This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Reasoning Text-to-Video Retrieval for Operating Room Clips via Action-Driven Digital Twins

Yiqing Shen[0000−0001−7866−3339], Hao Ding, and Mathias Unberath(✉)

Johns Hopkins University, Baltimore, MD, USA
{yshen92,unberath}@jhu.edu

Abstract. Text-to-video retrieval in operating rooms (OR) is an enabling technology for OR safety, as it allows stakeholders to retrieve and inspect recordings of specific events. However, because the most safety critical events may not follow the common structure, to unlock its full potential text-to-video retrieval must be able to handle implicit queries that require reasoning to identify the right videos (*e.g.*, "*the step right before clipping*"). However, existing methods rely on global embeddings that cannot reason over such queries. We propose OR^3, a text-to-video retrieval method that converts clips into action-driven digital twins (ActDTs), grouping concurrent subject-action-object triplets under non-overlapping temporal intervals. Moreover, rather than cross-modal matching through paired encoders, OR^3 performs imagination-based retrieval where an LLM generates hypothetical ActDTs from queries. This enables intra-modal matching via a single encoder trained with ActDT-tailored hard negatives. Finally, evidence-grounded refinement revises imagined ActDTs based on discrepancies with top candidates to capture procedure-specific patterns. We construct a benchmark from MM-OR with 276 implicit queries across four reasoning categories over 386 clips from robotic knee procedures. OR^3 achieves 57.6% R@1 and 77.3% R@5, outperforming the strongest baseline. These results demonstrate that OR^3 enables fine-grained discrimination between visually similar OR video clips through temporal action reasoning.

Keywords: Text-to-Video Retrieval · Reasoning Retrieval · Digital Twin · Operating Room Analysis · Large Language Models

1 Introduction

Operating rooms (ORs) can produce thousands of hours of video from room-mounted cameras, yet users typically need only short clips that last seconds to minutes [10, 16]. For example, surgical trainees may search for a specific dissection technique [26, 9], while quality teams may need to locate moments of miscommunication [8, 22]. Because entire OR videos typically span multiple hours, clip-level text-to-video retrieval is a natural formulation where the OR videos are first segmented by surgical phase and steps [27]. Unlike conventional text-to-video retrieval where queries explicitly describe visual content, OR retrieval must handle implicit queries demanding reasoning (*e.g.*, "*the step right before*

clipping" or "*the action that caused bleeding*") [22, 21]. Moreover, the OR video retrieval demands fine-grained discrimination because these clips from the same procedure appear visually near-identical, differing only in actions performed and entity interactions [18].

Existing text-to-video retrieval methods learn aligned text and video embeddings based on global similarity, but offer no mechanism to reason over implicit queries [28, 24]. Recent work on reasoning text-to-video retrieval addresses this by converting videos into digital twins (DTs) over which large language models (LLMs) can reason during retrieval [20, 22]. However, its object-centric DTs focus on per-frame entities and their visual attributes, which fail to discriminate between consecutive OR clips that share identical objects, but differ in actions and state transitions [20]. Moreover, its cross-modal matching between DT representation and query through paired encoders struggles to bridge the semantic gap between abstract implicit queries and concrete DT representation in OR contexts [3].

We propose OR^3 (operating room reasoning retrieval), a reasoning retrieval method for OR clips. OR^3 represents each clip as an action-driven digital twin (ActDT), a two-level structure that groups concurrent object interactions under non-overlapping temporal intervals. This action-centric representation captures state transitions and their dynamics, enabling discrimination between visually near-identical clips. Rather than matching queries against ActDTs through cross-modal paired encoders, OR^3 performs imagination-based retrieval by generating a hypothetical ActDT from the query itself. This casts retrieval as intra-modal matching and eliminates the semantic gap between abstract implicit queries and concrete ActDTs. When discrepancies emerge between imagined and actual ActDTs, evidence-grounded refinement revises the imagined ActDT to reflect procedure-specific patterns.

The contributions of this paper are three-fold. First, we introduce text-to-video reasoning retrieval for OR clips, a task requiring fine-grained discrimination between visually near-identical clips through implicit queries that demand reasoning. Correspondingly, we propose OR^3, which introduces ActDTs to encode temporal object interaction dynamics for OR clip retrieval. Second, we introduce imagination-based retrieval to eliminate cross-modal semantic gaps, and evidence-grounded refinement to adapt query interpretation to procedure-specific patterns. Third, we construct a benchmark for OR text-to-video reasoning retrieval.

