding beyond standard retrieval metrics. **Avg. Turns:** average number of turns per session. **Avg. W/S:** average number of words per session.

| Benchmark | Task | Data Source | Multi-turn | History-aware | Multi-scenario | Intent Eval. | Avg. Turns | Avg. W/S | Annotation |
|---|---|---|---|---|---|---|---|---|---|
| MSR-VTT (Xu et al., 2016) | TVR | Web Videos | ✗ | ✗ | ✗ | ✗ | - | 9.0 | Manual |
| WebVid-CoVR (Ventura et al., 2024a) | CVR | WebVid-2M | ✗ | ✗ | ✗ | ✗ | - | 4.8 | Auto&Manual |
| EgoCVR (Hummel et al., 2024) | CVR | Ego4D | ✗ | ✗ | ✗ | ✗ | - | 4.0 | Manual |
| Dense-WebVid-CoVR (Thawakar et al., 2025) | CVR | WebVid | ✗ | ✗ | ✗ | ✗ | - | 31.16 | Auto |
| MT Shoes (Pal et al., 2023) | CIR | Shoes | ✓ | ✗ | ✗ | ✗ | 3.3 | 17.2 | Auto |
| IVCR-200K (Han et al., 2025) | TVR | Video Datasets | ✓ | ✓ | ✗ | ✗ | 1.1 | 27.0 | Auto&Manual |
| FashionMT (Chen et al., 2025) | CIR | Fashion Images | ✓ | ✓ | ✗ | ✗ | 3 | 72.9 | Auto |
| **LTCVR (Ours)** | **CVR** | **WebVid-2M** | ✓ | ✓ | ✓ | ✓ | **5** | **101.3** | **Auto&Manual** |




(a) Modification Type Categories (b) Video Content Categories (c) Scenario Distribution

Figure 3: Data diversity of LTCVR. (a) shows the distribution of modification types. (b) illustrates the hierarchical taxonomy of video content, spanning eight major categories and their fine-grained subcategories. (c) shows the distribution of history-dependency scenarios.

supporting long-term CVR, diverse history-dependency scenarios, and explicit intent-level evaluation. Additional discussion is provided in Appendix A.

## 3 LTCVR: Long-term Composed Video Retrieval Benchmark

Existing multi-turn composed retrieval benchmarks are primarily designed for images and cover limited forms of history dependency, making them insufficient for systematically studying long-term CVR. To address this gap, we introduce LTCVR, a benchmark designed to evaluate how retrieval models reason over evolving user intent and interaction history in video retrieval. LTCVR contains 4,000 five-turn sessions covering five representative history-dependency patterns: *Progressive Refinement*, *Rollback*, *Cross-turn Combination*, *Local Correction*, and *Preference Change*. As shown in Figure 3, LTCVR further spans diverse modification types and video content, providing a controlled and diverse testbed for long-term history-aware CVR.

**Scenario-Controlled Benchmark Construction.** We construct LTCVR from WebVid-2M (Bain et al., 2021) through a scenario-controlled pipeline. We first group videos into semantic neighborhoods and retain moderately similar reference–target pairs, filtering out near-duplicates and overly distant pairs. For each retained pair, we keep the target fixed and generate a coherent five-turn progressive trajectory. At each turn, an MLLM produces user-style feedback describing visible changes toward the target, and the retrieved video becomes the reference for the next turn. Starting from these progressive trajectories, we construct four additional scenarios by modifying how the final intent depends on interaction history. For *Rollback*, the final request returns to an earlier reference; for *Cross-turn Combination*, it combines constraints from the latest reference and an earlier interaction turn; for *Local Correction*, it revises a specific mismatch in the latest result while preserving the overall intent; and for *Preference Change*, the user shifts to a new target, making earlier constraints partly or entirely obsolete. Together, these scenarios enable systematic evaluation of how retrieval models handle diverse history dependencies as user intent evolves.

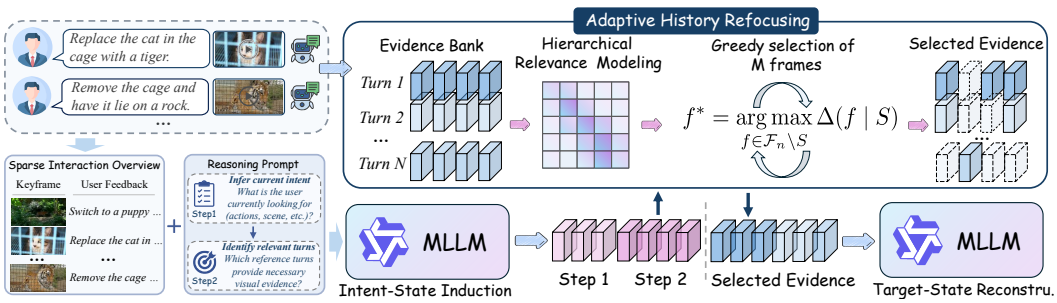



Figure 4: Overview of the InAHR framework for intent-driven adaptive history refocusing.

**Quality Control and Evaluation Protocol.** To ensure the quality of LTCVR, we employ a multi-stage filtering and verification process. We first remove trajectories containing invalid video transitions or redundant feedback. The remaining sessions are screened by an MLLM for visual grounding, instruction clarity, history dependency, and scenario consistency. We then perform approximately balanced sampling over video content and modification types. Finally, three independent human reviewers verify the accuracy, history dependency, and scenario consistency of each session. Samples that fail the review criteria are revised or replaced, resulting in the final 4,000 high-quality sessions. For evaluation, we use Recall@$K$ as the primary retrieval metric. However, retrieval accuracy alone does not reveal whether a model correctly understands the evolving intent or uses the appropriate interaction history. We therefore introduce an Intent Understanding Evaluation, which evaluates model outputs at the description, reasoning, and retrieval levels. At each level, we assess three complementary dimensions: *Intent Consistency*, measuring alignment with the current retrieval intent; *History Grounding*, measuring whether relevant historical information is correctly used while obsolete context is ignored; and *Scenario Faithfulness*, measuring whether the intended history-dependency pattern is correctly followed. Additional statistics, evaluation details, full construction details, prompts, and review criteria are provided in Appendix B.

## 4 InAHR: The Intent-Driven Adaptive History Refocusing Framework

Given the latest feedback $u_n$ and the interaction context, consisting of the latest reference video $v_n$ and preceding interaction history $\mathcal{H}_{<n} = \{(v_i, u_i)\}_{i=1}^{n-1}$, long-term composed video retrieval aims to retrieve the target video $v^* \in \mathcal{V}$ that best matches the current user intent, where $\mathcal{V}$ denotes the candidate video set. As shown in Figure 4, we propose **InAHR**, which reorganizes long-term historical information in a coarse-to-fine manner. Specifically, Intent-State Induction infers the current intent and relevant turns from a sparse interaction overview. Adaptive History Refocusing uses these turns as a soft prior to select relevant and non-redundant frame-level evidence. Finally, Target-State Reconstruction generates multiple complementary descriptions of the same target for retrieval.

**Intent-State Induction** Inspired by human cognitive reasoning, where people first form a coarse understanding before selectively revisiting relevant prior context (Sabo et al., 2024; Moccia et al., 2025), we first obtain a compact view of the evolving retrieval intent before accessing fine-grained historical evidence. Specifically, for each interaction turn $i$, we pair the middle frame $\bar{f}_i$ of the reference video $v_i$ with the corresponding user feedback $u_i$ to construct a sparse interaction overview: $\mathcal{O}_{\leq n} = \{(\bar{f}_i, u_i)\}_{i=1}^{n}$. This lightweight context provides a global view of the interaction trajectory while avoiding the redundancy of dense historical video inputs. Given $\mathcal{O}_{\leq n}$, the MLLM jointly infers a description of the current retrieval intent $c_n$ and the minimal set of turns $\mathcal{I}_n$ needed to determine it: $(c_n, \mathcal{I}_n) = MLLM_\theta(\mathcal{O}_{\leq n})$, $\mathcal{I}_n \subseteq \{1, \ldots, n\}$. Here, $c_n$ consolidates the visual and textual constraints that remain relevant while excluding superseded information, whereas $\mathcal{I}_n$ identifies the smallest sufficient set of turns required to resolve the current target. These outputs provide the semantic guidance and coarse history prior for the subsequent stages.

**Adaptive History Refocusing** Given the current intent $c_n$ and the relevant-turn set $\mathcal{I}_n$, we next refocus on fine-grained visual evidence from the interaction history. To measure visual evidence against the induced intent in a unified space, we encode $c_n$ with a frozen CLIP text encoder $\phi_{\text{text}}$ and

uniformly sample $U$ frames from each reference video $v_i$, which are encoded by the corresponding CLIP image encoder $\phi_{\text{img}}$: $h_n = \text{Norm}_2(\phi_{\text{text}}(c_n))$, $z_{i,k} = \text{Norm}_2(\phi_{\text{img}}(f_{i,k}))$, where $i = 1, \ldots, n$ and $k = 1, \ldots, U$.

We select visual evidence from three complementary aspects: *turn-level relevance*, which uses the predicted relevant-turn set $\mathcal{I}_n$ as a soft prior and further calibrates turn importance with visual-semantic consistency; *frame-level relevance*, which favors frames semantically aligned with the current intent; and *representativeness*, which encourages complementary evidence while suppressing redundancy. For *turn-level relevance*, we first convert $\mathcal{I}_n$ into a normalized soft prior: $\pi_i = \frac{1}{2}\left(\frac{1}{n} + \frac{\mathbf{1}[i \in \mathcal{I}_n]}{|\mathcal{I}_n|}\right)$, when $\mathcal{I}_n \neq \emptyset$, and set $\pi_i = 1/n$ otherwise. We then aggregate the sampled frames of each turn into a turn representation $\bar{z}_i = \text{Norm}_2\left(\frac{1}{U}\sum_{k=1}^{U} z_{i,k}\right)$, and calibrate the prior according to its visual-semantic consistency with the current intent:

$$r_i = \text{Softmax}_i\left(\log \pi_i + \frac{\cos(\bar{z}_i, h_n)}{\tau}\right), \tag{1}$$

where $\tau$ controls the sharpness of the relevance distribution. In this way, $\mathcal{I}_n$ provides coarse semantic guidance, while visual consistency further adjusts the importance of each interaction turn. For *frame-level relevance*, we further measure how well each frame within turn $i$ matches the current intent:

$$p_{i,k} = \frac{\exp(\cos(z_{i,k}, h_n)/\tau)}{\sum_{l=1}^{U} \exp(\cos(z_{i,l}, h_n)/\tau)}. \tag{2}$$

We combine the two levels as $\gamma_{i,k} = r_i p_{i,k}$, so that each frame is weighted jointly by the importance of its turn and its own relevance to the current intent.

However, selecting frames solely according to $\gamma_{i,k}$ may concentrate the evidence on a few highly similar frames, often from the same interaction turn, while missing complementary information from other relevant turns. Inspired by facility location-based submodular selection (Lin et al., 2009; Lin & Bilmes, 2011), we further encourage the selected evidence to cover diverse and complementary visual information while avoiding redundancy. Let $\mathcal{F}_n = \{f_{i,k} \mid i = 1, \ldots, n,\ k = 1, \ldots, U\}$ denote the sampled reference frames up to turn $n$, and define their visual similarity as $\kappa(f_{i,k}, f_{j,l}) = \frac{1+\cos(z_{i,k}, z_{j,l})}{2}$. Given an evidence budget $M$, we select

$$S_n^* = \arg\max_{S \subseteq \mathcal{F}_n,\ |S| \leq M} \sum_{i=1}^{n} \sum_{k=1}^{U} \gamma_{i,k} \max_{s \in S} \kappa(f_{i,k}, s). \tag{3}$$

Here, $\gamma_{i,k}$ assigns greater coverage priority to intent-relevant evidence, while the facility-location term encourages representative coverage and naturally suppresses redundant frames. Since the objective is monotone submodular, we use standard greedy maximization to obtain an approximate solution $\widehat{S}_n$, which serves as the final refocused frame set for Target-State Reconstruction.

