LOOKING ONE STEP AHEAD: AN ANTICIPATION-GUIDED AGENT FOR OBJECT STATE READOUT IN VIDEOS

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding the progress of an activity requires tracking how the objects involved change over time. Object state readout provides this information by describing a target object's states and transitions across video frames. Although multimodal large language models (MLLMs) generate detailed visual descriptions, these descriptions may leave the attributes needed to distinguish successive states insufficiently specified. In this paper, we propose look-ahead state readout (LASER), an anticipation-guided agent for object state readout that uses possible next-step changes to determine which attributes to inspect in the present. Given temporally ordered frames and a target object, LASER converts anticipated changes into targeted visual questions and inspects the current frame to establish attribute values. It then checks state readouts and forecasts against independent observations of subsequent frames, refining a state record used to generate the final description. Reusable textual observation skills guide this process and are refined by tracing discrepancies with reference captions to the responsible instructions. Candidate revisions are selected on separate videos based on improved caption quality without increased contradictions with reference facts. Model parameters remain frozen throughout adaptation and inference, and the revised skills remain fixed during evaluation. Experiment results and further analysis on OSCaR and PLM RCap show consistent improvements of LASER over strong baselines and existing approaches, which demonstrates the effectiveness of our approach.[1]

1 INTRODUCTION

Object state readout describes how a specified object's condition changes across visual observations, providing a language account of its evolving state (Nguyen et al., 2024; Cho et al., 2025). Whereas image captioning describes the content of a single image, state readout tracks object attributes across time to distinguish successive states and their transitions. This makes state readout an important task for procedural planning and robotic interaction, where understanding how actions change objects supports subsequent decisions (Nguyen et al., 2024). The central challenge is deciding which attributes reveal a meaningful change, rather than merely recognizing the object or naming its activity.

Multimodal large language models (MLLMs) provide a foundation for describing visual observations in natural language. Recent studies improve their video descriptions through detailed spatial and temporal grounding (Xue et al., 2025; Cho et al., 2025; Kazakos et al., 2025; Athar et al., 2025), while temporal benchmarks evaluate their understanding of motion and event order (Liu et al., 2024; Cai et al., 2024). Related agent studies explore anticipation-guided observation (Yang et al., 2025) and feedback-driven refinement of reusable instructions (Zhao et al., 2024; Zhang et al., 2026). These capabilities support object state readout, but accurate grounding and temporal ordering alone do not specify which attributes a state description needs to retain for subsequent comparison. A description may correctly identify an object and its surrounding activity while leaving the property that changes insufficiently specified, weakening the basis for describing its transition across images. Selecting these attributes requires considering which properties distinguish the object's possible subsequent states. For example, for a drawer, opening, closing, and remaining unchanged make its current extension

[1] Code is available at `https://anonymous.4open.science/r/LASER-EC56`.

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

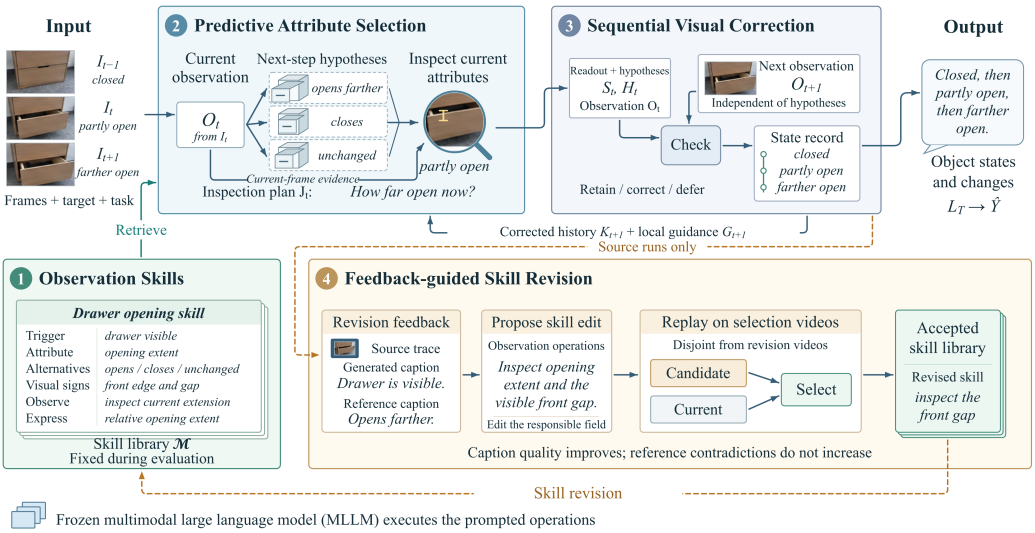

Figure 1: Overview of LASER. (1) Observation skills guide (2) predictive attribute selection, which anticipates possible changes and inspects current attributes. (3) Sequential visual correction checks the readout and forecast against independent observations, updates the state record, and guides subsequent inspection. The final record supports the output description. (4) Feedback-guided skill revision uses source-video traces and reference captions to propose edits, then compares candidate and current libraries on separate videos, selecting edits that improve caption quality without increasing reference contradictions. Solid arrows indicate information flow; orange dashed arrows indicate skill adaptation. All operations use a frozen MLLM, and the skill library remains fixed during evaluation.

relevant, even when that detail receives little emphasis in a general scene description. Anticipated changes thus provide a basis for deciding what to inspect, while visual evidence establishes the current attribute value. The next image provides further evidence for assessing those changes and identifying attributes that require closer inspection. This connection between possible changes and observable properties motivates using anticipation for attribute selection, subsequent evidence to refine state descriptions, and feedback on those descriptions to improve reusable observation instructions.