2 Methods

Overview. Consider an OR video segmented into an ordered clip sequence $\mathcal{C} = \{C_1, C_2, \ldots, C_M\}$ based on surgical phases and steps [16, 5]. Given an implicit text query q that requires reasoning, our goal is to retrieve a ranked list of clips $\{(C_i, r_i)\}_{i=1}^{K}$, where $C_i \in \mathcal{C}$, $r_i \in [0, 1]$ denotes the relevance score, and K is the number of clips returned [20]. To address this, OR^3 (shown in Fig. 1) first converts each clip into an action-driven digital twin (ActDT) that encodes

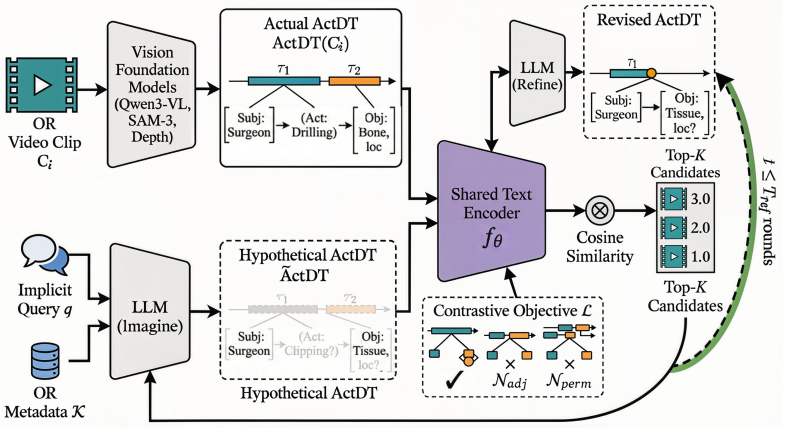

Fig. 1. Overview of OR^3. Each OR video clip C_i is converted into an actual ActDT through vision foundation models (Qwen3-VL, SAM-3, DepthAnythingV3), encoding subject-action-object triplets within non-overlapping temporal intervals. Given an implicit query q and OR metadata $\mathcal{K}$, an LLM imagines a hypothetical ActDT $\widehat{\text{ActDT}}_q$ that predicts the action primitives the target clip would contain. A shared text encoder f_θ, trained with a contrastive objective over adjacent-clip ($\mathcal{N}_{\text{adj}}$) and temporal-permutation ($\mathcal{N}_{\text{perm}}$) hard negatives [17], encodes both representations and ranks clips by cosine similarity. The top-K candidates are then passed to an LLM that performs evidence-grounded refinement, revising the hypothetical ActDT across up to T_{ref} rounds to align with procedure-specific patterns observed in the retrieved candidates.

concurrent object interactions as temporally structured action primitives [6] serialized in JSON. Imagination-based retrieval then employs an LLM to generate a hypothetical ActDT from the query itself, and matches this imagined representation against actual clip ActDTs using a single text encoder. Evidence-grounded refinement further revises the imagined ActDT based on discrepancies with top-ranked candidates, adapting to procedure-specific patterns. The refined ActDT is re-encoded and matched against clip ActDTs to produce the final ranked results.

Action-Driven Digital Twin. We represent each clip C_i as an action-driven digital twin (ActDT), a sequence of action primitives that encode the object and their interactions. Each ActDT is serialized as structured JSON text so that LLMs can process it directly without visual token compression of the raw video clips [21, 22, 19]. Formally, ActDT can be defined as $\text{ActDT}(C_i) = \big[\{\,\texttt{interval}: \tau_k,\ \texttt{actions}: \big[\{\texttt{subj}: \mathbf{s}_{k,n},\ \texttt{act}: v_{k,n},\ \texttt{obj}: \mathbf{o}_{k,n}\}\big]_{n=1}^{N_k}\}\big]_{k=1}^{T_i}$, where $T_i \in \mathbb{N}^+$ denotes the number of temporal intervals in clip C_i, and $N_k \in \mathbb{N}^+$ is the number of concurrent interactions within the k-th interval. Each interval $\tau_k = [t_k^{\text{s}}, t_k^{\text{e}}] \in \mathbb{R}^2$ is defined by its start t_k^{s} and end t_k^{e} timestamps, with intervals not overlapping and their union spanning the entire duration of C_i. The subject