**Target-State Reconstruction** The current target in long-term interactions may be jointly defined by complementary cues distributed across different interaction turns. Although the preceding stages recover intent-relevant visual evidence, compressing these heterogeneous cues into a single description may obscure fine-grained or less salient constraints. We therefore adopt a multi-faceted target reconstruction strategy. For each selected frame in $\widehat{S}_n$, we retain the feedback associated with its source turn and construct the refocused evidence $\mathcal{E}_n = \{(f_{i,k}, u_i) \mid f_{i,k} \in \widehat{S}_n\}$, ordered by the original turn and frame indices. Given the latest feedback $u_n$, the current intent $c_n$, and $\mathcal{E}_n$, the MLLM generates $J$ retrieval-oriented descriptions of the same target state: $D_n = \{d_n^{(j)}\}_{j=1}^{J} = \text{MLLM}_\theta(u_n, c_n, \mathcal{E}_n)$. where different descriptions emphasize complementary semantics such as subjects, actions, scenes, and local visual details while remaining consistent with the same retrieval intent. Each description is independently encoded by the retrieval text encoder, and the resulting query vectors are averaged to form the final query representation: $q_n = \text{Norm}_2\left(\frac{1}{J}\sum_{j=1}^{J} \phi_{\text{text}}(d_n^{(j)})\right)$. The target video is then retrieved by matching $q_n$ against candidate video representations. This reconstruction allows complementary historical cues to be explicitly captured before their aggregation into the final retrieval query.

Table 2: Comparison results across ALL, *Combination*, *Correction*, *Preference Change*, *Progressive* and *Rollback* subsets. The best and second-best scores are in bold and underlined, respectively.

| Model | ALL | | Combination | | Correction | | Pref. Change | | Progressive | | Rollback | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | R@1 | R@5 | R@1 | R@5 | R@1 | R@5 | R@1 | R@5 | R@1 | R@5 | R@1 | R@5 |
| CLIP | 8.6 | 26.5 | 10.5 | 32.3 | 8.2 | 31.2 | 2.8 | 15.8 | 9.8 | 26.2 | 7.6 | 23.1 |
| BLIP | 15.8 | 40.2 | 18.4 | 44.5 | 17.0 | 43.5 | 6.8 | 22.5 | 17.3 | 41.4 | 14.1 | 40.0 |
| LinCIR | 27.4 | 50.2 | 27.8 | 53.9 | 30.3 | 56.0 | 8.0 | 22.5 | 29.2 | 53.4 | 31.5 | 50.6 |
| CoVR | 31.0 | 58.7 | 30.4 | 57.8 | 35.0 | 62.0 | 12.2 | <u>38.0</u> | 35.8 | 64.2 | 29.7 | 57.8 |
| CoVR-2 | 30.2 | 59.3 | 30.5 | 59.1 | 35.3 | 64.5 | 10.2 | 31.5 | 32.9 | 63.0 | 31.5 | 63.0 |
| CVRDM | 31.6 | 60.4 | 32.5 | 62.1 | 37.3 | 66.5 | 10.7 | 33.3 | 34.3 | 63.4 | 32.1 | 62.5 |
| IP-CIR | 24.6 | 48.9 | 21.0 | 45.9 | 31.0 | 58.2 | 16.3 | 32.3 | 27.1 | 51.9 | 23.0 | 48.1 |
| OSrCIR | 37.1 | 60.6 | 33.1 | 59.8 | <u>44.2</u> | <u>66.7</u> | <u>21.7</u> | 37.5 | 41.1 | 65.1 | 36.7 | 60.4 |
| AutoCIR | 36.9 | 60.9 | 38.2 | 64.5 | 38.8 | 63.2 | 16.0 | 29.7 | 41.3 | 65.4 | 37.0 | 63.2 |
| CoTMR | 36.9 | <u>61.5</u> | 36.7 | 62.7 | 42.8 | 65.8 | 15.8 | 31.8 | 39.6 | 66.3 | <u>38.5</u> | <u>63.4</u> |
| ImageScope | <u>37.8</u> | 59.9 | <u>39.9</u> | <u>65.2</u> | 42.0 | 63.0 | 11.5 | 23.8 | <u>44.0</u> | <u>66.6</u> | 34.6 | 58.5 |
| **Ours** | **41.9** | **66.7** | **40.9** | **65.4** | **46.7** | **72.7** | **27.2** | **51.8** | **45.2** | **67.6** | **41.1** | **69.5** |

# 5 EXPERIMENTS

For evaluation, each LTCVR session contains a fixed five-turn interaction history and a target video corresponding to the final user intent. All methods are given the same interaction history and retrieve this predefined target from a shared gallery of 3,160 unique videos, obtained by deduplicating the targets of all 4,000 sessions. We compare InAHR with two groups of baselines. Embedding-based methods, including CLIP, BLIP, LinCIR (Gu et al., 2024), CoVR (Ventura et al., 2024a), CoVR-2 (Ventura et al., 2024b), and CVRDM (Thawakar et al., 2025), perform retrieval by composing the reference video and modification text into a joint query representation. MLLM-based methods, including IP-CIR (Li et al., 2025), OSrCIR (Tang et al., 2025), CoTMR (Sun et al., 2025), AutoCIR (Cheng et al., 2025), and ImageScope (Luo et al., 2025), use multimodal reasoning to produce retrieval-oriented target representations. Details are in Appendix D.

## 5.1 MAIN RESULTS

Table 2 presents the main comparison results on LTCVR. Overall, embedding-based methods degrade sharply, showing that joint embedding matching alone is insufficient for long-term history-dependent CVR. Existing MLLM-based methods perform better due to stronger contextual reasoning, but still provide limited gains under complex long-term interactions. In contrast, InAHR consistently outperforms all baselines by refocusing on intent-relevant historical evidence, establishing a strong baseline for long-term CVR. Additional generalization results are provided in Appendix F.

We next analyze performance across history-dependency scenarios. Embedding-based methods remain relatively competitive in the simpler *Correction* scenario, but struggle in more history-dependent scenarios such as *Combination*, *Progressive*, *Preference Change*, and *Rollback*. This shows that aggregating turn-level features is insufficient for deciding which historical cues should be preserved, combined, or discarded. MLLM-based methods improve over embedding-based methods, but their gains remain limited in *Preference Change* and *Rollback*, indicating that simply feeding the full history into an MLLM does not reliably resolve evolving user intent.

InAHR achieves clear gains across all scenarios by explicitly refocusing on intent-relevant historical evidence. It performs particularly well in *Preference Change*, where obsolete history should be ignored, and in *Rollback*, where useful evidence must be recovered from earlier turns. These results demonstrate that LTCVR exposes key failure modes of existing methods and that InAHR provides an effective solution for history-aware long-term CVR.

## 5.2 ABLATION STUDY

As shown in Table 4, we first examine the contribution of the three main stages of InAHR. For *w/o Intent-State Induction*, we directly use the latest feedback as the current intent ($c_n = u_n$) and treat all

Table 3: History Sensitivity Analysis on our benchmark. Δ measures the relative change from Last-turn to Full history. Red/green cells denote gains/losses after adding history.

| Model | Setting | ALL | | Combination | | Correction | | Pref. Change | | Progressive | | Rollback | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | R@1 | R@5 | R@1 | R@5 | R@1 | R@5 | R@1 | R@5 | R@1 | R@5 | R@1 | R@5 |
| CoVR | Last-turn | 27.5 | 52.8 | 22.3 | 44.8 | 30.0 | 60.0 | 38.3 | 61.8 | 31.4 | 58.1 | 18.8 | 41.6 |
| | Full history | 31.0 | 58.7 | 30.4 | 57.8 | 35.0 | 62.0 | 12.2 | 38.0 | 35.8 | 64.2 | 29.7 | 57.8 |
| | Δ(%) | +13% | +11% | +36% | +29% | +17% | +3% | -68% | -39% | +14% | +10% | +58% | +39% |
| CoTMR | Last-turn | 30.4 | 54.6 | 19.6 | 40.7 | 35.3 | 60.3 | 39.0 | 62.5 | 32.7 | 58.5 | 29.2 | 53.2 |
| | Full history | 36.9 | 61.5 | 36.7 | 62.7 | 42.8 | 65.8 | 15.8 | 31.8 | 39.6 | 66.3 | 38.5 | 63.4 |
| | Δ(%) | +21% | +13% | +87% | +54% | +21% | +9% | -59% | -49% | +21% | +13% | +32% | +19% |
| ImageScope | Last-turn | 30.1 | 55.6 | 20.4 | 45.5 | 34.8 | 62.3 | 37.5 | 67.7 | 34.6 | 59.0 | 24.7 | 48.7 |
| | Full history | 37.8 | 59.9 | 39.9 | 65.2 | 42.0 | 63.0 | 11.5 | 23.8 | 44.0 | 66.6 | 34.6 | 58.5 |
| | Δ(%) | +26% | +8% | +96% | +43% | +21% | +1% | -69% | -65% | +27% | +13% | +40% | +20% |
| **Ours** | Last-turn | 33.6 | 57.2 | 25.0 | 46.2 | 38.8 | 64.8 | 39.8 | 68.3 | 36.6 | 59.5 | 29.8 | 52.7 |
| | Full history | 41.9 | 66.7 | 40.9 | 65.4 | 46.7 | 72.7 | 27.2 | 51.8 | 45.2 | 67.6 | 41.1 | 69.5 |
| | Δ(%) | +25% | +17% | +64% | +42% | +20% | +12% | -32% | -24% | +23% | +14% | +38% | +32% |

turns as relevant ($I_n = \{1, \ldots, n\}$), which decreases Avg. by 2.1 points. For *w/o History Refocusing*, we bypass the fine-grained refocusing stage and directly use the sparse interaction overview as evidence for Target-State Reconstruction, resulting in the largest drop of 7.0 points. This suggests that coarse interaction summaries alone cannot recover sufficient fine-grained visual evidence from the history. Finally, *Single-Facet Reconstruction* reduces the number of target descriptions from $J = 3$ to $J = 1$, leading to a 0.9-point drop and showing the benefit of complementary target descriptions.

We further analyze the key designs within Adaptive History Refocusing. *Hard Turn Mask* restricts the candidate frame pool to turns predicted as relevant, $F'_n = \{f_{i,k} \mid i \in I_n\}$, instead of using them as a soft prior, reducing Avg. by 1.4 points. For *w/o Frame Relevance*, we set $p_{i,k} = 1/U$, so that frames within the same turn share the same relevance weight; this causes a 2.3-point drop, confirming the importance of fine-grained intent–frame alignment. For *w/o Representative Selection*, we remove the facility-location objective and directly select the top-$M$ frames according to $\gamma_{i,k}$, which decreases Avg. by 1.8 points. This indicates that relevance-only selection tends to provide less complementary evidence than the proposed representative selection. Overall, these ablations demonstrate the complementary roles of the three stages and validate the key designs within History Refocusing.