In this paper, we propose look-ahead state readout (LASER), an anticipation-guided agent that guides an MLLM to inspect an object's current state by anticipating possible next-step changes. Given a sequence of video frames and a target object, LASER uses observation skills, expressed as reusable textual instructions, to anticipate alternative changes and identify the attributes that distinguish them. Predictive attribute selection turns these attributes into visual questions and examines the current image to establish their values. Sequential visual correction then checks the resulting descriptions and anticipated changes against subsequent images, maintaining an evidence-supported state record for the final description. To improve the observation skills, feedback-guided skill revision traces discrepancies between generated and reference captions to the instructions responsible for observation and description. It proposes focused edits and retains those that improve caption quality on separate source videos without increasing contradictions with the references. This process adapts observation instructions while keeping all model parameters fixed, and the resulting skills remain fixed during evaluation. We evaluate LASER on English benchmarks across various MLLMs, where our approach outperforms strong baselines and existing studies, demonstrating the effectiveness of our approach.

2 The Approach

We propose look-ahead state readout (LASER), an anticipation-guided agent that considers possible next-step changes to guide observation of an object's current state. Figure 1 presents its overall architecture, comprising four components: *observation skills*, *predictive attribute selection*, *sequential visual correction*, and *feedback-guided skill revision*. LASER organizes observation and reasoning operations and maintains the state record and local guidance; a frozen MLLM executes the visual inspection and language reasoning within these operations. Specifically, given T temporally ordered video frames $V = (I_1, \ldots, I_T)$, a target prompt P identifying the object, and a task instruction

$\mathcal{T}$ specifying the description scope, LASER processes the observations in sequence. For each observation I_t, with $t = 1, \ldots, T$, LASER retrieves *observation skills* from the skill library $\mathcal{M}$ to guide which attributes of the target object to inspect and how to describe them. *Predictive attribute selection* uses these skills to propose possible next-step changes, converts their distinguishing attributes into visual questions, and checks the current image to establish the attribute values. When the next observation arrives, *sequential visual correction* checks earlier descriptions and anticipated changes against the new evidence. It updates a state record L_t of observed attributes, changes, and supporting evidence, and local guidance G_t containing unresolved questions for subsequent inspection. After the final observation, LASER composes the description $\widehat{Y}$ from the final state record $L = L_T$. During adaptation, *feedback-guided skill revision* uses discrepancies in these readout sequences and reference captions to propose instruction edits, then selects edits through replay on separate videos. LASER returns the description $\widehat{Y}$ and its supporting state record L:

$$(\widehat{Y}, L) = \mathrm{LASER}(V, P, \mathcal{T}; \mathcal{M}) \tag{1}$$

The following subsections describe observation skills, predictive attribute selection, sequential visual correction, and feedback-guided skill revision, respectively.

2.1 Observation Skills

An observation skill is a structured textual instruction that tells the agent what to inspect about an object and how to describe the visual evidence. Each skill contains six fields: a trigger specifying when the instruction applies, an attribute scope identifying what to inspect, alternative changes suggesting possible next states, visual signs distinguishing those alternatives, observation operations specifying how to check the current image, and expression constraints governing the resulting description. We organize these observation instructions into a skill library $\mathcal{M}$. For each observation, the MLLM selects relevant skills to guide inspection (see Section 2.2 for details). Feedback-guided skill revision updates these instructions using caption feedback (Section 2.4).

2.2 Predictive Attribute Selection

Predictive attribute selection uses possible next-step changes to determine which attributes matter for describing the object's current state. For each image I_t, the MLLM first observes the target using the task instruction $\mathcal{T}$ and the target information available up to that observation, denoted by $P_{\leq t}$. It produces an observation description O_t containing visible attributes and their values, supporting image regions, and any uncertainty. This initial observation receives no predicted changes, so its description is grounded in the image. For $t > 1$, LASER checks earlier descriptions against the new observation before selecting skills. The resulting K_t contains the earlier states and observed changes after evidence-supported corrections, while local guidance G_t records questions that still require inspection. Both are empty for the first observation. The MLLM selects skills whose triggers and attribute scopes match O_t, $\mathcal{T}$, K_t, and G_t, forming $\mathcal{M}_t$. When no skill matches, the MLLM follows the task instruction and directly inspects observable attributes of the target object.

Using the selected skills, the MLLM anticipates how the target might change by the next observation. For $t < T$, it generates a set of hypotheses H_t from the current description and the information accumulated so far:

$$H_t = \mathrm{Anticipate}(O_t, K_t, G_t, \mathcal{M}_t, \mathcal{T}) \tag{2}$$

Each hypothesis identifies an object attribute, a possible next value or change direction, and a visual sign that would distinguish that outcome. For a drawer, alternatives might be opening further, closing, or remaining unchanged; they make its current extension an informative attribute to inspect. These hypotheses concern the next available observation and use only the images observed so far.

The MLLM converts the hypotheses into an inspection plan J_t: a set of questions about the target's current attributes. Each question specifies what visual evidence to examine and how that evidence supports a description. Questions remain neutral about the expected answer; for example, the drawer hypothesis leads to asking how far the drawer is currently extended. The plan also includes task-relevant attributes already identified in O_t and unresolved questions in G_t. We denote the hypotheses available to planning by H_t^{use}: these are H_t for $t < T$ and the preceding hypotheses H_{T-1} for the final observation. The plan is constructed as:

$$J_t = \mathrm{Plan}(O_t, H_t^{\mathrm{use}}, \mathcal{M}_t, G_t, \mathcal{T}) \tag{3}$$

The MLLM inspects the current image according to J_t and produces the state readout S_t by $S_t = \mathrm{Read}(I_t, P_{\leq t}, J_t, O_t)$. The readout records attribute values and their supporting visual evidence, leaving values unresolved when the image is insufficient. Subsequent observations check this readout and its associated hypotheses through sequential visual correction.