$\mathbf{s}_{k,n} = \{\, \texttt{cat} : c,\ \texttt{attr} : d,\ \texttt{loc} : \ell \,\}$ and object $\mathbf{o}_{k,n}$ share the same format, each containing three fields. The category $c \in \mathcal{E}$ is drawn from an OR-specific vocabulary $\mathcal{E}$ of surgical roles, instruments, and equipment. The attribute d is a natural-language summary of properties such as posture or instrument state, generated by a VLM [1] from the RGB frames of C_i. The location $\ell \in \mathbb{R}^5$ encodes the bounding-box coordinates derived from the SAM-3 [2] segmentation masks and the depth prediction in the object centroids of DepthAnythingV3 [12]. The action $v_{k,n} \in \mathcal{V}$ records the interaction linking $\mathbf{s}_{k,n}$ to $\mathbf{o}_{k,n}$, also produced by the VLM and drawn from an OR-specific action vocabulary $\mathcal{V}$.

Imagination-Based Retrieval. OR^3 reformulates retrieval as intra-modal matching rather than cross-modal alignment. Given a query q, an LLM generates a hypothetical ActDT $\widehat{\text{ActDT}}_q = \text{LLM}_{\text{imagine}}(q, \mathcal{K})$, where $\mathcal{K}$ is procedural information drawn from OR video metadata. The LLM does not rephrase q but instead predicts the action primitives the target clip would contain. The hypothetical ActDT shares the same interval-action structure as actual ActDTs but omits ℓ and replaces VLM-generated attributes d with procedure-informed properties predicted by the LLM, as no visual frames are available at query time. Because both hypothetical and actual ActDTs are serialized as JSON text, we encode them with a single text encoder f_θ and compute

$$r_i = \text{sim}\Big(f_\theta(\widehat{\text{ActDT}}_q),\ f_\theta(\text{ActDT}(C_i)) \Big), \tag{1}$$

where $\text{sim}(\cdot,\cdot)$ denotes cosine similarity and r_i is the relevance score for clip C_i. We train f_θ with a contrastive objective combining two types of hard negatives tailored to ActDT. The first type is temporally adjacent clips, which share nearly identical entities but differ in actions, forcing the encoder to attend to action-level distinctions. The second type is temporal-permutation negatives, where a random permutation σ reorders the intervals of each ActDT to produce $\text{ActDT}^\sigma(C_i) = [\text{ActDT}(C_i)_{\sigma(k)}]_{k=1}^{T_i}$, preserving action primitives while disrupting their temporal arrangement. To bridge the structural gap between hypothetical and actual ActDTs, we apply field-dropout augmentation, independently masking each field with probability p_{drop} to produce $\widetilde{\text{ActDT}}(C_i) = \mathcal{M}(\text{ActDT}(C_i), p_{\text{drop}})$, where $\mathcal{M}$ replaces selected fields with a null token. This makes f_θ tolerant to the absence of the field in hypothetical ActDTs. The training loss [15, 4] over the positive pair (q, C^+) where C^+ is the ground-truth clip for q, adjacent negatives $\mathcal{N}_{\text{adj}}(C^+)$ drawn from temporally neighboring clips, and permutation negatives $\mathcal{N}_{\text{perm}}(C^+)$ constructed by interval reordering is

$$\mathcal{L} = -\sum_{(q,C^+)} \log \frac{\exp(r^+/\gamma)}{\exp(r^+/\gamma) + \sum_{C^- \in \mathcal{N}_{\text{adj}} \cup \mathcal{N}_{\text{perm}}} \exp(r^-/\gamma)}, \tag{2}$$

where γ is a temperature parameter, and $r^+ = \text{sim}(f_\theta(\widehat{\text{ActDT}}_q),\ f_\theta(\widetilde{\text{ActDT}}(C^+)))$, $r^- = \text{sim}(f_\theta(\widehat{\text{ActDT}}_q),\ f_\theta(\widetilde{\text{ActDT}}(C^-)))$. The top-$K$ candidates ranked by r_i proceed to evidence-grounded refinement.