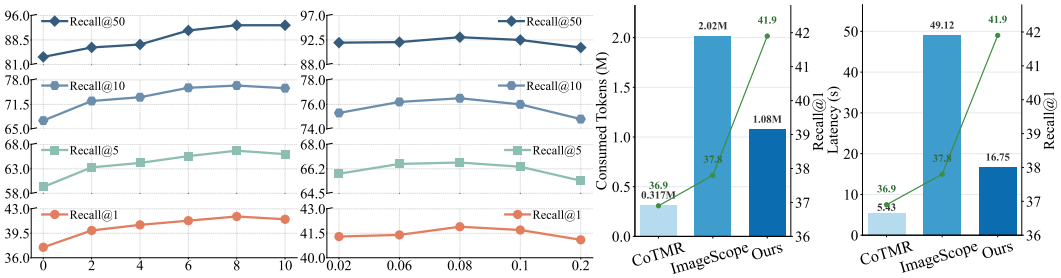
Table 4: Ablation study of InAHR on our benchmark. Δ denotes the Avg. difference relative to the default InAHR.

| Method | R@1 | R@5 | R@10 | Avg. | Δ |
|---|---|---|---|---|---|
| **InAHR** | **41.9** | **66.7** | **76.5** | **61.7** | 0.0 |
| w/o Intent-State Induction | 40.1 | 64.5 | 74.2 | 59.6 | -2.1 |
| w/o History Refocusing | 37.5 | 59.3 | 67.2 | 54.7 | -7.0 |
| Single-Facet Reconstruction | 41.3 | 65.8 | 75.3 | 60.8 | -0.9 |
| Hard Turn Mask | 40.9 | 65.2 | 74.8 | 60.3 | -1.4 |
| w/o Frame Relevance | 40.3 | 64.2 | 73.8 | 59.4 | -2.3 |
| w/o Representative Selection | 40.5 | 64.7 | 74.4 | 59.9 | -1.8 |

### 5.3 In-depth Analysis

**History Sensitivity Analysis.** We analyze history sensitivity by comparing two settings: *Last-turn*, where the model observes only the latest reference video and latest feedback, and *Full history*, where the model observes the entire interaction sequence up to the latest feedback. As shown in Table 3, Full history improves performance in most history-dependent scenarios, especially *Combination*, *Rollback*, and *Progressive*, where earlier constraints remain useful. However, it hurts *Preference Change*, indicating that outdated historical constraints can mislead retrieval when intent shifts. Compared with baselines, InAHR better exploits useful history while being less affected by distracting context, achieving clear Full-history gains in history-dependent scenarios and a smaller drop under *Preference Change*. These results show that LTCVR requires adaptive history utilization rather than simply longer context.

**Parameter Analysis.** We analyze two hyperparameters in the refocusing module: the evidence-frame budget $M$ and the intent relevance temperature $\tau$. As shown in Figure 5(a), retrieval performance improves as $M$ increases from 0 to 8, indicating that re-injecting more intent-relevant historical evidence helps the MLLM generate more grounded target descriptions. However, increasing $M$ brings limited gains and can slightly degrade performance, since excessive evidence may reintroduce




(a) Evidence Budget $M$ (b) Relevance Temperature $\tau$ (c) Token Consumption (d) Latency

Figure 5: Parameter and efficiency analysis of InAHR.

redundant or weakly relevant visual context. This suggests that InAHR benefits from a compact evidence coreset rather than using more historical frames. Figure 5(b) shows a similar sensitivity pattern for $\tau$. A very small $\tau$ overly concentrates the hierarchical relevance on a few highly matched turns and frames, whereas a large $\tau$ weakens visual intent conditioning and places greater emphasis on representative coverage. The best performance is achieved at a moderate $\tau$, suggesting that effective history refocusing requires an appropriate balance between intent relevance and evidence coverage.

**Latency and Efficiency Analysis.** We further compare representative MLLM-based methods by total output tokens and average inference latency. CoTMR is fastest, requiring 5.43s per session and 0.317M tokens, but has limited retrieval accuracy with 36.9 R@1. ImageScope incurs the highest cost, consuming 2.02M tokens and 49.12s per session, while reaching only 37.8 R@1. InAHR uses 16.75s and 1.08M tokens, higher than CoTMR but far lower than ImageScope, and achieves the best R@1 of 41.9. This shows that selecting and re-injecting compact intent-relevant evidence adds moderate cost but offers a stronger accuracy–efficiency trade-off.

**Intent Understanding Evaluation.** Table 5 reports the Intent Understanding Evaluation on the ALL split. We observe that InAHR consistently achieves the best scores across all three levels, including description quality, reasoning quality, and retrieval-level intent satisfaction. The gains are especially clear on *History Grounding* and *Scenario Faithfulness* in both the description and reasoning stages, showing that our method more accurately determines which historical information should be preserved, revised, or discarded under different intent-evolution scenarios. Although all methods exhibit a larger performance drop at the retrieval level, reflecting the additional difficulty of translating intent understanding into discriminative retrieval, InAHR still obtains the highest retrieval-level scores, indicating that its advantage comes not only from generating more plausible explanations, but from forming more faithful and retrieval-ready intent representations.

Table 5: Intent Understanding Evaluation on our benchmark. **Desc.**, **Reas.**, and **Retr.** denote the description, reasoning, and retrieval levels, respectively; **IC**, **HG**, and **SF** denote Intent Consistency, History Grounding, and Scenario Faithfulness.

| Model | Desc. | | | Reas. | | | Retr. | | |
|---|---|---|---|---|---|---|---|---|---|
| | IC | HG | SF | IC | HG | SF | IC | HG | SF |
| OSrCIR | 8.6 | 7.9 | 7.2 | 8.8 | 8.1 | 7.6 | 5.9 | 6.1 | 5.9 |
| AutoCIR | 8.3 | 7.6 | 6.8 | 8.6 | 7.8 | 7.1 | 5.8 | 5.9 | 5.6 |
| CoTMR | 8.5 | 8.0 | 7.4 | 8.9 | 8.5 | 8.1 | 5.8 | 6.0 | 5.7 |
| ImageScope | 8.2 | 7.7 | 6.9 | 8.7 | 8.2 | 7.8 | 6.2 | 6.4 | 6.1 |
| **Ours** | **8.8** | **8.6** | **8.2** | **9.2** | **9.1** | **8.8** | **6.8** | **7.0** | **6.8** |

## 6 Conclusion

We present LTCVR, a benchmark for long-term composed video retrieval with five representative history-dependency patterns that capture how user intent evolves across interactions. Evaluation on LTCVR shows that existing embedding-based and MLLM-based methods struggle to adaptively utilize long-term interaction history. To address this challenge, we propose InAHR, an Intent-Driven Adaptive History Refocusing framework that first induces the current intent, then adaptively refocuses historical evidence from the turn level to the frame level, and finally reconstructs the target through multiple complementary retrieval-oriented descriptions. Extensive experiments, ablations, and intent-understanding analyses demonstrate the effectiveness and generality of InAHR across diverse long-horizon scenarios. Future work will extend LTCVR to longer and more diverse interactions and explore more efficient history refocusing for large-scale retrieval.

### AI USE STATEMENT

In this work, we used generative AI tools for language editing and polishing, literature retrieval and discovery, and synthetic data generation for the LTCVR benchmark. Specifically, generative AI was used to generate progressive feedback, rewrite interactions into different history-dependency scenarios, and assist with automatic quality screening and semantic-equivalence checking. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

### REPRODUCIBILITY STATEMENT

We are committed to the reproducibility of our research. We provide the source code, environment setup instructions, and a step-by-step guide for reproducing our results in our anonymous GitHub repository. Further details on the LTCVR benchmark construction and quality control, InAHR implementation, prompt design, baseline adaptations, and hyperparameter settings are provided in the appendices.

## REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1728–1738, 2021.

Shengcao Cao, Tanmaya Shekhar Dabral, Zhongli Ding, Madhuri Shanbhogue, Kaifeng Chen, Zhe Li, Mojtaba Seyedhosseini, Yu-Xiong Wang, and Liang-Yan Gui. CoCo-IR: Contextual composed image retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 364–382. Springer, 2026.

Yanzhe Chen, Zhiwen Yang, Jinglin Xu, and Yuxin Peng. MAI: A multi-turn aggregation-iteration model for composed image retrieval. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

Zhangtao Cheng, Yuhao Ma, Jian Lang, Kunpeng Zhang, Ting Zhong, Yong Wang, and Fan Zhou. Generative thinking, corrective action: User-friendly composed image retrieval via automatic multi-agent collaboration. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 334–344, 2025.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2829, 2023.

Google DeepMind. Gemini 3 Flash model card. Technical report, Google DeepMind, 2025. URL `https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Flash-Model-Card.pdf`.

Geonmo Gu, Sanghyuk Chun, Wonjae Kim, Yoohoon Kang, and Sangdoo Yun. Language-only efficient training of zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13225–13234, 2024.

Animesh Gupta, Jay Parmar, Ishan Rajendrakumar Dave, and Mubarak Shah. From play to replay: Composed video retrieval for temporally fine-grained videos. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 147996–148011, 2025.

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 ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pp. 127–137, 2025.

Thomas Hummel, Shyamgopal Karthik, Mariana-Iuliana Georgescu, and Zeynep Akata. EgoCVR: An egocentric benchmark for fine-grained composed video retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 1–17, 2024.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 12888–12900. PMLR, 2022.

You Li, Fan Ma, and Yi Yang. Imagine and seek: Improving composed image retrieval with an imagined proxy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3984–3993, 2025.

Hui Lin and Jeff Bilmes. A class of submodular functions for document summarization. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL)*, pp. 510–520. Association for Computational Linguistics, 2011.

Hui Lin, Jeff Bilmes, and Shasha Xie. Graph-based submodular selection for extractive summarization. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 381–386. IEEE, 2009.

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 Web Conference (WWW)*, pp. 1666–1682, 2025.

Arianna Moccia, Matthew Plummer, Ivor Simpson, and Alexa M. Morcom. The effects of external cue overlap and internal goals on selective memory retrieval as revealed by electroencephalographic (EEG) neural pattern reinstatement. *European Journal of Neuroscience*, 62(1):e70194, 2025. doi: 10.1111/ejn.70194.

OpenAI. GPT-4o system card, 2024. URL `https://openai.com/index/gpt-4o-system-card/`.

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 (ICCV)*, pp. 11323–11334, 2023.

Bill Psomas, George Retsinas, Nikos Efthymiadis, Panagiotis Filntisis, Yannis Avrithis, Petros Maragos, Ondrej Chum, and Giorgos Tolias. Instance-level composed image retrieval. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 76737–76782, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 8748–8763. PMLR, 2021.

Melinda Sabo, Edmund Wascher, and Daniel Schneider. The neural basis of attentional selection in goal-directed memory retrieval. *Scientific Reports*, 14:20937, 2024. doi: 10.1038/s41598-024-71691-x.

Kuniaki Saito, Kihyuk Sohn, Xiang Zhang, Chun-Liang Li, Chen-Yu Lee, Kate Saenko, and Tomas Pfister. Pic2Word: Mapping pictures to words for zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19305–19314, 2023.

Zelong Sun, Dong Jing, and Zhiwu Lu. CoTMR: Chain-of-thought multi-scale reasoning for training-free zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22675–22684, 2025.

Yuanmin Tang, Jue Zhang, Xiaoting Qin, Jing Yu, Gaopeng Gou, Gang Xiong, Qingwei Lin, Saravan Rajmohan, Dongmei Zhang, and Qi Wu. Reason-before-retrieve: One-stage reflective chain-of-thoughts for training-free zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14400–14410, 2025.

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 (CVPR)*, 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 (ICCV)*, 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 (AAAI)*, 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 (CVPR)*, pp. 6439–6448, 2019.

Yue Wu, Zhaobo Qi, Yiling Wu, Junshu Sun, Yaowei Wang, and Shuhui Wang. Learning fine-grained representations through textual token disentanglement in composed video retrieval. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. MSR-VTT: A large video description dataset for bridging video and language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5288–5296, 2016.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Yifei Yuan and Wai Lam. Conversational fashion image retrieval via multiturn natural language feedback. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pp. 839–848, 2021.

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 (ICCV)*, pp. 22120–22130, 2025.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

## APPENDIX CONTENTS


**A Additional Related Work** . . . 13
**B Additional Details of LTCVR** . . . 14
B.1 Reference-Target Pair Matching . . . 14
B.2 Seed Interaction Generation . . . 14
B.3 Scenario-specific Interaction Generation . . . 15
B.4 Quality Control . . . 18
B.5 Intent Understanding Evaluation . . . 20
**C Additional Details of InAHR** . . . 22
C.1 Prompt Details . . . 23
**D Additional Details of Experimental Settings** . . . 25
D.1 Baselines . . . 25
D.2 Implementation Details . . . 26
**E Relevant-Turn Identification Analysis** . . . 26
**F Generalization Analysis** . . . 27
F.1 Results on LTCVR using different MLLM backbones. . . . 27
F.2 Results on LTCVR using different retrieval backbones. . . . 27
F.3 Results on WebVid-CoVR test set. . . . 27
F.4 Robustness to Longer Interaction Histories . . . 28
**G Qualitative Analysis** . . . 29
**H Limitations** . . . 29


## A ADDITIONAL RELATED WORK

**Multi-turn Composed Image Retrieval.** Multi-turn composed image retrieval extends conventional CIR by allowing the retrieval target to evolve through successive reference–feedback interactions. Early works such as MT-FashionIQ (Yuan & Lam, 2021) study iterative refinement in the fashion domain, while FashionNTM (Pal et al., 2023) introduces memory-based history modeling to accumulate information across turns. FashionMT (Chen et al., 2025) further considers retrospective dependencies, showing that later requests may depend on earlier interactions rather than only the latest turn. More recently, CoCo-IR (Cao et al., 2026) models richer contextual dependencies over interleaved image–text histories. These studies establish the importance of history modeling in composed retrieval, but remain centered on static images and cover a relatively limited set of history-dependency patterns. In contrast, LTCVR targets long-term video retrieval and explicitly controls five forms of intent evolution—progressive refinement, rollback, cross-turn combination, local correction, and preference change—thereby evaluating whether a model can accumulate, revisit, combine, revise, or discard historical information while reasoning over dynamic visual content.