2.3 Sequential Visual Correction

When the agent processes image I_t ($t > 1$), sequential visual correction checks the preceding state description S_{t-1} and predicted changes H_{t-1} against the new visual evidence. The MLLM uses the observation descriptions O_{t-1} and O_t, which are obtained without predicted values. If these descriptions provide insufficient evidence, the MLLM examines the two images again to answer the unresolved visual questions. This additional inspection produces W_t without receiving the preceding claimed value or predicted outcome; otherwise, W_t is empty. The comparison produces feedback E_t:

$$E_t = \mathrm{Compare}(S_{t-1}, H_{t-1}, O_{t-1}, O_t, W_t, \mathcal{T}) \tag{4}$$

The feedback identifies supported or contradicted hypotheses, observed state changes, errors in earlier descriptions, and questions that the available evidence does not resolve. If an object previously observed as standing is now seated, the agent records the transition and rejects a prediction of continued standing; it does not change the earlier standing observation. An earlier description is corrected only when visual evidence contradicts that description. When the object's identity across the two images is uncertain, the agent leaves the transition unresolved.

The state record and local guidance start empty: $L_0 = \varnothing$ and $G_0 = \varnothing$. For $t > 1$, the correction operation applies evidence-supported repairs to the preceding state record L_{t-1}, producing K_t, while the guidance operation updates the unresolved questions in G_t:

$$K_t = \mathrm{Correct}(L_{t-1}, E_t), \quad G_t = \mathrm{Guide}(G_{t-1}, E_t) \tag{5}$$

Here, K_t contains the corrected history before the current state is added. Predictive attribute selection then uses K_t and G_t to guide inspection of the current image, yielding S_t. The agent adds this readout to the corrected history to obtain the updated state record:

$$L_t = \mathrm{Update}(K_t, S_t) \tag{6}$$

For the first image, E_1, K_1, and G_1 are empty, and S_1 initializes the state record.

After all T images have been processed, the MLLM generates the final description from L_T. It selects supported states and changes within the object and temporal scope specified by $\mathcal{T}$:

$$\widehat{Y} = \mathrm{Compose}(L_T, \mathcal{T}) \tag{7}$$

2.4 Feedback-guided Skill Revision

Feedback-guided skill revision improves observation instructions using existing reference captions while keeping model parameters fixed. We divide the source videos into a revision set $\mathcal{D}_{\mathrm{rev}}$ for identifying problems and proposing edits, and a selection set $\mathcal{D}_{\mathrm{sel}}$ for evaluating those edits. The sets contain different videos to test whether edits generalize beyond the examples that motivate them. In each round, LASER processes the revision videos with the current library. The MLLM compares generated captions with references and examines the corresponding observations, hypotheses, inspection questions, and state records to identify discrepancy sources. The comparison identifies omissions and explicit contradictions; details unmentioned in the reference remain unjudged.

When a relevant attribute is never inspected, it examines whether the skill's trigger or alternative changes fail to direct attention to that attribute. When inspection produces an unsupported value, it examines the visual signs and observation operations. When a supported fact appears in the state record but is missing from the final caption, it examines the expression constraints. For a discrepancy linked to an existing skill, the MLLM proposes an edit to one responsible field; if no skill covers the situation, it proposes a new skill. Each proposal produces a candidate library $\mathcal{M}'$ derived independently from the current library $\mathcal{M}$. The MLLM proposes no edit when the available evidence does not identify a responsible instruction.

LASER compares each candidate with the current library on $\mathcal{D}_{\mathrm{sel}}$. For an example x, which contains the images, target information, and task instruction, let Y_x^* denote its reference caption. The score q

measures agreement between the generated and reference captions, giving the mean caption quality:

$$Q(\mathcal{M}) = \frac{1}{|\mathcal{D}_{\text{sel}}|} \sum_{x \in \mathcal{D}_{\text{sel}}} q\big(\text{Caption}(x; \mathcal{M}), Y_x^*\big) \tag{8}$$

Here, Caption returns the description produced by Equation 1; references are used for scoring rather than caption generation. To check contradictions, the MLLM extracts explicit facts from each reference caption once and uses the same facts to assess every candidate. An MLLM judge determines whether a generated caption contains a claim incompatible with a reference fact concerning the same object and temporal scope. Omissions, details absent from the reference, and uncertain comparisons do not count as contradictions. The contradiction score $X(\mathcal{M})$ is the fraction of selection examples containing at least one explicit contradiction.

Selection requires an improvement in caption quality without increasing reference contradictions. For a positive quality margin ϵ, the acceptance rule is:

$$\text{Accept}(\mathcal{M}') = \mathbf{1}[Q(\mathcal{M}') - Q(\mathcal{M}) \geq \epsilon,\ X(\mathcal{M}') \leq X(\mathcal{M})] \tag{9}$$

The accepted candidate with the highest caption quality replaces the current library; if none is accepted, the library is retained. The resulting skill library remains fixed during evaluation.

3 Experiment Setting

3.1 Datasets

We evaluate object-conditioned captioning on OSCaR (Nguyen et al., 2024) and PLM RCap (Cho et al., 2025). We use OSCaR's released EPIC state-change slot, containing 7K training clips and 500 test clips. Each example pairs three ordered frames with a named object and a reference description of its state change. PLM RCap contains 179K training, 4K development, and 10K test region–interval captions. Given a target masklet and its video interval, the task describes the target's actions and interactions. For each benchmark, we select 500 examples from its training split and use their reference captions to adapt observation skills. We divide these examples into video-disjoint revision and selection sets, using the former to propose instruction edits and the latter to evaluate them.

3.2 Baselines and Comparing Models

We compare LASER with the following baselines. **Direct Captioning** generates a target description from the input frames and task instruction in one pass. **Sequential Readout** describes object states in temporal order and summarizes their changes without anticipating future states. **Generic Reinspection** extends sequential readout by checking for omissions or inconsistencies. **Attribute Checklist** selects current-state attributes for inspection and retains visual correction, without next-step hypotheses. **Free Anticipation** predicts possible future events without identifying their distinguishing attributes and visual signs. **LASER (No Skill Revision)** uses unadapted initial observation skills. **LASER (GEPA Optimization)** uses GEPA (Agrawal et al., 2026) to optimize the same prompt interfaces instead of feedback-guided skill revision. **LASER (Full)** is our full approach.