Evidence-Grounded Refinement. The initial hypothetical ActDT relies on the LLM's generic knowledge, which may not capture the conventions of a specific procedure [20, 19]. OR^3 revises the imagined ActDT itself based on evidence from retrieved candidates. After the initial ranking produces the top candidates K with actual ActDTs $\{\mathrm{ActDT}(C_{i_k})\}_{k=1}^{K}$, the LLM receives the original query q, the current hypothetical $\widehat{\mathrm{ActDT}}_q$, and the candidate ActDTs, then identifies discrepancies between the imagined and observed action primitives. Afterwards, the LLM generates a revised hypothetical, namely $\widehat{\mathrm{ActDT}}_q^{(t+1)} = \mathrm{LLM}_{\mathrm{refine}}\left(q,\ \widehat{\mathrm{ActDT}}_q^{(t)},\ \{\mathrm{ActDT}(C_{i_k})\}_{k=1}^{K}\right)$, where t indexes the refinement round and $\widehat{\mathrm{ActDT}}_q^{(0)} = \widehat{\mathrm{ActDT}}_q$ is the initial imagined ActDT. The revised hypothetical ActDT is re-encoded by f_θ, and relevance scores r_i are recomputed over $\mathcal{C}$ to produce an updated ranking. This process iterates for at most T_{ref} rounds or terminates early when the top-ranked clip remains unchanged.

Benchmark. We construct an OR reasoning retrieval benchmark from the MM-OR dataset [16] to evaluate clip-level retrieval under implicit queries. Each of the 17 full-length recordings is segmented into non-overlapping clips according to the annotated surgical phases and procedural steps, yielding 386 clips with durations ranging from 20 seconds to 10 minutes. Two annotators with surgical domain knowledge independently come up implicit queries in four reasoning categories: temporal (*e.g.*, "*the step right before bone sawing*"), causal (*e.g.*, "*the action that triggered a sterility breach*"), procedural (*e.g.*, "*the first use of the drill after robot calibration*"), and role-based (*e.g.*, "*what the nurse prepared while the surgeon was sawing*"). Each query is mapped to one or more ground-truth clips and cross-verified by the other annotator, with disagreements resolved through discussion to ensure unambiguous correspondence. The final benchmark contains 276 implicit queries over 386 clips, split into training (162 queries, 218 clips), validation (48 queries, 72 clips), and test (66 queries, 96 clips) following the original MM-OR data partitions.

3 Experiments

Implementation Details. We construct ActDTs using Qwen3-VL-8B [1] as the VLM for generating action descriptions and entity attributes, SAM-3 [2] for instance segmentation masks, and DepthAnythingV3 [12] for depth estimation at object centroids. For imagination-based retrieval and evidence-grounded refinement, we employ Gemini-3-Pro as the LLM backbone, with top-K=10 candidates forwarded to refinement and a maximum of T_{ref}=3 refinement rounds. The text encoder f_θ is initialized from the BERT-base model [7] and trained with AdamW and learning rate 2×10^{-5}, weight decay 0.01 for 80 epochs on a 8 NVIDIA GeForce RTX 4090 GPUs, with temperature γ=0.07, field-dropout probability p_{drop}=0.3, and two temporally adjacent clips on each side as hard negatives. We report Recall at rank K (R@K) for $K \in \{1, 5, 10\}$, measuring

Table 1. Retrieval performance on the OR reasoning retrieval benchmark. Methods are grouped into embedding-based retrieval, LLM approaches, reasoning retrieval, and our OR^3. Bold and underline mark the best and second-best results.