**Interactive and Dialogue-based Video Retrieval.** Another related line of work studies interactive video retrieval through natural-language dialogue. IVCR-200K (Han et al., 2025) formulates Interactive Video Corpus Retrieval as a multi-turn conversational task and supports both whole-video and video-moment retrieval; its accompanying InterLLaVA framework leverages MLLMs and dialogue context to improve retrieval and provide explainable responses. Recent interactive text-to-video retrieval methods such as UMIVR (Zhang et al., 2025) instead generate targeted clarification questions to iteratively reduce ambiguity in the user's textual query. These settings differ from long-term CVR in the form of interaction. In LTCVR, each turn is a *composed multimodal query* consisting of a reference video and a textual modification, so the model must determine which visual properties should be preserved, revised, recovered, combined, or discarded across turns. Existing dialogue-based video retrieval settings are primarily designed to progressively clarify a textual search intent, rather than to systematically evaluate such reference-dependent multimodal history relations. Accordingly, LTCVR complements these works by providing controlled history-dependency scenarios and explicit intent-level evaluation for long-term composed video retrieval.

# B Additional Details of LTCVR

## B.1 Reference-Target Pair Matching

This subsection provides the detailed procedure used to obtain semantically suitable reference-target pairs for constructing LTCVR sessions. The goal is to reduce two degenerate cases: pairs that are almost identical and can be solved without meaningful modification reasoning, and pairs that are too far apart and would require unrealistic semantic jumps. We therefore first group videos into local semantic neighborhoods and then retain only reference-target pairs whose similarity lies in a moderate neighborhood-specific range.

Given the raw video collection $\mathcal{V}_{src} = \{v_j\}_{j=1}^{N_{\text{src}}}$ with $N_{\text{src}} = 120{,}000$ videos, we uniformly sample $U_{\text{match}} = 10$ frames from each video and encode them with the CLIP(ViT-G/14) visual encoder $\phi_{\text{match}}(\cdot)$. The video-level matching embedding is obtained by mean pooling and $\ell_2$ normalization:

$$\mathbf{m}_j = \text{Norm}_2\left(\frac{1}{U_{\text{match}}}\sum_{\ell=1}^{U_{\text{match}}}\phi_{\text{match}}(\tilde{f}_{j,\ell})\right), \tag{4}$$

where $\tilde{f}_{j,\ell}$ denotes the $\ell$-th sampled frame used only for reference-target matching. We then partition the videos into $G = 600$ semantic neighborhoods using $k$-means on $\ell_2$-normalized matching embeddings:

$$\{\Omega_g, \boldsymbol{\mu}_g\}_{g=1}^{G} = \arg\min_{\{\Omega_g,\boldsymbol{\mu}_g\}_{g=1}^{G}} \sum_{g=1}^{G}\sum_{v_j\in\Omega_g} \left\|\mathbf{m}_j - \boldsymbol{\mu}_g\right\|_2^2, \quad \|\mathbf{m}_j\|_2 = 1. \tag{5}$$

where $\Omega_g$ denotes the $g$-th semantic neighborhood, and $\boldsymbol{\mu}_g$ is its cluster centroid. This clustering step restricts pair sampling to semantically coherent neighborhoods, making the resulting reference-target pairs more suitable for composed video retrieval.

For each semantic neighborhood $\Omega_g$, we enumerate candidate ordered pairs

$$\mathcal{Y}_g = \{(v_a, v_b) \mid v_a, v_b \in \Omega_g,\ a \neq b\}, \tag{6}$$

where the first element $v_a$ serves as the initial reference video and the second element $v_b$ serves as the target video. The semantic similarity of a candidate pair is computed as cosine similarity:

$$\sigma(v_a, v_b) = \mathbf{m}_a^\top \mathbf{m}_b. \tag{7}$$

Instead of using a single global threshold, we define the moderate-similarity band within each semantic neighborhood to account for different intra-neighborhood density. Let

$$\rho_g^- = Q_{q_{\text{low}}}\left(\{\sigma(v_a, v_b) : (v_a, v_b) \in \mathcal{Y}_g\}\right), \quad \rho_g^+ = Q_{q_{\text{high}}}\left(\{\sigma(v_a, v_b) : (v_a, v_b) \in \mathcal{Y}_g\}\right), \tag{8}$$

where $Q_q(\cdot)$ denotes the $q$-quantile of the similarity distribution in semantic neighborhood $\Omega_g$. Here, $q_{\text{low}}$ and $q_{\text{high}}$ denote the lower and upper quantile levels, respectively; we set $q_{\text{low}} = 0.2$ and $q_{\text{high}} = 0.8$ in our implementation. We keep only pairs whose similarity lies within this neighborhood-wise moderate range:

$$\widetilde{\mathcal{Y}}_g = \{(v_a, v_b) \in \mathcal{Y}_g \mid \rho_g^- \leq \sigma(v_a, v_b) \leq \rho_g^+\}. \tag{9}$$

The lower bound removes pairs that are semantically too distant, while the upper bound removes near-duplicate pairs that require little modification reasoning. The final reference-target pool is

$$\widetilde{\mathcal{Y}} = \bigcup_{g=1}^{G} \widetilde{\mathcal{Y}}_g. \tag{10}$$

All subsequent seed trajectories are generated from pairs sampled from $\widetilde{\mathcal{Y}}$.

## B.2 Seed Interaction Generation

The retained reference-target pairs are first converted into the *Progressive Refinement* scenario, where the target video remains fixed and each turn introduces one or two additional modifications from the latest reference toward the target.

For a retained pair $(v_1, v^*) \in \widetilde{\mathcal{Y}}$, we use $v_1$ as the initial reference video and keep $v^*$ fixed as the session target. At turn $n$, MLLM (GPT-4o) generates a user-style feedback instruction conditioned on the latest reference video, the target video, and the preceding interaction history:

$$u_n = MLLM_{\text{seed}}\left(v_n, v^*, \mathcal{H}_{<n}\right), \tag{11}$$

where

$$\mathcal{H}_{<n} = \{(v_i, u_i)\}_{i=1}^{n-1} \tag{12}$$

denotes the preceding interaction turns. The retrieval system then returns the next reference video:

$$v_{n+1} = \arg\max_{v \in \mathcal{V}} \text{sim}\left(\Psi(v_n, u_n), \Phi(v)\right), \tag{13}$$

where $\Psi(\cdot)$ denotes the query representation induced by the latest reference video and the corresponding feedback, and $\Phi(\cdot)$ denotes the candidate video representation. The retrieved video $v_{n+1}$ becomes the reference for the next turn, while the target $v^*$ remains fixed. We repeat this process for $N = 5$ turns to obtain a progressive seed interaction:

$$\mathcal{T}^{\text{prog}} = \{(v_n, u_n)\}_{n=1}^{N}, \quad \text{with target } v^*. \tag{14}$$

We discard a seed interaction if its retrieved reference sequence contains repeated videos. After filtering, we obtain $40{,}000$ five-turn progressive seed interactions.

**Prompt for seed feedback generation.** We use the following prompt template to generate each progressive feedback instruction. The placeholders are filled with uniformly sampled frames and captions of the latest reference video, the target video, and the preceding turns.

**Prompt 1: Seed Feedback Generation**

**# SCENARIO**
You are a user trying to find a specific video (the "Target Scene") through an iterative video retrieval process. You have received the latest retrieved result and want to refine it toward the target scene.
**## 1. THE GOAL (What you want)**
Target Scene:
- Topic: `[[TARGET_CAPTION]]`
- Visual Details: `[[TARGET_FRAMES]]`
**## 2. THE LOG (Previous refinement history)**
Turn `[[PREV_TURN_IDX]]`
- Previous dialogue history: `[[HISTORY]]`
- Result: The assistant returned the latest reference scene below.
**## 3. LATEST STATUS (The latest retrieved result)**
Latest Reference Scene:
- Topic: `[[LATEST_REFERENCE_CAPTION]]`
- Visual Details: `[[LATEST_REFERENCE_FRAMES]]`
**## Step 1: Visual Difference Analysis**
Compare the latest reference video with the fixed target video. Identify one or two salient visible changes involving scene, subject, object, action, appearance, camera style, or temporal dynamics.
**## Step 2: User Feedback Writing**
Write one concise user-style feedback sentence that moves the search from the latest reference video toward the target video. The feedback must be grounded in visible differences and suitable as a retrieval query.
**# Constraints**
Do not directly mention that you are comparing two videos. Do not reveal the target video identity. Do not introduce objects, actions, or attributes that are not visible in the target. Do not repeat a previous feedback instruction unless the new sentence adds a clear refinement.
**# Output Format**
Return a JSON object {"feedback": "..."}

### B.3 Scenario-specific Interaction Generation

Progressive seed trajectories provide coherent multi-turn interactions, but mainly capture monotonic refinement. To cover more realistic history-dependent intent evolution, we rewrite seed interactions into four additional scenarios: *Rollback*, *Cross-turn Combination*, *Local Correction*, and *Preference Change*. In all cases, the rewritten feedback is required to remain visually grounded, natural, and consistent with the intended scenario.

**Rollback.** *Rollback* simulates the case where the user returns to an earlier interaction state and continues from that earlier reference state. Given a progressive seed interaction $\mathcal{T}^{\text{prog}}$, we first sample an anchor turn from the intermediate history: $a \sim \text{Uniform}(\{2, \ldots, N-1\})$, which corresponds to turns 2, 3, and 4 when $N = 5$. We then rewrite the final feedback to explicitly reactivate the earlier reference $v_a$ rather than the most recent reference $v_N$. Formally, the rewritten feedback is

$$u_N^{\text{roll}} = MLLM_{\text{roll}}\left(v_a, v_N, v^*, \mathcal{H}_{<N}\right). \tag{15}$$

The target remains $v^*$, but the final intent depends on correctly identifying the anchor reference.

**Prompt 2: Rollback Scenario Rewriting**

**# Task Description**
You are rewriting a long-term video retrieval dialogue into a rollback scenario. The latest user intent should return to an earlier reference video and continue the search from that earlier reference state.
**## Inputs**
- Most recent reference video: `[[LATEST_REFERENCE_FRAMES]]`
- Anchor reference video: `[[ANCHOR_REFERENCE_FRAMES]]`
- Anchor turn index: `[[ANCHOR_TURN]]`
- Fixed target video: `[[TARGET_FRAMES]]`
- Previous dialogue history: `[[HISTORY]]`

**## Step 1: Anchor Identification**
Inspect the anchor reference video and determine which visible properties should be reused as the active reference state. The most recent reference should not be treated as the only source of context.
**## Step 2: Rollback Feedback Writing**
Write the latest user feedback so that the user clearly goes back to the earlier reference video and then asks for one or two visible modifications toward the target. Use natural expressions such as "Compared to the most recent turn, I still prefer the video from turn `[[ANCHOR_TURN]]`", "use that earlier result", or "Building on that".
**## Constraints**
The feedback must not be answerable from the most recent reference alone. The key reference state must be the anchor reference, not the latest reference. Do not introduce target details that are not visible.
**## Output Format**
Return a JSON object {"rewritten_feedback": "..."}.