We also compare LASER with five existing approaches. **InstanceCap** (Fan et al., 2025) organizes video instances into structured descriptions of their appearance, actions, motion, and position. **Matters** (Peng et al., 2025) describes local image patches and hierarchically aggregates their details through semantic filtering. **RICO** (Wang et al., 2025a) reconstructs images from captions and revises the captions by comparing the reconstructions with the original images. **ReAgent-V** (Zhou et al., 2026) combines task-relevant frame selection and tool-derived evidence with critical feedback and multi-perspective answer revision. **ScaleCap** (Xing et al., 2026) expands captions through targeted visual questions and filters unsupported sentences through contrastive scoring. We adapt these approaches to object-conditioned captioning and evaluate them on the benchmarks.

3.3 Implementation Details

Our backbones MLLMs comprise Qwen3.5-4B, Qwen3.5-9B, and Qwen3.5-35B-A3B (Team, 2026), plus Ministral-3-3B (Liu et al., 2026). The two dense Qwen models each have 32 language layers, with hidden dimensions of 2,560 and 4,096 for the 4B and 9B variants, respectively. The 35B-A3B

Table 1: Comparison of our approach with various baselines across four backbone MLLMs on the test set of OSCaR and PLM RCap, where "+ LASER (full)" denotes our full model.

Approach	OSCaR				PLM RCap
	BLEU↑	ROUGE-1↑	ROUGE-2↑	ROUGE-L↑	Score↑
Qwen3.5-4B	7.7	28.4	8.7	25.2	24.7
+ Sequential Readout	8.0	29.2	8.8	26.2	26.1
+ Generic Reinspection	8.0	29.8	8.7	25.9	25.3
+ Attribute Checklist	8.6	29.6	9.2	26.7	26.8
+ Free Anticipation	8.1	29.9	9.0	25.2	25.3
+ LASER (No Skill Revision)	8.7	30.1	8.9	26.7	25.9
+ LASER (GEPA Optimization)	9.0	**31.1**	9.1	**29.0**	**27.5**
+ LASER (Full)	**9.3**	30.9	**9.3**	28.9	27.3
Qwen3.5-9B	8.6	29.8	9.5	26.9	25.6
+ Sequential Readout	9.1	30.6	9.5	28.5	26.7
+ Generic Reinspection	9.3	31.1	9.9	28.7	25.8
+ Attribute Checklist	9.6	30.8	9.9	27.7	27.2
+ Free Anticipation	9.0	30.5	9.6	28.5	26.4
+ LASER (No Skill Revision)	9.1	30.5	9.8	28.2	26.8
+ LASER (GEPA Optimization)	9.5	31.6	**10.1**	**29.9**	27.5
+ LASER (Full)	**9.9**	**32.0**	**10.1**	29.6	**28.1**
Qwen3.5-35B-A3B	8.8	30.4	9.7	27.7	26.3
+ Sequential Readout	9.0	30.3	10.0	27.9	27.5
+ Generic Reinspection	9.0	31.0	9.8	29.4	27.0
+ Attribute Checklist	9.6	31.5	10.1	29.3	27.3
+ Free Anticipation	9.2	31.0	9.7	28.2	27.3
+ LASER (No Skill Revision)	9.7	30.9	10.0	28.9	28.1
+ LASER (GEPA Optimization)	10.3	32.2	10.7	**30.7**	**29.0**
+ LASER (Full)	**10.6**	**33.1**	**10.8**	30.4	28.8
Ministral-3-3B	6.8	27.1	6.1	24.0	22.9
+ Sequential Readout	7.6	28.2	6.0	24.9	23.8
+ Generic Reinspection	8.1	28.4	6.3	25.4	23.8
+ Attribute Checklist	8.1	28.6	7.0	25.5	23.6
+ Free Anticipation	7.4	27.7	7.4	25.4	23.4
+ LASER (No Skill Revision)	8.0	28.7	6.9	25.9	23.3
+ LASER (GEPA Optimization)	8.1	29.1	8.0	26.4	24.7
+ LASER (Full)	**8.8**	**29.9**	**8.2**	**27.0**	**25.2**

variant uses 40 language layers with a hidden dimension of 2,048; its mixture-of-experts layers select 8 of 256 routed experts per token and include a shared expert, with an intermediate dimension of 512 for each expert. Ministral-3-3B has 26 language layers with a hidden dimension of 3,072 and a 24-layer vision encoder.

We divide the adaptation examples into revision and selection sets at a 4:1 ratio, keeping examples from the same video in one set. For PLM RCap, we uniformly sample eight frames from each target interval; OSCaR uses its three provided frames. At each observation, LASER selects up to three relevant skills and generates up to three alternative next-step changes. Skill adaptation runs for at most five rounds, with up to three candidate libraries evaluated per round. We normalize caption quality scores to $[0, 1]$ and require an improvement of at least $\epsilon = 0.005$ without increasing reference contradictions. Adaptation stops when no candidate satisfies these criteria. All model parameters remain frozen, and the adapted skill library remains fixed during evaluation.

For evaluation, we use different metrics for the two benchmarks. For OSCaR, we report BLEU (Papineni et al., 2002), ROUGE-1, ROUGE-2, and ROUGE-L (Lin, 2004), with ROUGE-L as the primary metric. For PLM RCap, we use the official language-model-judged reference-caption similarity score (Cho et al., 2025). Higher values indicate better caption agreement on both tasks. Evaluation references are accessible only to the scorer.

Table 2: Comparison of our approach with existing studies on using Qwen3.5-9B as the backbone MLLM. The best and second-best results are marked by boldface and underlines, respectively.