Method	Temporal			Causal			Procedural			Role-based			Overall		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
CLIP4Clip [13]	$10.5_{\pm 2.4}$	$26.3_{\pm 3.4}$	$36.8_{\pm 3.8}$	$5.3_{\pm 1.9}$	$15.8_{\pm 3.1}$	$23.7_{\pm 3.6}$	$8.8_{\pm 2.1}$	$20.6_{\pm 3.0}$	$29.4_{\pm 3.4}$	$11.8_{\pm 2.6}$	$26.5_{\pm 3.6}$	$38.2_{\pm 3.9}$	$9.1_{\pm 1.8}$	$22.7_{\pm 2.7}$	$31.8_{\pm 3.0}$
X-CLIP [14]	$14.2_{\pm 2.7}$	$31.6_{\pm 3.6}$	$42.1_{\pm 3.8}$	$7.9_{\pm 2.3}$	$18.4_{\pm 3.3}$	$28.9_{\pm 3.8}$	$11.8_{\pm 2.4}$	$23.5_{\pm 3.2}$	$32.4_{\pm 3.5}$	$14.7_{\pm 2.9}$	$32.4_{\pm 3.8}$	$44.1_{\pm 4.0}$	$12.1_{\pm 2.1}$	$25.8_{\pm 2.8}$	$36.4_{\pm 3.1}$
InternVideo2 [25]	$17.6_{\pm 3.0}$	$36.8_{\pm 3.8}$	$47.4_{\pm 3.9}$	$10.5_{\pm 2.6}$	$23.7_{\pm 3.6}$	$34.2_{\pm 4.0}$	$14.7_{\pm 2.6}$	$32.4_{\pm 3.5}$	$41.2_{\pm 3.7}$	$17.6_{\pm 3.1}$	$38.2_{\pm 3.9}$	$52.9_{\pm 4.0}$	$15.2_{\pm 2.3}$	$33.3_{\pm 3.0}$	$43.9_{\pm 3.2}$
TeachCLIP [23]	$12.3_{\pm 2.5}$	$28.1_{\pm 3.5}$	$40.4_{\pm 3.8}$	$5.3_{\pm 1.9}$	$15.8_{\pm 3.1}$	$26.3_{\pm 3.7}$	$11.8_{\pm 2.4}$	$23.5_{\pm 3.2}$	$32.4_{\pm 3.5}$	$12.9_{\pm 2.7}$	$29.4_{\pm 3.7}$	$41.2_{\pm 4.0}$	$10.6_{\pm 2.0}$	$24.2_{\pm 2.7}$	$34.8_{\pm 3.1}$
Qwen3-VL-8B [1]	$21.1_{\pm 3.2}$	$42.1_{\pm 3.8}$	$52.6_{\pm 3.9}$	$15.8_{\pm 3.1}$	$28.9_{\pm 3.8}$	$39.5_{\pm 4.1}$	$17.6_{\pm 2.8}$	$35.3_{\pm 3.6}$	$47.1_{\pm 3.7}$	$23.5_{\pm 3.4}$	$44.1_{\pm 4.0}$	$55.9_{\pm 4.0}$	$19.7_{\pm 2.5}$	$37.9_{\pm 3.1}$	$48.5_{\pm 3.2}$
Gemini3	$15.8_{\pm 2.8}$	$33.3_{\pm 3.7}$	$45.6_{\pm 3.9}$	$10.5_{\pm 2.6}$	$21.1_{\pm 3.5}$	$31.6_{\pm 3.9}$	$11.8_{\pm 2.4}$	$29.4_{\pm 3.4}$	$38.2_{\pm 3.7}$	$17.6_{\pm 3.1}$	$35.3_{\pm 3.9}$	$47.1_{\pm 4.0}$	$13.6_{\pm 2.2}$	$30.3_{\pm 2.9}$	$40.9_{\pm 3.2}$
ReasonT2V [20]	$\underline{26.3}_{\pm 3.4}$	$\underline{47.4}_{\pm 3.9}$	$\underline{57.9}_{\pm 3.8}$	$\underline{15.8}_{\pm 3.1}$	$\underline{31.6}_{\pm 3.9}$	$\underline{42.1}_{\pm 4.2}$	$\underline{23.5}_{\pm 3.2}$	$\underline{41.2}_{\pm 3.7}$	$\underline{50.0}_{\pm 3.7}$	$\underline{29.4}_{\pm 3.7}$	$\underline{47.1}_{\pm 4.0}$	$\underline{61.8}_{\pm 3.9}$	$\underline{24.2}_{\pm 2.8}$	$\underline{42.4}_{\pm 3.2}$	$\underline{53.0}_{\pm 3.2}$
OR^3 (Ours)	$\mathbf{63.2}_{\pm 3.8}$	$\mathbf{82.5}_{\pm 3.0}$	$\mathbf{89.5}_{\pm 2.4}$	$\mathbf{42.1}_{\pm 4.2}$	$\mathbf{63.2}_{\pm 4.1}$	$\mathbf{73.7}_{\pm 3.7}$	$\mathbf{58.8}_{\pm 3.7}$	$\mathbf{76.5}_{\pm 3.2}$	$\mathbf{85.3}_{\pm 2.6}$	$\mathbf{67.6}_{\pm 3.8}$	$\mathbf{85.3}_{\pm 2.9}$	$\mathbf{91.2}_{\pm 2.3}$	$\mathbf{57.6}_{\pm 3.2}$	$\mathbf{77.3}_{\pm 2.7}$	$\mathbf{84.8}_{\pm 2.3}$

the fraction of queries for which at least one ground-truth clip appears within the top-K results. All metrics are averaged over five runs with different random seeds, and we report mean $\pm$ standard deviation.