**Cross-turn Combination.** *Cross-turn Combination* requires the final retrieval intent to combine constraints from different interaction turns. Given a progressive seed interaction $\mathcal{T}^{\text{prog}}$, we sample the earlier anchor turn in the same way as Rollback: $a \sim \text{Uniform}(\{2, \ldots, N-1\})$. We then rewrite the final feedback to combine visual constraints from the earlier reference $v_a$ and the latest reference $v_N$. Formally, the rewritten feedback is

$$u_N^{\text{comb}} = MLLM_{\text{comb}}\left(v_a, v_N, v^*, \mathcal{H}_{<N}\right). \tag{16}$$

The generated instruction preserves an attribute or visual constraint from $v_a$ while also incorporating information from the latest reference $v_N$.

**Prompt 3: Cross-Turn Combination Rewriting**

**# Task Description**
You are rewriting a long-term video retrieval dialogue into a cross-turn combination scenario. The desired target should combine visual constraints from the latest reference video and a selected earlier reference video.
**## Inputs**
- Latest reference video: `[[LATEST_REFERENCE_FRAMES]]`
- Earlier reference video: `[[EARLIER_REFERENCE_FRAMES]]`
- Anchor turn index: `[[ANCHOR_TURN]]`
- Target video: `[[TARGET_FRAMES]]`
- Previous dialogue history: `[[HISTORY]]`

**## Step 1: Constraint Extraction**
Identify one visible constraint that should be kept from the latest reference video and one visible constraint that should be recovered or preserved from the earlier reference video.

**## Step 2: Combination Feedback Writing**
Write the latest user feedback so that the desired target combines both constraints. Use natural language such as "keep ... from the latest reference, but bring back ... from turn `[[ANCHOR_TURN]]`" or "combine ... with ... from the earlier video".
**## Constraints**
The instruction must require cross-turn reasoning and should not be solvable from only one reference video. Describe one or two target-grounded modifications and avoid unsupported details.
**## Output Format**
Return a JSON object {"rewritten_feedback": "..."}.

**Local Correction.** *Local Correction* simulates the case where the user corrects a specific error in the latest reference while preserving the overall retrieval intent. Given a progressive seed interaction $\mathcal{T}^{\text{prog}}$, we rewrite the final feedback based on the latest reference $v_N$, the target video $v^*$, the original final feedback $u_N$, and the previous interaction history:

$$u_N^{\text{corr}} = MLLM_{\text{corr}}\left(v_N, v^*, u_N, \mathcal{H}_{<N}\right). \tag{17}$$

Unlike *Rollback* or *Cross-turn Combination*, this scenario mainly depends on the most recent reference $v_N$, and the rewritten feedback is expressed as a targeted correction rather than a generic refinement.

**Prompt 4: Local Correction Rewriting**

**# Task Description**
You are rewriting a long-term video retrieval dialogue into a local correction scenario. The user should correct one specific mismatch in the latest retrieved result while preserving the broader search intent.
**## Inputs**
- Latest reference video: `[[LATEST_REFERENCE_FRAMES]]`
- Target video: `[[TARGET_FRAMES]]`
- Original feedback instruction: `[[ORIGINAL_FEEDBACK]]`
- Previous dialogue history: `[[HISTORY]]`

**## Step 1: Local Mismatch Identification**
Identify one visible attribute, object, action, scene detail, or style difference that should be corrected between the latest reference and the target.
**## Step 2: Correction Feedback Writing**
Rewrite the latest feedback as a self-correction. The sentence should sound like a user correcting a detail, for example using "not ..., but ...", "more accurately, ...", "make it closer by changing ...", or "keep the rest, but fix ...".
**## Constraints**
The feedback must preserve the overall retrieval direction and must not reset the whole target. It must not contradict the previous trajectory or introduce unsupported details.
**## Output Format**
Return a JSON object {"rewritten_feedback": "..."}.

**Preference Change.** *Preference Change* simulates the case where the user shifts to a new target, making previous constraints partly or entirely obsolete. Given a progressive seed interaction $\mathcal{T}^{\text{prog}}$, we sample a new target $v^*_{\text{new}}$ from the same semantic neighborhood as the original target $v^*$, while requiring it to be different from $v^*$ and moderately similar to the latest reference $v_N$:

$$v^*_{\text{new}} \in \left\{ v \in \Omega_g \mid v \neq v^*,\ \rho_g^- \leq \sigma(v_N, v) \leq \rho_g^+ \right\}, \quad v^* \in \Omega_g. \tag{18}$$

We then rewrite the final feedback to express the new preference:

$$u_N^{\text{pref}} = MLLM_{\text{pref}}\left(v_N, v^*_{\text{new}}, \mathcal{H}_{<N}\right). \tag{19}$$

The session target is replaced by $v^*_{\text{new}}$ in the rewritten trajectory.

**Prompt 5: Preference-Change Rewriting**

**# Task Description**
You are rewriting a long-term video retrieval dialogue into a preference-change scenario. The user should shift to a new target, making previous constraints partly or entirely obsolete.
**## Inputs**
- Previous dialogue history: `[[HISTORY]]`
- Latest reference video: `[[LATEST_REFERENCE_FRAMES]]`
- Newly selected target video: `[[NEW_TARGET_FRAMES]]`
**## Step 1: New Target Analysis**
Compare the latest reference video with the newly selected target video. Identify one or two visible modifications that redirect the search toward the new target.
**## Step 2: Preference-Shift Feedback Writing**
Write the latest user feedback so that the user clearly changes preference. Use natural expressions such as "actually, I now prefer ...", "instead, I want ...", or "forget the previous direction and look for ...".
**## Output Format**
Return a JSON object {"rewritten_feedback": "..."}.

### B.4 Quality Control

We apply a four-stage quality-control pipeline consisting of structural pre-filtering, MLLM-based automatic screening, category-balanced sampling, and human review. The pipeline is designed to ensure that each retained session forms a valid retrieval trajectory, contains non-redundant feedback, exhibits the intended history dependency, and remains visually consistent with the final target.

**Structural pre-filtering.** We first remove sessions with invalid video-transition structures. For a session with reference sequence $\mathcal{V}_{\text{ref}} = (v_1, v_2, \ldots, v_N)$, we discard it if the sequence contains repeated references:

$$\exists\, i < j, \quad v_i = v_j. \tag{20}$$

This removes cyclic trajectories such as $A \to B \to A$ as well as trajectories that repeatedly revisit the same retrieved result without meaningful progression.

We further remove sessions containing redundant feedback. Specifically, we use Qwen3-8B (Yang et al., 2025) as an LLM-based semantic equivalence checker to compare feedback instructions within the same session:

$$\eta_{ij} = LLM(u_i, u_j), \quad 1 \le i < j \le N, \tag{21}$$

where $\eta_{ij} = 1$ indicates that $u_i$ and $u_j$ express the same or nearly identical modification intent. A session is discarded if

$$\exists\, i < j, \quad \eta_{ij} = 1. \tag{22}$$

This filtering ensures that each turn contributes a non-trivial update to the evolving retrieval intent.

**MLLM-based automatic screening.** After structural filtering, each candidate session is independently screened by GPT-4o. The screening stage evaluates four dimensions on a 1–5 scale: *visual grounding*, *instruction clarity*, *history dependency*, and *scenario consistency*. The scoring rubric is defined as follows:

- **5**: fully valid; the feedback is visually grounded, clear, non-redundant, and strongly consistent with the assigned scenario.
- **4**: mostly valid; only minor wording or grounding imperfections are present.
- **3**: acceptable but imperfect; the intended scenario is recognizable, although some grounding or dependency may remain ambiguous.
- **2**: weak; the feedback is vague, redundant, weakly grounded, or only loosely follows the intended history dependency.
- **1**: invalid; the feedback is hallucinated, contradicts the visual evidence, or fails to follow the assigned scenario.

A session is retained only if all four scores are at least 3. After this screening stage, 16,190 candidate sessions remain.

Because GPT-4o is also used during session generation, we explicitly separate the generation and screening stages through different prompts and task roles. During generation, GPT-4o is instructed to produce grounded user feedback toward a specified target, whereas during screening it acts only as an evaluator of the completed interaction trajectory. The final inclusion decision is further subjected to independent human review, reducing reliance on a single model-based filtering stage.

**Prompt 6: Automatic Quality Screening**

**# Task Description**
You are validating a long-term composed video retrieval benchmark sample. Your task is to decide whether the sample is visually grounded, clear, history-dependent, and consistent with its assigned scenario.
**## Inputs**
- Scenario label: `[[SCENARIO]]`
- Interaction history: `[[INTERACTION_HISTORY]]`
- Target video: `[[TARGET_FRAMES]]`
**## Step 1: Score the Sample**
Rate the sample from 1 to 5 on four dimensions: visual grounding, instruction clarity, history dependency, and scenario consistency. A score of 5 means fully correct; 4 means mostly correct; 3 means acceptable but imperfect; 2 means weak; and 1 means invalid.
**## Step 2: Reflection**
Before making the final decision, reflect on whether your scores are consistent with the visual evidence, the interaction history, and the assigned scenario. Check whether the feedback is grounded in the videos, whether it depends on the required historical context, and whether it matches the intended scenario. Revise your scores if necessary.
**## Acceptance Rule**
Return `accept` only if all four scores are at least 3 after reflection. If any score is below 3, return `reject` with a short reason.
**## Output Format**
Return a JSON object:
{ `"visual_grounding"`: 1-5,
`"instruction_clarity"`: 1-5,
`"history_dependency"`: 1-5,
`"scenario_consistency"`: 1-5,
`"decision"`: `"accept"` or `"reject"`,
`"reason"`: `"..."` }.

**Balanced sampling.** To improve data diversity, we categorize the 16,190 retained sessions along two axes: video content category and modification type. The content taxonomy covers major visual domains including animals, nature, people, urban scenes, objects, vehicles, arts, and industry. Modification types include scene-level, subject-level, appearance-level, action-level, camera/style-level, and other modifications.

For session $q$, we denote the assigned labels as

$$(c^{(q)}, m^{(q)}) = MLLM_{\text{cat}}(q), \tag{23}$$

where $c^{(q)}$ and $m^{(q)}$ denote its content category and modification type, respectively.

Based on these labels, we perform approximately balanced sampling across content categories and modification types while maintaining coverage of all five history-dependency scenarios. This step prevents LTCVR from being dominated by a small number of frequent visual domains or edit patterns. We finally sample 4,000 candidate sessions for human review.

**Human review.** All 4,000 sampled sessions are independently evaluated by three human reviewers. For each session, reviewers are shown the assigned scenario, the complete ordered reference-feedback history, and the final target video. Each reviewer rates the session on a 1–5 scale along three dimensions:

- **Accuracy**: whether the feedback and the active historical constraints are visually consistent with the final target video.
- **History dependency**: whether resolving the current intent requires the intended use of interaction history, such as accumulating, revisiting, combining, revising, or discarding previous information.
- **Scenario consistency**: whether the interaction follows its assigned history-dependency pattern rather than another scenario type.

Table 6: Inter-reviewer agreement computed from the initial independent human ratings before revision.

| Dimension | Fleiss' $\kappa$ | Exact Agree. | Within-1 Agree. |
|---|---|---|---|
| Accuracy | 0.53 | 67.2% | 86.1% |
| History Dependency | 0.48 | 62.8% | 84.3% |
| Scenario Consistency | 0.52 | 66.5% | 85.7% |

A session is directly accepted only if the mean score for every dimension across the three reviewers is at least 4 and no individual reviewer assigns a score below 3 to any dimension. Under this criterion, 2,283 sessions (57.1%) are accepted without modification.

The remaining 1,717 sessions are approximately evenly assigned to the three reviewers for manual curation. Among them, 1,185 sessions (29.6%) are judged repairable and revised according to the initial review comments. Typical revisions include clarifying ambiguous wording, strengthening the intended history dependency, and correcting minor cross-turn inconsistencies. The other 532 sessions (13.3%) contain non-repairable problems, including invalid visual grounding, inconsistency with the final target, incorrect history dependency, or irreparable cross-turn contradictions. These sessions are discarded and replaced with newly sampled candidates from the same history-dependency scenario.