Approach	OSCaR				PLM RCap
	BLEU↑	ROUGE-1↑	ROUGE-2↑	ROUGE-L↑	Score↑
Qwen3.5-9B	8.6	29.8	9.5	26.9	25.6
InstanceCap (Fan et al., 2025)	<u>9.6</u>	31.1	<u>9.9</u>	27.6	27.4
Matters (Peng et al., 2025)	9.1	30.2	9.4	27.1	25.4
RICO (Wang et al., 2025a)	8.9	30.1	<u>9.9</u>	27.5	25.5
ReAgent-V (Zhou et al., 2026)	9.2	<u>31.4</u>	9.8	28.7	<u>27.8</u>
ScaleCap (Xing et al., 2026)	9.3	30.1	9.8	<u>29.4</u>	26.6
LASER	**9.9**	**32.0**	**10.1**	**29.6**	**28.1**

4 Results and Analysis

4.1 Overall Results

Table 1 presents the results of our approach and baselines on OSCaR and PLM RCap across four backbone MLLMs. There are several observations. First, LASER consistently outperforms direct captioning, sequential readout, and generic reinspection across all metrics. These gains suggest that accurate state descriptions benefit from selecting informative attributes beyond organizing observations temporally or inspecting them repeatedly. Second, LASER achieves higher scores than both attribute checklists and free anticipation across the evaluated backbones. This supports connecting possible changes to specific visual attributes, so that anticipation guides what to inspect in the current image. Third, feedback-guided skill revision improves every reported metric over LASER with the initial skill library. This indicates that caption feedback helps refine reusable observation instructions without updating model parameters. Fourth, LASER improves over direct captioning for all four backbones, spanning different model sizes, model families, and dense and mixture-of-experts architectures, which demonstrates the robustness of our approach.

Table 2 compares LASER with five existing approaches adapted to object-conditioned captioning using Qwen3.5-9B as the backbone MLLM. LASER outperforms all compared approaches on both benchmarks, which demonstrates the effectiveness of using anticipated changes to identify state-distinguishing attributes and refining their descriptions through subsequent visual evidence and caption feedback.

4.2 Ablation Study

We examine whether anticipated changes help identify current-state evidence and whether independent observation limits reliance on predicted values. Using Qwen3.5-9B as the backbone MLLM on both benchmarks, we independently remove current reinspection, remove discriminating visual signs, or expose the initial observer to preceding hypotheses, retaining the full approach's revised skills in each case. Table 3 compares these variants with full LASER and direct captioning. Removing current reinspection lowers every metric, supporting the use of anticipated changes to guide a second examination of the present image. Removing discriminating signs also lowers every metric and produces the largest ROUGE-L decline among the ablations. This pattern suggests that identifying possible changes is more useful when those changes are connected to explicit visual criteria for inspection. Exposing the initial observer to predicted changes reduces all reported scores and produces the largest decline in RCap similarity among the ablations. The result supports keeping initial visual observations independent of forecasts, so that later comparisons rely on separately obtained evidence. All three ablated variants remain above direct captioning, indicating that each removed operation contributes to the full approach while the remaining observation structure retains value.

4.3 Case Study

We use two region-captioning cases from source videos to illustrate how LASER handles supported and contradicted predictions. Figure 2 shows the input frames and target regions in panel (a), state

Table 3: Results of ablation study of predictive observation with Qwen3.5-9B.

Approach	OSCaR				PLM RCap
	BLEU↑	ROUGE-1↑	ROUGE-2↑	ROUGE-L↑	Score↑
LASER (Full)	**9.9**	**32.0**	**10.1**	**29.6**	**28.1**
w/o Current Reinspection	9.1	31.2	10.0	28.9	27.4
w/o Discriminating Signs	9.8	31.7	9.8	28.1	27.5
w/o Independent Observation	9.3	31.8	9.7	29.0	27.3
Qwen3.5-9B (Base)	8.6	29.8	9.5	26.9	25.6

estimates and predictions in panel (b), subsequent visual checks in panel (c), and the generated and reference captions in panels (d) and (e). Green contours identify the target whose state is tracked.

In Case 1, the anticipated forward lean directs inspection toward the target's torso angle relative to the sink. The later image supports this change, and the final caption describes the lean while preserving the ongoing dishwashing activity. The reference mentions dishwashing without explicitly describing the lean, so the displayed images provide the evidence for this additional posture detail. The case illustrates how an anticipated change becomes a specific visual criterion used in the final description. In Case 2, LASER anticipates continued standing with a slight forward lean, but the later observation shows the target seated on the bicycle. Sequential visual correction rejects the forecast, retains the earlier standing state, and records the newly observed transition. The final caption describes the target sitting on the bicycle, consistent with the reference's account of getting onto it. The case illustrates that an incorrect forecast still leaves room for evidence-supported state readout: subsequent observations determine the transition without changing a supported earlier state.

5 Related Work

Object-centered video understanding studies state localization (Xue et al., 2024; Souček et al., 2022), transformation graphs (Sun et al., 2026), and segmentation through appearance changes (Yu et al., 2023). Region-captioning approaches emphasize object representations (Yuan et al., 2025) and local–global context (Lian et al., 2025), while OSCaR (Nguyen et al., 2024) and PerceptionLM (Cho et al., 2025) provide state-change and region-captioning tasks, respectively. Agent approaches organize reasoning and interaction (Yao et al., 2022), execute visual programs (Gupta & Kembhavi, 2023), and gather evidence through iterative inspection (Wang et al., 2024a), keyframe selection (Wang et al., 2025b), or evidence-sufficiency assessment (Wang et al., 2026). StreamAgent anticipates where and when task-relevant information will appear to guide evidence acquisition and response timing in streaming video question answering (Yang et al., 2025). Other uses of possible outcomes include counterfactual supervision for video representation learning (Kung et al., 2025) and robot behavior forecasting in TRACE (Puthumanaillam et al., 2025). Feedback-based approaches revise responses (Madaan et al., 2023) or retain experience as reflections (Shinn et al., 2023), workflows (Wang et al., 2024b), manuals (Chen et al., 2024), and evolving contextual playbooks (Zhang et al., 2026). Prompt optimization uses scored candidates (Yang et al., 2024), textual critiques (Pryzant et al., 2023), modular instruction search (Khattab et al., 2023; Opsahl-Ong et al., 2024), and feedback propagation (Yuksekgonul et al., 2025); GEPA uses execution traces to reflect on and evolve prompts (Agrawal et al., 2026). LASER focuses on the role of anticipation in determining what constitutes an informative description of the present: alternative next-step changes identify the object attributes and visual signs to inspect in the current image, whose evidence establishes their values. Subsequent observations check these values and transitions, while caption discrepancies guide revisions to the responsible attribute-selection, visual-inspection, or expression instructions, making the observation criteria themselves the target of adaptation.