Retrieval Performance. Table 1 compares OR^3 with embedding-based retrieval [13, 14, 23], LLM-based approaches [1, 25], and ReasonT2V [20] on our OR reasoning retrieval benchmark. OR^3 achieves 57.6% R@1, outperforming ReasonT2V [20] by 33.4 percentage points and confirming that ActDT captures action-level differences while imagination-based retrieval bridges the semantic gap between abstract implicit queries and digital twin representations. Embedding-based methods [13, 14, 23] remain below 16% R@1, as global similarity matching cannot reason over implicit queries nor discriminate between visually near-identical OR clips differing only in actions. LLM-based approaches improve through partial language reasoning, yet visual token compression [11, 20] discards fine-grained action and temporal structure needed for OR clip retrieval. ReasonT2V [20] reaches 24.2% R@1 through digital twin representations, but its object-centric design fails to capture action-level state transitions distinguishing consecutive surgical clips. Across query categories, causal queries show

Fig. 2. Qualitative comparison on the OR reasoning retrieval benchmark across four query categories. Red borders indicate incorrect retrievals where methods return clips that do not match the query, while green borders denote correct retrievals by OR^3.

Table 2. Ablation study on the OR reasoning retrieval benchmark (R@1). ActDT: action-driven digital twin (replaced by object-centric digital twin of ReasonT2V when absent). Imag.: imagination-based retrieval (replaced by cross-modal paired-encoder matching when absent). Ref.: evidence-grounded refinement. Adj.: adjacent-clip hard negatives. Perm.: temporal-permutation negatives. FD: field-dropout augmentation.

ActDT	Imag.	Ref.	Adj.	Perm.	FD	Temp.	Caus.	Proc.	Role	Overall
✗	✓	✓	✓	✓	✓	47.4$_{\pm 3.9}$	28.9$_{\pm 3.8}$	41.2$_{\pm 3.7}$	52.9$_{\pm 4.0}$	42.4$_{\pm 3.2}$
✓	✗	✓	✓	✓	✓	50.9$_{\pm 3.9}$	31.6$_{\pm 3.9}$	44.1$_{\pm 3.7}$	55.9$_{\pm 4.0}$	45.5$_{\pm 3.2}$
✓	✓	✗	✓	✓	✓	54.4$_{\pm 3.9}$	34.2$_{\pm 4.0}$	50.0$_{\pm 3.7}$	61.8$_{\pm 3.9}$	50.0$_{\pm 3.2}$
✓	✓	✓	✗	✓	✓	57.9$_{\pm 3.9}$	36.8$_{\pm 4.1}$	50.0$_{\pm 3.7}$	61.8$_{\pm 3.9}$	51.5$_{\pm 3.1}$
✓	✓	✓	✓	✗	✓	59.6$_{\pm 3.8}$	39.5$_{\pm 4.1}$	55.9$_{\pm 3.7}$	58.8$_{\pm 3.9}$	53.0$_{\pm 3.1}$
✓	✓	✓	✓	✓	✗	56.1$_{\pm 3.9}$	36.8$_{\pm 4.1}$	55.9$_{\pm 3.7}$	61.8$_{\pm 3.9}$	52.3$_{\pm 3.2}$
✓	✓	✓	✓	✓	✓	**63.2**$_{\pm 3.8}$	**42.1**$_{\pm 4.2}$	**58.8**$_{\pm 3.7}$	**67.6**$_{\pm 3.8}$	**57.6**$_{\pm 3.2}$

most challenging, with OR^3 achieving 42.1% versus 15.8% for ReasonT2V, as resolving cause-effect chains requires both temporal ordering and action-level reasoning. Temporal and procedural queries show the largest absolute gains of 36.9 and 35.3 R@1 points over ReasonT2V, reflecting ActDT's interval-structured action primitives for queries referencing sequential ordering or workflow position. Role-based queries yield the highest performance at 67.6% R@1, since identifying agent-action pairs aligns naturally with the subject-action-object triplets in each ActDT interval.