To avoid self-validation, a reviewer who revises a session does not participate in its subsequent validation. Each revised or replacement session is independently re-evaluated by the other two reviewers using the same three-dimensional rubric and acceptance criterion. All 1,717 revised or replacement sessions pass this re-evaluation, yielding the final benchmark of 4,000 sessions.

We further quantify inter-reviewer agreement using the initial independent ratings before any revision. As reported in Table 6, Fleiss' $\kappa$ values are 0.53, 0.48, and 0.52 for Accuracy, History Dependency, and Scenario Consistency, respectively. The corresponding pairwise exact agreement rates are 67.2%, 62.8%, and 66.5%, while the within-one-point agreement rates reach 86.1%, 84.3%, and 85.7%. These results show consistent reviewer agreement across all three dimensions, with high within-one-point agreement indicating that rating differences are generally small.

### B.5 Intent Understanding Evaluation

This subsection describes the judge-based Intent Understanding Evaluation used in Sec. 3. This evaluation checks whether a model correctly understands the current retrieval intent in a multi-turn session, rather than only whether the annotated target video appears in the retrieved list. For each sample, the judge is given the full interaction history, the scenario label, and the target video as reference information. It then scores the model output at three levels: the generated target descriptions, the intermediate reasoning, and the top-5 retrieved videos.

**Model input and output.** All MLLM-based methods use the same LTCVR input format. The input contains the preceding interaction history, where each turn includes uniformly sampled frames from the reference video and the corresponding user feedback, together with the latest reference video and latest feedback for the current retrieval request. For this evaluation, we record three outputs from each method: its reasoning text, its final target descriptions, and its top-5 retrieved videos. If a method does not have an explicit reasoning field, we use the textual rationale generated before the final target description.

**Scoring dimensions.** At each level, the judge assigns three scores from 0 to 10. *Intent Consistency* measures whether the evaluated output matches the final retrieval intent. *History Grounding* measures

Table 7: Input and output fields used in Intent Understanding Evaluation.

| Stage | Input | Output |
|---|---|---|
| Evaluated model | Preceding interaction history $\mathcal{H}_{<n} = \{(v_i, u_i)\}_{i=1}^{n-1}$, latest reference video $v_n$ and latest feedback $u_n$. | Reasoning text $r_n$, retrieval-oriented target description $d_n$, and top-5 retrieved videos $\{\hat{v}_{n,j}\}_{j=1}^{5}$. |
| Description-level judge | Scenario label, full interaction history, target-video frames and caption, and generated target description $d_n$. | Scores for Intent Consistency, History Grounding, and Scenario Faithfulness, with a short rationale. |
| Reasoning-level judge | Scenario label, full interaction history, target-video frames and caption, and model reasoning $r_n$. | Scores for Intent Consistency, History Grounding, and Scenario Faithfulness, with a short rationale. |
| Retrieval-level judge | Scenario label, full interaction history, target-video frames and caption, latest feedback $u_n$, and top-5 retrieved video frames and captions. | Scores for Intent Consistency, History Grounding, and Scenario Faithfulness, with a short rationale. |

whether the output uses the correct historical turns and ignores obsolete or irrelevant turns. *Scenario Faithfulness* measures whether the output follows the dependency pattern implied by the scenario label, such as combining constraints across turns, rolling back to an earlier reference, correcting a local mismatch, or changing preference. We provide the reasoning-level judge prompt used in our implementation below. The description-level and retrieval-level judge prompts follow the same scoring dimensions and output format, with only the evaluated field changed to the generated target descriptions or the top-5 retrieved videos, respectively. The final Intent Understanding Evaluation scores for each method are obtained by averaging the judge scores over all evaluated samples.

**Prompt 7: Reasoning-level Intent Understanding Judge**

**# Task Description**
You are evaluating the intermediate reasoning of a model for long-term composed video retrieval. The model is given a multi-turn retrieval interaction and must infer the user's final target video. Your task is to judge whether the model's reasoning correctly explains the evolving user intent before retrieval.
**## Inputs**
- Scenario label: `[[SCENARIO]]`
One of `Progressive Refinement`, `Rollback`, `Cross-turn Combination`, `Local Correction`, or `Preference Change`.
- Interaction history: `[[INTERACTION_HISTORY]]`
The ordered reference-video frames and user feedback from the preceding turns, together with the latest reference-video frames.
- Latest feedback: `[[LATEST_FEEDBACK]]`
- Target video: `[[TARGET_VIDEO_FRAMES_AND_CAPTION]]`
- Model reasoning under evaluation: `[[MODEL_REASONING]]`
The intermediate reasoning produced by the model before generating the final retrieval description.
**## Step 1: Infer the Ground-truth Intent**
Read the interaction history, latest feedback, scenario label, and target video. Determine which parts of the interaction history should be preserved, revised, revisited, combined, or ignored to form the final retrieval intent.
**## Step 2: Evaluate the Model Reasoning**
Evaluate the model reasoning along three dimensions:
- **Intent Consistency**: whether the reasoning correctly identifies the final target intent and matches the visible target video.
- **History Grounding**: whether the reasoning correctly uses the relevant turns, feedback, and visual evidence, while avoiding irrelevant or obsolete history.
- **Scenario Faithfulness**: whether the reasoning follows the required history-dependency pattern of the scenario, such as progressive refinement, rollback, cross-turn combination, local correction, or preference change.
**## Scoring Rule**

Assign each dimension a score from 0 to 10. A score of 10 means the reasoning is fully correct, visually grounded, and uses the appropriate history. A score of 7–9 means the reasoning is mostly correct with only minor missing details. A score of 4–6 means the reasoning captures part of the final intent but misses important historical or visual constraints. A score of 1–3 means the reasoning is largely inconsistent with the final intent or relies on wrong history. A score of 0 means the reasoning is invalid, unrelated, or unavailable.

**## Output Format**

Return a JSON object:

```
{
"intent_consistency": 0-10,
"history_grounding": 0-10,
"scenario_faithfulness": 0-10,
"rationale": "brief explanation of the assigned scores"
}
```

## C Additional Details of InAHR

This subsection provides additional theoretical and implementation details for the intent-conditioned facility-location objective used in Adaptive History Refocusing. We first establish its monotone submodularity and greedy approximation guarantee, and then describe the computational complexity and deterministic implementation used for evidence selection.

**Monotone submodularity.** Recall that each sampled historical frame $f_{i,k}$ is assigned a hierarchical relevance weight $\gamma_{i,k} = r_i p_{i,k}$, where $r_i$ denotes the relevance of interaction turn $i$ and $p_{i,k}$ measures the relevance of frame $k$ within that turn. Since $r_i$ is normalized across turns and $p_{i,k}$ is normalized within each turn, we have

$$r_i \geq 0, \qquad p_{i,k} \geq 0, \qquad \sum_{i=1}^{n} r_i = 1, \qquad \sum_{k=1}^{U} p_{i,k} = 1, \tag{24}$$

and therefore

$$\gamma_{i,k} \geq 0, \qquad \sum_{i=1}^{n} \sum_{k=1}^{U} \gamma_{i,k} = 1. \tag{25}$$

Importantly, the predicted relevant-turn set $\mathcal{I}_n$ is used only to construct the soft turn prior $\pi_i$ and does not hard-filter the frame set. Thus, the selection ground set remains

$$\mathcal{F}_n = \{f_{i,k} \mid i = 1, \ldots, n,\ k = 1, \ldots, U\}. \tag{26}$$

For two historical frames, the visual coverage kernel is

$$\kappa(f_{i,k}, f_{j,\ell}) = \frac{1 + \cos(\mathbf{z}_{i,k}, \mathbf{z}_{j,\ell})}{2}, \tag{27}$$

which satisfies $0 \leq \kappa(\cdot, \cdot) \leq 1$. For analysis, we write the objective in Eq. (3) as

$$\mathcal{R}_n(S) = \sum_{i=1}^{n} \sum_{k=1}^{U} \gamma_{i,k} \max_{s \in S} \kappa(f_{i,k}, s), \qquad S \subseteq \mathcal{F}_n, \tag{28}$$

with $\mathcal{R}_n(\emptyset) = 0$. Define the current coverage of frame $f_{i,k}$ by a selected set $S$ as

$$c_S(i,k) = \max_{s \in S} \kappa(f_{i,k}, s), \qquad c_\emptyset(i,k) = 0. \tag{29}$$

Because adding an element cannot decrease $c_S(i,k)$ and all $\gamma_{i,k}$ are non-negative, $\mathcal{R}_n(S)$ is monotone.

To establish submodularity, consider two sets $A \subseteq B \subseteq \mathcal{F}_n$ and a candidate $y \in \mathcal{F}_n \setminus B$. The marginal gain of adding $y$ to $S$ is

$$\begin{aligned} \Delta\mathcal{R}_n(y \mid S) &= \mathcal{R}_n(S \cup \{y\}) - \mathcal{R}_n(S) \\ &= \sum_{i=1}^{n} \sum_{k=1}^{U} \gamma_{i,k} \left[\max(c_S(i,k), \kappa(f_{i,k}, y)) - c_S(i,k)\right]. \end{aligned} \tag{30}$$

Since $A \subseteq B$ implies $c_A(i,k) \le c_B(i,k)$ for every frame, we obtain $\Delta\mathcal{R}_n(y \mid A) \ge \Delta\mathcal{R}_n(y \mid B)$. Hence, $\mathcal{R}_n$ satisfies the diminishing-return property and is a normalized monotone submodular function. Under the cardinality constraint $|S| \le M$, standard greedy maximization returns $\widehat{S}_n$ satisfying $R_n(\widehat{S}_n) \ge (1 - 1/e)R_n(S_n^*)$, where $S_n^*$ denotes the optimal solution to Eq. (3).

**Computational complexity.** Let $L = |\mathcal{F}_n| = nU$ denote the total number of sampled historical frames, and let $d$ be the dimension of the normalized frame embeddings $\mathbf{z}_{i,k}$. Once the frame and intent embeddings are available, computing the turn-level and frame-level relevance scores requires $\mathcal{O}(Ld)$ operations.

The pairwise visual coverage kernel over all historical frames can be precomputed in $\mathcal{O}(L^2 d)$ time and stored in $\mathcal{O}(L^2)$ memory. During greedy maximization, we maintain a coverage value $c_m(x)$ for each frame $x \in \mathcal{F}_n$ after $m$ frames have been selected. Given the cached similarity matrix, evaluating the marginal gain of one candidate requires $\mathcal{O}(L)$ time, and evaluating all remaining candidates requires $\mathcal{O}(L^2)$ time per greedy step. Selecting $M$ evidence frames therefore requires $\mathcal{O}(ML^2)$ additional time. The overall evidence-selection complexity is thus

$$\mathcal{O}\big(L^2 d + ML^2\big), \qquad L = nU, \tag{31}$$

with $\mathcal{O}(L^2)$ memory for the cached coverage kernel.

In our default setting, $U = 20$ and $M = 8$. Since LTCVR contains five-turn sessions, the largest evidence pool contains only $L = 100$ sampled frames, making the greedy refocusing step lightweight relative to MLLM inference.

**Deterministic greedy selection.** At greedy step $m$, given the previously selected set $S_{m-1}$, we select

$$x_m^* = \arg \max_{x \in \mathcal{F}_n \setminus S_{m-1}} \Delta\mathcal{R}_n(x \mid S_{m-1}). \tag{32}$$

When multiple candidates attain the same marginal gain, we first prefer the frame with the larger hierarchical relevance weight $\gamma_{i,k}$. If a tie still remains, we use a fixed ordering by turn index and then by frame index.

After selecting $x_m^*$, the cached coverage values are updated as

$$c_m(y) = \max(c_{m-1}(y), \kappa(y, x_m^*)), \qquad y \in \mathcal{F}_n. \tag{33}$$

This incremental update avoids repeatedly evaluating the full facility-location objective. After selection, the $M$ evidence frames are sorted by their original turn and frame indices before being re-injected into the MLLM, preserving the temporal organization of the interaction history.