6 Conclusion

In this paper, we propose look-ahead state readout (LASER), an anticipation-guided agent that uses possible next-step changes to determine which object attributes deserve attention in the present.

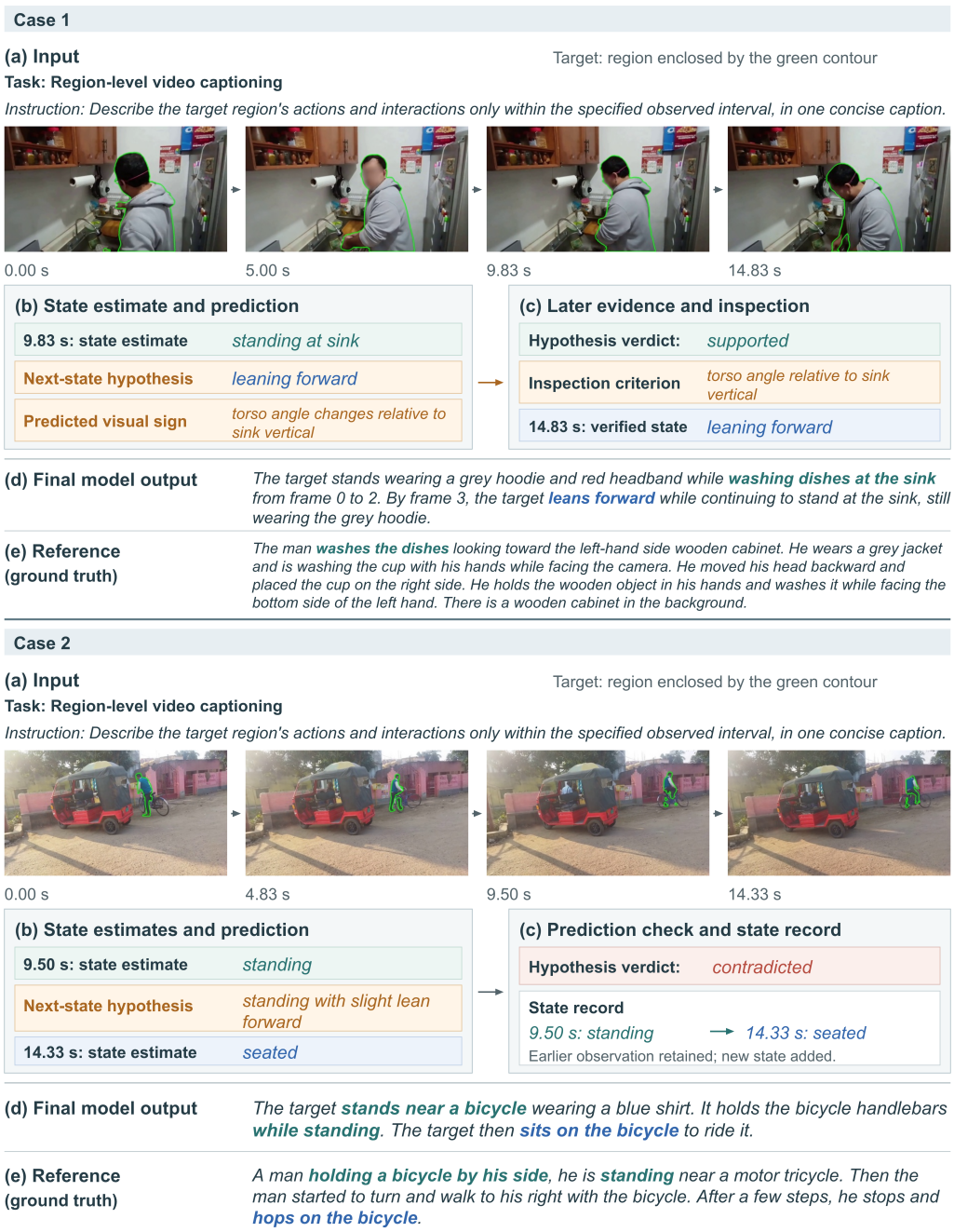

Figure 2: Two region-captioning cases illustrating selected posture changes. Each case shows the input frames and target region, an anticipated change, later visual evidence, the model caption, and the reference. Case 1 links an anticipated forward lean to a visual inspection criterion. Case 2 rejects a standing forecast and preserves the observed transition to sitting. Green contours identify the target.

LASER converts anticipated changes into current-state inspection questions and uses subsequent visual evidence to refine the state record for description. Feedback-guided skill revision improves reusable observation instructions through caption feedback while keeping all model parameters fixed. Experiments on OSCaR and PLM RCap demonstrate consistent improvements over direct captioning and alternative observation strategies across four MLLM backbones, and LASER outperforms the five compared existing approaches. Ablations support the importance of inspecting current attributes, identifying distinguishing visual signs, and checking predictions against independent observations.

AI USE STATEMENT

We used generative AI tools to support manuscript editing and the collection and organization of research information. For manuscript editing, these tools helped refine wording, improve sentence flow, and maintain consistent terminology. For research preparation, they assisted with identifying relevant literature, summarizing its key points, and organizing information by research topic.