Fig. 2 shows qualitative results in all four query categories. Embedding-based methods [13, 14, 23] and video-LLMs consistently retrieve visually similar but incorrect clips, as consecutive OR phases share nearly identical room layouts and personnel configurations. ReasonT2V [20] narrows the gap through digital

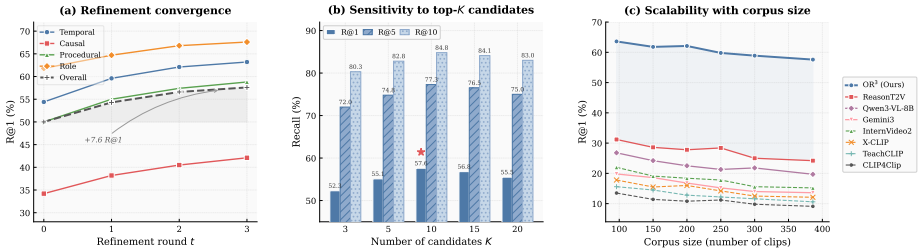

Fig. 3. Additional analyses of OR^3. (a) R@1 per query category as a function of evidence-grounded refinement round t, where t=0 corresponds to the initial imagined ActDT before any refinement. (b) Sensitivity to the number of top-K candidates forwarded from imagination-based retrieval to refinement, with the star marking the selected operating point (K=10). (c) R@1 of all compared methods as the retrieval corpus grows from 96 to 386 clips; the shaded region indicates the margin of OR^3 over the strongest baseline.

twin reasoning yet still confuses clips whose objects overlap but whose actions differ. OR^3 retrieves the correct clip in each case, confirming that action-driven digital twins and imagination-based matching together resolve the fine-grained distinctions that other approaches miss.

Ablation Study. Table 2 isolates each component by removing one at a time and reporting R@1 across query categories. Replacing ActDT with ReasonT2V's object-centric digital twin representation [20] causes the largest drop of 15.2% overall, with procedural queries falling 17.6% and temporal queries 15.8%, confirming that interval-structured action primitives drive fine-grained clip discrimination. This aligns with the large performance gap in Table 1 between OR^3 and ReasonT2V [20] on temporal and procedural queries, where sequential action ordering determines relevance. Eliminating imagination-based retrieval reduces overall R@1 by 12.1%, validating that matching of the same-space eliminates the semantic gap between abstract queries and concrete ActDT representations. Without evidence-grounded refinement, overall R@1 drops 7.6%, with causal queries declining 7.9%, as procedure-specific conventions matter most for resolving cause-effect chains. Among training strategies, adjacent-clip hard negatives contribute 6.1% by forcing the encoder to distinguish action-level differences between temporally neighboring clips. Field-dropout augmentation adds 5.3% by tolerating missing visual attributes in hypothetical ActDTs, while temporal-permutation negatives yield 4.6% by preserving interval ordering. All components produce additive gains exceeding their individual contributions, indicating mutual reinforcement. Fig. 3 presents additional analyses of refinement convergence, candidate pool sensitivity, and corpus scalability. Evidence-grounded refinement yields monotonic R@1 gains across all query categories, with causal queries benefiting the most (34.2% $\rightarrow$ 42.1% over three rounds) and diminishing returns after t=2 supporting the early-stopping criterion. Performance peaks at K=10 candidates, as smaller pools risk excluding the ground-truth clip while larger pools dilute the evidence set presented to the LLM. As the corpus grows from 96 to 386 clips, OR^3 loses only 6.0 R@1 points compared to 7.0 for ReasonT2V and up to half the initial accuracy for embedding-based methods, confirming that ActDT-based discrimination scales more gracefully with increasing visual ambiguity.

4 Conclusion

We present OR^3, a reasoning retrieval method that addresses the gap between how users search for OR clips and how the existing text-to-video retrieval method operates. By shifting from object-centric to action-driven digital twins, our approach captures the temporal dynamics and state transitions that distinguish visually similar OR clips. Imagination-based retrieval further reframes cross-modal matching as an intra-modal problem, while evidence-grounded refinement adapts query interpretation to the conventions of individual procedures. Future work can explore evaluation of other procedure types and larger clip collections.