### C.1 Prompt Details

InAHR uses structured MLLM prompts in Intent-State Induction and Target-State Reconstruction. The first prompt reasons over the sparse interaction overview to resolve the evolving user intent and identify the minimal set of relevant reference turns that provide useful visual evidence. The second prompt re-injects the selected frame-level evidence and generates multiple complementary retrieval-oriented descriptions of the same target state.

**Prompt 8: Intent-State Induction**

**# Task Description**
You are given an ordered multi-turn composed video retrieval history. Each turn contains a reference-video snapshot and the corresponding user feedback. In the final turn, the reference video is the latest retrieved result and the feedback represents the current request.
Your task is to: (1) reason about the current retrieval intent from the interaction history; (2) reconstruct a concise description of the desired target video; and (3) identify the minimal set of reference turns that provide necessary visual evidence for the target.
**## Reasoning**

Determine which information from the interaction history remains relevant to the current request and which information is no longer needed. Use the history only when it is necessary to resolve the current target.
**## Current Intent Reconstruction**
Write one concise and self-contained description of the desired target video.
Describe the final target state rather than the interaction process. Preserve historical constraints only when they remain consistent with the current request, and exclude information that has been superseded.
Focus on visible and retrieval-relevant semantics, such as subjects, objects, appearance, actions, scenes, spatial relations, viewpoint, and temporal cues.
**## Relevant-Turn Identification**
Select the smallest sufficient set of turns whose reference videos provide visual information required by the final target.
Do not select a turn solely because it is recent, related to the query, or useful for understanding the dialogue. Exclude redundant turns when their required visual information is already represented by another selected turn.
**## Output Format**
Return only:
{ `"reasoning"`: `"..."`,
`"current_intent"`: `"..."`,
`"relevant_turns"`: [...] }

**Prompt 9: Multi-Faceted Target-State Reconstruction**

**# Task Description**
You are given: (1) the latest user feedback, (2) the resolved current retrieval intent, and (3) selected visual evidence from the interaction history.
Each selected frame is accompanied by the feedback associated with its source turn. Your task is to reconstruct the desired target video and generate exactly $J$ complementary retrieval-oriented descriptions of the same target state.
**## Target Reconstruction**
Use the selected visual evidence to recover concrete visual details that are relevant to the current target. Interpret the evidence together with its associated feedback, while treating the Current Retrieval Intent and Latest User Feedback as the primary specification of the final target.
Apply the requested modification rather than simply describing the visual references. Use the Current Retrieval Intent to complement target information that is not fully expressed by the latest feedback or selected evidence.
**## Multi-Faceted Description Generation**
Generate exactly $J$ concise and self-contained descriptions of the desired target video.
All descriptions must represent the same target state, but should emphasize complementary semantic aspects, such as subjects and attributes, actions and interactions, scenes and spatial context, or fine-grained visual and temporal details.
Each description should preserve the essential target constraints and be independently usable as a retrieval query. The descriptions should differ in semantic emphasis rather than simple wording.
**## Constraints**
Use visual details only when they are consistent with the Current Retrieval Intent and Latest User Feedback.
Do not introduce unsupported information or retain historical details that are no longer relevant to the current target.
Describe only the final target video. Do not mention interaction history, turn indices, reference frames, selected evidence, or the reasoning process.
**## Output Format**
Return only:
{ `"descriptions"`: [ `"description 1"`, ..., `"description` $J$`"` ] }

# D Additional Details of Experimental Settings

## D.1 Baselines

This subsection provides additional details for the baselines used in Sec. 5. Since most existing methods are designed for single-turn composed retrieval, we adapt them to LTCVR with minimal task-specific changes. All methods are evaluated on the same fixed LTCVR sessions and a shared retrieval gallery of 3,160 unique videos, obtained by deduplicating the targets of all 4,000 sessions. During evaluation, the interaction history of each session is kept fixed rather than regenerated by individual methods. The target video and scenario label are used only for benchmark annotation and evaluation, and are not exposed to the retrieval models. For methods that output a textual retrieval query, we use the same OpenCLIP retrieval encoder and the same precomputed candidate-video index, so performance differences mainly reflect how each method interprets long-term interaction history.

**Embedding-based methods.** We evaluate six embedding-based baselines. **CLIP** (Radford et al., 2021) composes the reference-video representation and modification text in the CLIP embedding space. We use the open-source pretrained CLIP checkpoint without further fine-tuning. **BLIP** follows the same embedding-composition framework as CLIP, but replaces CLIP with a large COCO-fine-tuned BLIP image–text encoder. It extracts frame-level visual features from sampled video frames and text features from the modification text, without further fine-tuning on LTCVR. **LinCIR** (Gu et al., 2024) converts visual features into language-compatible pseudo tokens and combines them with the modification text for composed retrieval. We use the official released LinCIR weights and evaluate the model on LTCVR in a zero-shot manner. **CoVR** (Ventura et al., 2024a) learns to map a reference video and modification text into the target-video embedding space. **CoVR-2** (Ventura et al., 2024b) extends CoVR by using a BLIP-2 backbone and enriching target-video representations with both visual features and target captions. **CVRDM** (Thawakar et al., 2025) further models fine-grained video–text modifications for dense composed video retrieval. For CoVR, CoVR-2, and CVRDM, we follow the official training recipes and train each model on WebVid-CoVR (Ventura et al., 2024a) for 5 epochs. The resulting checkpoints are evaluated on LTCVR without additional fine-tuning.

To adapt these single-turn methods to a five-turn LTCVR session, we first compute a turn-level query embedding $\mathbf{q}_i$ from each reference-feedback pair $(v_i, u_i)$. We then aggregate turn-level queries with a deterministic recency-weighted average:

$$\mathbf{q}_{\text{full}} = \text{Norm}_2\left(\sum_{i=1}^{n} w_i \mathbf{q}_i\right), \quad w_i = \frac{i}{\sum_{j=1}^{n} j}. \tag{34}$$

This weighting gives larger weights to more recent turns while still preserving earlier evidence. For the *Last-turn* setting in Table 3, we use only $\mathbf{q}_n$; for the *Full history* setting, we use $\mathbf{q}_{\text{full}}$.

**MLLM-based methods.** We evaluate five MLLM-based baselines. **IP-CIR** (Li et al., 2025) is a training-free composed image retrieval method that imagines proxy images aligned with the query image and modification text to enhance retrieval. We adapt it to our video setting by generating three target-video proxy frames from the sampled reference frames and user feedback, and use their visual features to construct the retrieval query. **OSrCIR** (Tang et al., 2025) performs one-step reasoning over the visual reference and textual modification to generate a target description; we extend its input from a single turn to the chronological full history. **CoTMR** (Sun et al., 2025) uses chain-of-thought modification reasoning before producing the final retrieval query. We provide the full interaction history and ask it to summarize the final target state before retrieval. **AutoCIR** (Cheng et al., 2025) uses an automatic reasoning-and-evaluation pipeline to improve the generated retrieval query. **ImageScope** (Luo et al., 2025) decomposes the query into fine-grained visual requirements, verifies candidate images against these requirements, and reranks candidates through pairwise comparison. We adapt it to our video setting by applying the same verification and reranking procedure to candidate videos using sampled frames and the interaction history. For all MLLM-based baselines, we preserve the original reasoning style of each method and only modify the input packaging so that the model can observe all turns in a long-term video interaction.

### D.2 Implementation Details

The default MLLM backbone for MLLM-based methods is Qwen2.5-VL (Bai et al., 2025b), and the default retrieval encoder is OpenCLIP ViT-B/32 (Radford et al., 2021; Cherti et al., 2023). For the backbone comparison in Table 10, we replace the retrieval encoder with OpenCLIP ViT-L/14 and ViT-G/14 while keeping the MLLM reasoning pipeline unchanged.

**Video and text encoding.** For retrieval, each candidate video is represented by uniformly sampled 10 frames and encoded by the selected retrieval backbone. Frame embeddings are mean-pooled and $\ell_2$-normalized to obtain the video embedding. Textual outputs produced by MLLM-based methods are encoded by the text encoder of the same OpenCLIP backbone. Retrieval is performed by cosine similarity between the normalized text query and normalized candidate-video embeddings. We report Recall@1, Recall@5 and Recall@10 where applicable.

Table 8: Relevant-turn identification analysis on key history-dependent scenarios. **Turn-P:** precision of the predicted relevant-turn set. **Turn-R:** recall with respect to the ground-truth relevant-turn set. **Turn-F1:** the F1 score computed for each session and then averaged across all sessions. **Exact Match:** whether the predicted relevant-turn set exactly matches the ground-truth set derived from LTCVR construction metadata. **Random Turns**, **All Turns**, and **CLIP-Similarity** are diagnostic reference strategies for relevant-turn selection.

| Scenario | Method | Turn-P ↑ | Turn-R ↑ | Turn-F1 ↑ | Exact Match ↑ |
|---|---|---|---|---|---|
| *Combination* | Random Turns | 0.40 | 0.60 | 0.45 | 0.02 |
| | All Turns | 0.40 | 1.00 | 0.57 | 0.00 |
| | CLIP-Similarity | 0.51 | 0.69 | 0.57 | 0.03 |
| | InAHR | 0.72 | 0.71 | 0.71 | 0.43 |
| *Rollback* | Random Turns | 0.20 | 0.62 | 0.29 | 0.03 |
| | All Turns | 0.20 | 1.00 | 0.33 | 0.00 |
| | CLIP-Similarity | 0.25 | 0.69 | 0.36 | 0.02 |
| | InAHR | 0.98 | 1.00 | 0.99 | 0.96 |
| *Pref. Change* | Random Turns | 0.18 | 0.59 | 0.27 | 0.02 |
| | All Turns | 0.20 | 1.00 | 0.33 | 0.00 |
| | CLIP-Similarity | 0.26 | 0.64 | 0.35 | 0.05 |
| | InAHR | 0.62 | 0.79 | 0.67 | 0.48 |

**InAHR hyperparameters.** The sparse interaction overview is constructed using the middle frame of each reference video together with its corresponding user feedback. For Adaptive History Refocusing, we uniformly sample $U = 20$ frames from each reference video and retain all sampled frames in the historical evidence set $\mathcal{F}_n$. We use a frozen OpenCLIP ViT-B/32 image–text encoder to encode the induced current intent $c_n$ and historical frames for relevance estimation. We set the relevance temperature to $\tau = 0.08$ for both the turn-level relevance in Eq. (1) and the frame-level relevance in Eq. (2). The evidence budget is set to $M = 8$, and the resulting facility-location objective is optimized by greedy selection over the complete historical frame set. The selected evidence frames are ordered by their original turn and within-turn frame indices, paired with the feedback from their source turns, and then re-injected into the MLLM for Target-State Reconstruction. For Target-State Reconstruction, we generate $J = 3$ complementary retrieval-oriented descriptions, average their text embeddings, and $\ell_2$-normalize the resulting representation to form the final retrieval query.

## E Relevant-Turn Identification Analysis

To evaluate whether InAHR can identify the relevant turns that provide necessary visual evidence for the current target, we conduct a relevant-turn identification analysis on three representative history-dependent scenarios: *Combination*, *Rollback*, and *Preference Change*. For each session, we derive the ground-truth relevant-turn set from the LTCVR construction metadata. Specifically, the ground-truth set contains the two source turns whose visual constraints are jointly required in *Combination*, the anchor reference turn in *Rollback*, and the latest reference turn in *Preference*

*Change*, where earlier visual constraints become obsolete. We measure the agreement between the predicted and ground-truth sets using Turn-P, Turn-R, Turn-F1, and Exact Match.

Since existing retrieval baselines do not explicitly predict relevant turns, we compare InAHR with three diagnostic reference strategies: Random Turns, All Turns, and CLIP-Similarity. Random Turns uniformly samples a subset of interaction turns, All Turns retains the complete history, and CLIP-Similarity ranks turns according to the similarity between the latest feedback and turn-level visual representations, selecting turns before the largest score drop.