REPRODUCIBILITY STATEMENT

The implementation details and experiment settings are described in Section 3.3 (Implementation Details). The code is available for review at `https://anonymous.4open.science/r/LASER-EC56`.

REFERENCES

Lakshya A Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziems, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, Michael J Ryan, Meng Jiang, et al. Gepa: Reflective prompt evolution can outperform reinforcement learning. In *International Conference on Learning Representations*, volume 2026, pp. 8479–8565, 2026.

Ali Athar, Xueqing Deng, and Liang-Chieh Chen. Vicas: A dataset for combining holistic and pixel-level video understanding using captions with grounded segmentation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19023–19035, 2025.

Mu Cai, Reuben Tan, Jianrui Zhang, Bocheng Zou, Kai Zhang, Feng Yao, Fangrui Zhu, Jing Gu, Yiwu Zhong, Yuzhang Shang, Yao Dou, Jaden Park, Jianfeng Gao, Yong Jae Lee, and Jianwei Yang. Temporalbench: Benchmarking fine-grained temporal understanding for multimodal video models. *arXiv preprint arXiv:2410.10818*, 2024.

Minghao Chen, Yihang Li, Yanting Yang, Shiyu Yu, Binbin Lin, and Xiaofei He. Automanual: Constructing instruction manuals by llm agents via interactive environmental learning. In *Advances in Neural Information Processing Systems*, volume 37, pp. 589–631, 2024.

Jang Hyun Cho, Andrea Madotto, Effrosyni Mavroudi, Triantafyllos Afouras, Tushar Nagarajan, Muhammad Maaz, Yale Song, Tengyu Ma, Shuming Hu, Suyog Jain, Miguel Martin, Huiyu Wang, Hanoona Bangalath, Peize Sun, Po-Yao Huang, Daniel Bolya, Nikhila Ravi, Shashank Jain, Tammy Stark, Seungwhan Moon, Babak Damavandi, Vivian Lee, Andrew Westbury, Salman H Khan, Philipp Krähenbühl, Piotr Dollár, Lorenzo Torresani, Kristen Grauman, and Christoph Feichtenhofer. PerceptionLM: Open-Access Data and Models for Detailed Visual Understanding. In *Advances in Neural Information Processing Systems*, volume 38, pp. 23475–23537, 2025.

Tiehan Fan, Kepan Nan, Rui Xie, Penghao Zhou, Zhenheng Yang, Chaoyou Fu, Xiang Li, Jian Yang, and Ying Tai. Instancecap: Improving text-to-video generation via instance-aware structured caption. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28974–28983, 2025.

Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14953–14962, 2023.

Evangelos Kazakos, Cordelia Schmid, and Josef Sivic. Large-scale pre-training for grounded video caption generation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 24434–24444, 2025.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. DSPy: Compiling Declarative Language Model Calls into Self-Improving Pipelines. *arXiv preprint arXiv:2310.03714*, 2023.

Chi-Hsi Kung, Frangil Ramirez, Juhyung Ha, Yi-Ting Chen, David Crandall, and Yi-Hsuan Tsai. What changed and what could have changed? state-change counterfactuals for procedure-aware video representation learning. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12294–12306, 2025.

Long Lian, Yifan Ding, Yunhao Ge, Sifei Liu, Hanzi Mao, Boyi Li, Marco Pavone, Ming-Yu Liu, Trevor Darrell, Adam Yala, et al. Describe anything: Detailed localized image and video captioning. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21766–21777, 2025.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Alexander H. Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan, Alexandre Sablayrolles, Amélie Héliou, Amos You, Andy Ehrenberg, Andy Lo, Anton Eliseev, Antonia Calvi, Avinash Sooriyarachchi, Baptiste Bout, Baptiste Rozière, Baudouin De Monicault, Clémence Lanfranchi, Corentin Barreau, Cyprien Courtot, Daniele Grattarola, Darius Dabert, Diego de las Casas, Elliot Chane-Sane, Faruk Ahmed, Gabrielle Berrada, Gaëtan Ecrepont, Gauthier Guinet, Georgii Novikov, Guillaume Kunsch, Guillaume Lample, Guillaume Martin, Gunshi Gupta, Jan Ludziejewski, Jason Rute, Joachim Studnia, Jonas Amar, Joséphine Delas, Josselin Somerville Roberts, Karmesh Yadav, Khyathi Chandu, Kush Jain, Laurence Aitchison, Laurent Fainsin, Léonard Blier, Lingxiao Zhao, Louis Martin, Lucile Saulnier, Luyu Gao, Maarten Buyl, Margaret Jennings, Marie Pellat, Mark Prins, Mathieu Poirée, Mathilde Guillaumin, Matthieu Dinot, Matthieu Futeral, Maxime Darrin, Maximilian Augustin, Mia Chiquier, Michel Schimpf, Nathan Grinsztajn, Neha Gupta, Nikhil Raghuraman, Olivier Bousquet, Olivier Duchenne, Patricia Wang, Patrick von Platen, Paul Jacob, Paul Wambergue, Paula Kurylowicz, Pavankumar Reddy Muddireddy, Philomène Chagniot, Pierre Stock, Pravesh Agrawal, Quentin Torroba, Romain Sauvestre, Roman Soletskyi, Rupert Menneer, Sagar Vaze, Samuel Barry, Sanchit Gandhi, Siddhant Waghjale, Siddharth Gandhi, Soham Ghosh, Srijan Mishra, Sumukh Aithal, Szymon Antoniak, Teven Le Scao, Théo Cachet, Theo Simon Sorg, Thibaut Lavril, Thiziri Nait Saada, Thomas Chabal, Thomas Foubert, Thomas Robert, Thomas Wang, Tim Lawson, Tom Bewley, Tom Bewley, Tom Edwards, Umar Jamil, Umberto Tomasini, Valeriia Nemychnikova, Van Phung, Vincent Maladière, Virgile Richard, Wassim Bouaziz, Wen-Ding Li, William Marshall, Xinghui Li, Xinyu Yang, Yassine El Ouahidi, Yihan Wang, Yunhao Tang, and Zaccharie Ramzi. Ministral 3. *arXiv preprint arXiv:2601.08584*, 2026.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos? In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 8731–8772, 2024.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. In *Advances in neural information processing systems*, volume 36, pp. 46534–46594, 2023.