Extending the framework to support multi-turn queries and real-time retrieval during live procedures can be another promising future direction.

Acknowledgments. Y. Shen was supported in part by the JHU Amazon Initiative for Artificial Intelligence (AI2AI) fellowship program.

Disclosure of Interests. The authors have no competing interests in the paper.

References

1. Shuai Bai et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.
2. Nicolas Carion et al. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025.
3. Zhanbin Che and Huaili Guo. Cross-modal video retrieval model based on video-text dual alignment. *International Journal of Advanced Computer Science & Applications*, 15(2), 2024.
4. Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020.
5. Tobias Czempiel et al. Tecno: Surgical phase recognition with multi-stage temporal convolutional networks. In *International conference on medical image computing and computer-assisted intervention*, pages 343–352. Springer, 2020.
6. Eadom Dessalene, Michael Maynord, Cornelia Fermüller, and Yiannis Aloimonos. Therbligs in action: Video understanding through motion primitives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10618–10626, 2023.
7. Jacob Devlin et al. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.
8. Alexander AJ Grüter et al. Video-based tools for surgical quality assessment of technical skills in laparoscopic procedures: a systematic review. *Surgical endoscopy*, 37(6):4279–4297, 2023.
9. Grant M Henning et al. A step toward modernization of urologic training: Incorporation of a novel surgical intelligence platform for robotic prostatectomy video review. *Journal of endourology*, 39(11):1204–1210, 2025.
10. Marc Levin et al. Surgical data recording in the operating room: a systematic review of modalities and metrics. *British Journal of Surgery*, 108(6):613–621, 2021.
11. Jianjian Li et al. Fcot-vl: Advancing text-oriented large vision-language models with efficient visual token compression. *arXiv preprint arXiv:2502.18512*, 2025.
12. Haotong Lin et al. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.
13. Huaishao Luo et al. Clip4clip: An empirical study of clip for end to end video clip retrieval. *arXiv preprint arXiv:2104.08860*, 2021.
14. Yiwei Ma et al. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In *Proceedings of the 30th ACM international conference on multimedia*, pages 638–647, 2022.

15. Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.
16. Ege Özsoy et al. Mm-or: A large multimodal operating room dataset for semantic understanding of high-intensity surgical environments. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19378–19389, 2025.
17. Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. *arXiv preprint arXiv:2010.04592*, 2020.
18. Saurav Sharma et al. fine-clip: Enhancing zero-shot fine-grained surgical action recognition with vision-language models. *arXiv preprint arXiv:2503.19670*, 2025.
19. Yiqing Shen et al. Online reasoning video segmentation with just-in-time digital twins. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 24698–24706, 2025.
20. Yiqing Shen et al. Reasoning text-to-video retrieval via digital twin video representations and large language models. *arXiv preprint arXiv:2511.12371*, 2025.
21. Yiqing Shen et al. Temporally-constrained video reasoning segmentation and automated benchmark construction. In *International Workshop on Foundation Models for General Medical AI*, pages 150–158. Springer, 2025.
22. Yiqing Shen, Chenjia Li, Bohan Liu, Cheng-Yi Li, Tito Porras, and Mathias Unberath. Operating room workflow analysis via reasoning segmentation over digital twins. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 415–424. Springer, 2025.
23. Kaibin Tian, Ruixiang Zhao, Hu Hu, Runquan Xie, Fengzong Lian, Zhanhui Kang, and Xirong Li. Teachclip: Multi-grained teaching for efficient text-to-video retrieval. *arXiv preprint arXiv:2308.01217*, 2023.
24. Xiaohan Wang et al. T2vlad: global-local sequence alignment for text-video retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5079–5088, 2021.
25. Yi Wang et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *European conference on computer vision*, pages 396–416. Springer, 2024.
26. Samy Cheikh Youssef et al. Learning surgical skills through video-based education: a systematic review. *Surgical Innovation*, 30(2):220–238, 2023.
27. Tong Yu et al. Live laparoscopic video retrieval with compressed uncertainty. *Medical Image Analysis*, 88:102866, 2023.
28. Haonan Zhang et al. Text-video retrieval with global-local semantic consistent learning. *IEEE Transactions on Image Processing*, 2025.