As shown in Table 8, InAHR achieves the most reliable relevant-turn identification across all three scenarios. In *Combination*, it provides the best precision–recall balance when recovering the two complementary visual sources. In *Rollback*, it almost perfectly identifies the required historical anchor, reaching 0.99 Turn-F1 and 0.96 Exact Match. In *Preference Change*, it more accurately identifies the currently active visual state while reducing the selection of obsolete historical turns. By contrast, All Turns achieves perfect recall by retaining the entire history, but suffers from low precision and zero Exact Match. These results show that InAHR provides meaningful turn-level guidance by selectively identifying target-relevant visual history rather than indiscriminately retaining all interaction turns.

# F Generalization Analysis

## F.1 Results on LTCVR using different MLLM backbones.

We evaluate the generality of InAHR across different MLLM backbones, including Qwen3-VL (Bai et al., 2025a), InternVL3-8B (Zhu et al., 2025), Gemini-3-Flash (Google DeepMind, 2025), and GPT-4o (OpenAI, 2024). We directly replace the MLLM used in Intent-State Induction and Target-State Reconstruction without any model-specific adaptation. As shown in Table 9, all alternative backbones improve over the default Qwen2.5-VL, with Avg. gains ranging from +1.9 to +5.3. The strongest MLLMs also yield the best retrieval performance, suggesting that InAHR can further benefit from improved MLLM reasoning capabilities while remaining broadly compatible across different model families.

Table 9: Results on LTCVR using different MLLM backbones. Δ denotes the Avg. difference relative to the default InAHR.

| MLLM Backbone | R@1 | R@5 | R@10 | Avg. | Δ |
|---|---|---|---|---|---|
| **Qwen2.5-VL** | **41.9** | **66.7** | **76.5** | **61.7** | 0.00 |
| *Impact of Different MLLMs* | | | | | |
| Qwen3-VL | 43.7 | 70.8 | 79.5 | 64.7 | +3.0 |
| InternVL3-8B | 42.9 | 69.0 | 78.9 | 63.6 | +1.9 |
| Gemini-3-Flash | 44.6 | 72.3 | 81.4 | 66.1 | +4.4 |
| GPT-4o | 46.0 | 73.0 | 82.1 | 67.0 | +5.3 |

## F.2 Results on LTCVR using different retrieval backbones.

To examine whether the gains of InAHR depend on a particular retrieval encoder, we evaluate representative MLLM-based methods with stronger OpenCLIP retrieval backbones while keeping the MLLM reasoning pipeline unchanged. As shown in Table 10, InAHR consistently achieves the best performance with both ViT-L/14 and ViT-G/14. With ViT-L/14, it improves the average score from the strongest baseline score of 59.5 to 63.9; with ViT-G/14, it further improves the average score from 62.7 to 66.7. These results indicate that the advantage of intent-driven adaptive history refocusing is complementary to stronger retrieval backbones rather than being tied to a specific encoder.

## F.3 Results on WebVid-CoVR test set.

We further evaluate InAHR on the WebVid-CoVR test set to examine its generalization to standard single-turn CVR. We compare against CoVR (Ventura et al., 2024a), CoVR-2 (Ventura et al., 2024b), CVRDM (Thawakar et al., 2025), and CoVR-R (Thawakar et al., 2026) under two settings: with and without task-specific fine-tuning on WebVid-CoVR. The baseline results in the latter setting are reported from Table A of Thawakar et al. (2026). Although InAHR is designed for long-term multi-turn retrieval, its three-stage pipeline can be directly applied to a single-turn query without structural modification. Given a reference video and modification text, Intent-State Induction resolves the target intent, Adaptive History Refocusing selects compact intent-relevant visual evidence from

Table 10: Comparison results on LTCVR using different retrieval backbones.

| Backbone | Model | R@1 | R@5 | R@10 | Avg. |
|---|---|---|---|---|---|
| ViT-L/14 | IP-CIR | 31.2 | 55.4 | 65.4 | 50.7 |
| | OSrCIR | 39.9 | 64.7 | 73.4 | 59.3 |
| | AutoCIR | 37.6 | 63.7 | 73.7 | 58.3 |
| | CoTMR | 40.7 | 64.6 | 73.1 | 59.5 |
| | ImageScope | 40.0 | 62.7 | 72.4 | 58.4 |
| | **Ours** | **44.1** | **69.3** | **78.2** | **63.9** |
| ViT-G/14 | IP-CIR | 35.3 | 59.4 | 68.9 | 54.5 |
| | OSrCIR | 44.2 | 67.6 | 76.3 | 62.7 |
| | AutoCIR | 43.3 | 67.1 | 74.9 | 61.8 |
| | CoTMR | 44.3 | 67.6 | 76.2 | 62.7 |
| | ImageScope | 42.4 | 65.4 | 74.5 | 60.8 |
| | **Ours** | **47.3** | **72.5** | **80.3** | **66.7** |

Table 11: Generalization analysis on WebVid-CoVR-Test set. † denotes results reported in Table A of Thawakar et al. (2026)

| Method | R@1 | R@5 | R@10 | R@50 |
|---|---|---|---|---|
| *Fine-tuned on WebVid-CoVR* | | | | |
| CoVR | 55.95 | 81.22 | 89.05 | 98.08 |
| CoVR-2 | 59.82 | 83.84 | 91.28 | 98.24 |
| CVRDM | **61.32** | **85.90** | **93.53** | **98.88** |
| *Without WebVid-CoVR Fine-tuning* | | | | |
| CoVR† | 45.46 | 70.46 | 79.54 | 93.27 |
| CoVR-2† | 45.66 | 71.71 | 81.30 | 94.80 |
| CVRDM† | 48.08 | 73.36 | 81.06 | 93.78 |
| CoVR-R† | 49.15 | 70.72 | 79.25 | 93.42 |
| **InAHR** | **49.80** | **75.66** | **84.97** | **96.09** |

Table 12: Robustness to longer interaction histories on extended LTCVR sessions. Each cell reports R@1 / R@5 (%). For each scenario, we extend the same 200 five-turn sessions to 7 and 9 turns while preserving the original target and history-dependency pattern. ALL is computed over 600 sessions at each history length.

| Turns | Method | ALL | Combination | Pref. Change | Rollback |
|---|---|---|---|---|---|
| 7 | CoTMR (Sun et al., 2025) | 25.00 / 43.83 | 31.50 / 55.00 | 15.50 / 29.00 | 28.00 / 47.50 |
| | ImageScope (Luo et al., 2025) | 26.00 / 45.33 | 38.00 / 59.00 | 8.00 / 23.50 | 32.00 / 53.50 |
| | **InAHR** | **33.50 / 55.33** | **40.00 / 61.50** | **25.00 / 46.50** | **35.50 / 58.00** |
| 9 | CoTMR (Sun et al., 2025) | 17.83 / 39.00 | 27.50 / 54.00 | 4.50 / 16.50 | 21.50 / 46.50 |
| | ImageScope (Luo et al., 2025) | 23.67 / 41.50 | 33.50 / 56.50 | 7.00 / 18.00 | 30.50 / 50.00 |
| | **InAHR** | **30.33 / 53.67** | **38.00 / 61.00** | **21.00 / 44.00** | **32.00 / 56.00** |

the reference video, and Target-State Reconstruction generates complementary retrieval-oriented descriptions, which are then encoded and aggregated into the final retrieval query.

As shown in Table 11, methods fine-tuned on WebVid-CoVR achieve stronger in-domain performance. More importantly, among methods evaluated without WebVid-CoVR-specific fine-tuning, InAHR achieves the highest performance across all four recall metrics. Compared with the strongest reported result at each cutoff, InAHR improves R@1, R@5, R@10, and R@50 by 0.65, 2.30, 3.67, and 1.29 points, respectively. These results indicate that the proposed intent-driven history refocusing framework can also generalize to standard single-turn composed video retrieval.

### F.4 Robustness to Longer Interaction Histories

To examine robustness beyond the default five-turn setting, we construct controlled 7- and 9-turn extensions of three history-dependent LTCVR scenarios: *Combination*, *Preference Change*, and *Rollback*. For each scenario, we sample 200 original five-turn sessions and extend the same sessions to both history lengths while preserving the original target and scenario-defining turns. Additional intermediate interactions are inserted and the final feedback is rewritten to maintain a coherent trajectory and the original dependency pattern, yielding 600 sessions for each history length. These scenarios provide complementary stress tests: relevant evidence becomes increasingly sparse in *Combination*, more temporally distant in *Rollback*, and potentially conflicting with obsolete constraints in *Preference Change*.

As shown in Table 12, InAHR consistently achieves the best performance across all three scenarios under both 7- and 9-turn settings. On the ALL split, it outperforms the strongest baseline by 7.50/10.00 R@1/R@5 points at 7 turns and 6.66/12.17 points at 9 turns. When extending the history from 7 to 9 turns, InAHR decreases from 33.50/55.33 to 30.33/53.67, while maintaining clear margins over

both CoTMR and ImageScope. The gains are particularly pronounced on *Preference Change*, where obsolete historical constraints become increasingly distracting, and remain consistent on *Combination* and *Rollback*, where useful evidence becomes sparse or temporally distant. These results show that InAHR remains effective as interaction histories grow longer and history-dependent evidence becomes harder to isolate.

## G Qualitative Analysis

Figure 6 compares InAHR with OSrCIR and CoTMR on two challenging scenarios. In the *Preference Change* case, the user shifts the target from a bear catching salmon in a stream to multiple bears playing in the snow. The baselines remain biased toward earlier evidence and generate descriptions involving salmon, streams, or rocky terrain, while InAHR correctly discards obsolete history and focuses on the new snowy-scene intent. In the *Rollback* case, the user explicitly returns to the second-turn video and asks for a saucier dish. OSrCIR and CoTMR fail to recover the correct historical anchor and drift toward a black plate or generic curry dishes, whereas InAHR preserves the white plate and food attributes from the referenced turn while incorporating the saucy modification. These examples show that InAHR can decide whether to ignore or revisit history and generate more faithful retrieval-oriented target descriptions.

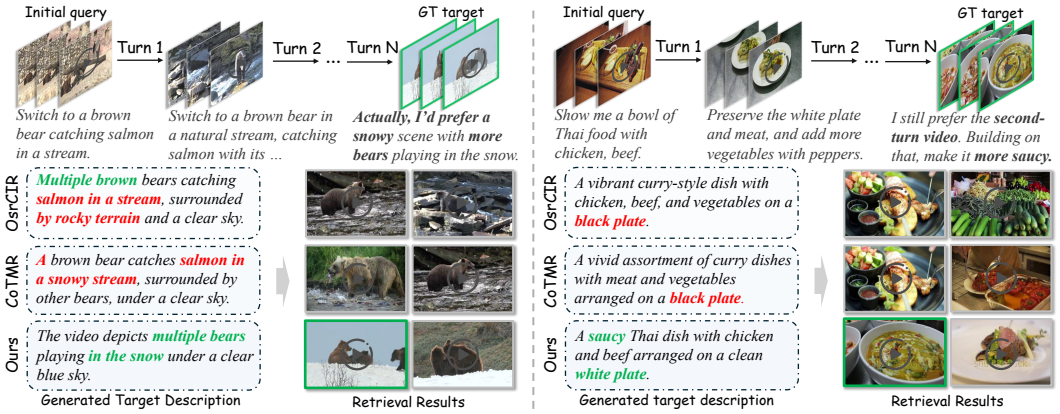



Figure 6: Qualitative results on the LTCVR benchmark: *Preference Change* on the left and *Rollback* on the right.

## H Limitations

Although LTCVR and InAHR provide a systematic framework for studying long-term CVR, several aspects could be further improved. First, GPT-4o is involved in both interaction generation and automatic quality screening. We mitigate potential model-specific tendencies by using different prompts and task roles for the two stages and by introducing independent human review, but some dependence on a single foundation model may still remain. Second, InAHR relies on MLLM reasoning during both Intent-State Induction and Target-State Reconstruction, which introduces additional inference overhead compared with lightweight embedding-based retrieval methods. Finally, the Intent Understanding Evaluation adopts an MLLM-as-a-judge protocol; while explicit prompts and scoring criteria are provided to improve transparency and consistency, future work could further strengthen the evaluation with multiple independent judges or broader human assessment.