Nguyen Nguyen, Jing Bi, Ali Vosoughi, Yapeng Tian, Pooyan Fazli, and Chenliang Xu. Oscar: Object state captioning and state change representation. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 3565–3576, 2024.

Krista Opsahl-Ong, Michael J Ryan, Josh Purtell, David Broman, Christopher Potts, Matei Zaharia, and Omar Khattab. Optimizing instructions and demonstrations for multi-stage language model programs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 9340–9366, 2024.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Ruotian Peng, Haiying He, Yake Wei, Yandong Wen, and Di Hu. Matters: Training-free fine-grained image caption enhancement via local perception. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3963–3973, 2025.

Reid Pryzant, Dan Iter, Jerry Li, Yin Lee, Chenguang Zhu, and Michael Zeng. Automatic prompt optimization with "gradient descent" and beam search. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 7957–7968, 2023.

Gokul Puthumanaillam, Paulo Padrao, Jose Fuentes, Pranay Thangeda, William E Schafer, Jae Hyuk Song, Karan Jagdale, Leonardo Bobadilla, and Melkior Ornik. Trace: A self-improving framework for robot behavior forecasting with vision-language models. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 11128–11135, 2025.

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

Tomáš Souček, Jean-Baptiste Alayrac, Antoine Miech, Ivan Laptev, and Josef Sivic. Look for the change: Learning object states and state-modifying actions from untrimmed web videos. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13936–13946, 2022.

Yihong Sun, Xinyu Yang, Jennifer Sun, and Bharath Hariharan. Tracking and understanding object transformations. In *Advances in Neural Information Processing Systems*, volume 38, pp. 81039–81065, 2026.

Qwen Team. Qwen3.5-omni technical report. *arXiv preprint arXiv:2604.15804*, 2026.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. Videoagent: Long-form video understanding with large language model as agent. In *European Conference on Computer Vision*, pp. 58–76, 2024a.

Yuchi Wang, Yishuo Cai, Shuhuai Ren, Sihan Yang, Linli Yao, Yuanxin Liu, Yuanxing Zhang, Pengfei Wan, and Xu Sun. Rico: Improving accuracy and completeness in image recaptioning via visual reconstruction. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 21796–21815, 2025a.

Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. Videotree: Adaptive tree-based video representation for llm reasoning on long videos. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3272–3282, 2025b.

Ziyang Wang, Honglu Zhou, Shijie Wang, Junnan Li, Caiming Xiong, Silvio Savarese, Mohit Bansal, Michael S Ryoo, and Juan Carlos Niebles. Active video perception: Iterative evidence seeking for agentic long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9088–9099, 2026.

Zora Zhiruo Wang, Jiayuan Mao, Daniel Fried, and Graham Neubig. Agent workflow memory. *arXiv preprint arXiv:2409.07429*, 2024b.

Long Xing, Qidong Huang, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Jinsong Li, Shuangrui Ding, Weiming Zhang, Nenghai Yu, Jiaqi Wang, Feng Wu, and Dahua Lin. Scalecap: Scalable image captioning via dual-modality debiasing. In *International Conference on Learning Representations*, volume 2026, pp. 32266–32285, 2026.

Zihui Xue, Kumar Ashutosh, and Kristen Grauman. Learning object state changes in videos: An open-world perspective. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18493–18503, 2024.

Zihui Xue, Joungbin An, Xitong Yang, and Kristen Grauman. Progress-aware video frame captioning. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13639–13650, 2025.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. Large language models as optimizers. In *International Conference on Learning Representations*, volume 2024, pp. 12028–12068, 2024.

Haolin Yang, Feilong Tang, Lingxiao Zhao, Xinlin Zhuang, Yifan Lu, Xiang An, Ming Hu, Xiaofeng Zhang, Abdalla Swikir, Junjun He, et al. Streamagent: Towards anticipatory agents for streaming video understanding. *arXiv preprint arXiv:2508.01875*, 2025.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

Jiangwei Yu, Xiang Li, Xinran Zhao, Hongming Zhang, and Yu-Xiong Wang. Video state-changing object segmentation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20382–20391, 2023.

Yuqian Yuan, Hang Zhang, Wentong Li, Zesen Cheng, Boqiang Zhang, Long Li, Xin Li, Deli Zhao, Wenqiao Zhang, Yueting Zhuang, Jianke Zhu, and Lidong Bing. Videorefer suite: Advancing spatial-temporal object understanding with video llm. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18970–18980, 2025.

Mert Yuksekgonul, Federico Bianchi, Joseph Boen, Sheng Liu, Pan Lu, Zhi Huang, Carlos Guestrin, and James Zou. Optimizing generative ai by backpropagating language model feedback. *Nature*, 639(8055):609–616, 2025.

Qizheng Zhang, Changran Hu, Shubhangi Upasani, Boyuan Ma, Fenglu Hong, Vamsidhar Kamanuru, Jay Rainton, Chen Wu, Mengmeng Ji, Hanchen Li, et al. Agentic context engineering: Evolving contexts for self-improving language models. In *International Conference on Learning Representations*, volume 2026, pp. 86069–86100, 2026.

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. Expel: Llm agents are experiential learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19632–19642, 2024.

Yiyang Zhou, Yangfan He, Yaofeng Su, Siwei Han, Joel Jang, Gedas Bertasius, Mohit Bansal, and Huaxiu Yao. Reagent-v: A reward-driven multi-agent framework for video understanding. *Advances in Neural Information Processing Systems*, 38:151454–151491, 2026.