CaReVOS: Causal Reasoning Across Missing Events for Video Object Segmentation

Anonymous authors
Paper under double-blind review

Abstract

Reasoning video object segmentation requires models to identify and segment objects from implicit queries that depend on understanding events and object relations. Existing settings, however, largely assume that the interactions defining these roles are observable. To study causal reasoning across missing events, we formulate a partial-observation setting in which models infer and ground the relevant participants from observations before and after an unobserved interaction. Based on this formulation, we introduce CaReVOS, a benchmark comprising 1,432 videos, 6,386 reasoning expressions, and 373,828 object mask annotations. It includes consequence, source, multi-solution, and counterfactual questions, with human-verified answers and masks. Because the interaction is hidden, answering these questions requires relating the observed setup to its outcome before selecting the target. To improve both this inference and the resulting segmentation, we develop Event-OPD. Supervised fine-tuning teaches target grounding, while privileged on-policy distillation lets a full-event teacher guide a student that sees only the opening and closing segments. Extensive evaluations show that existing referring and reasoning segmentation models struggle under this setting, while adaptation to CaReVOS yields substantial gains. Our distilled model achieves 58.33 $\mathcal{J}$, 61.98 $\mathcal{F}$, and 60.16 $\mathcal{J}\&\mathcal{F}$, exceeding the strongest trained baseline by 2.21 points and its supervised initialization by 1.62 points.

1 Introduction

Referring video object segmentation (RVOS) identifies and tracks objects described by natural language expressions (Wu et al., 2022; Ding et al., 2023). Reasoning video object segmentation extends this task to questions whose answers must be inferred from the video and commonsense knowledge (Yan et al., 2024; Bai et al., 2024; Deng et al., 2025). A target may be defined by the action it performs, the change it causes, or its relation to another object. Answering such a question requires both understanding the event and locating its participants. Segmentation therefore tests whether a model's reasoning leads to the correct objects and their locations throughout the video.

Existing work has broadened the evidence used to identify objects in video. MeViS (Ding et al., 2023) emphasizes motion, while VISA (Yan et al., 2024), VideoLISA (Bai et al., 2024), and GLUS (Lin et al., 2025) combine language reasoning with segmentation. GroundMoRe (Deng et al., 2025) grounds event questions in object masks, and CoT-RVS (Kao et al., 2026) reasons across frames to select targets. FeVOS (Lan et al., 2026) further identifies objects through anticipated future events. These event roles also connect segmentation to causal video reasoning, which asks why events occur and how outcomes would change under other conditions (Yi et al., 2020; Chen et al., 2024). Work on causal grounding further asks where the evidence for these explanations appears (Zhang et al., 2026b). Yet selecting participants of an unseen interaction requires relating the initial state to the observed outcome, a form of reasoning that remains less explored.

Figure 1 shows a wooden block tower before and after it collapses. The opening segment establishes the intact tower and nearby participants, while the closing segment reveals the fallen blocks. Together, these observations show the possible participants and the change to be explained, but the action linking them is hidden. A model must therefore combine the setup and outcome to infer who could have caused the collapse and segment that participant throughout the video.

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.

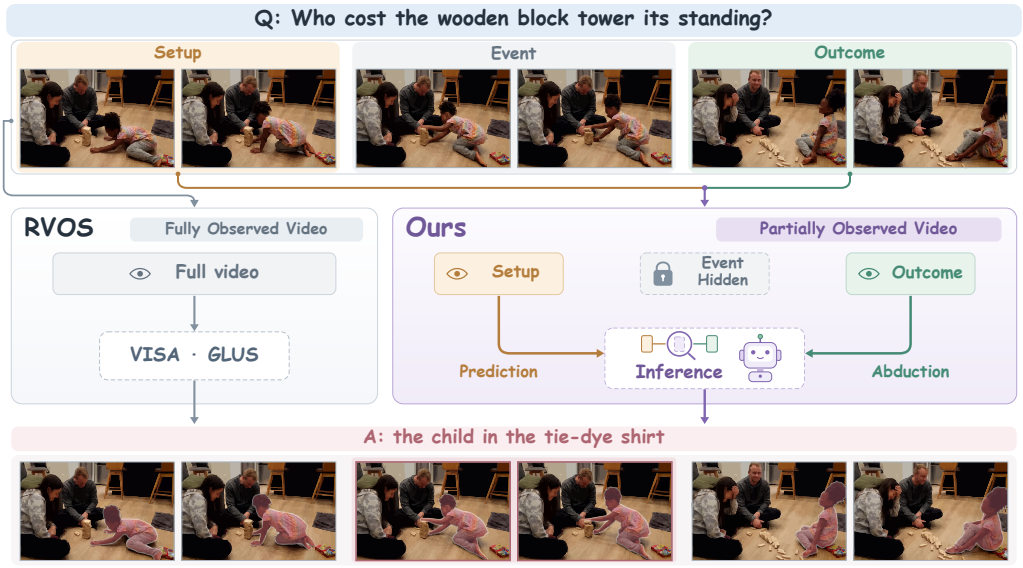

Figure 1: CaReVOS overview. From the setup and outcome, the model infers a participant while the event is hidden, then segments that participant throughout the full video.

To test this inference through object segmentation, we introduce CaReVOS. Given a reasoning expression and observations before and after an event, a model infers the participants that satisfy the query while the interaction itself is withheld. The target's identity therefore depends on its role in the unseen event, which the model infers by connecting the initial state to the observed outcome. For example, a question about what caused a change may refer to the responsible participant, whereas one about what was affected may refer to an object whose state has changed. The model then segments the selected targets throughout the full video, allowing us to evaluate both its choice of event participants and how accurately it locates and tracks them throughout the interaction.

CaReVOS contains 1,432 videos, 6,386 reasoning expressions, and 373,828 object mask annotations. Its questions range from identifying the source of a change to asking whose outcome would differ if an action had not occurred. Answers can refer to one object or several objects sharing the queried role. We build the dataset through automatic selection, manual clip refinement, expression and mask annotation, and final human verification. This process links each question to a concrete set of objects and their trajectories, supporting evaluation of reasoning and object localization.

To address both event inference and object grounding, we develop Event-OPD, a training framework that connects a vision-language model to a video segmenter through a learned object representation. Supervised fine-tuning first equips the model with the basic segmentation capability by linking reference answers to object representations for mask prediction. Building on this grounding ability, we distill from a frozen teacher that observes the complete event while the student sees only the opening and closing segments. The teacher uses this additional evidence to guide how the student continues its own generated answers. Alongside this token-level guidance, we align the student and teacher object representations to extend distillation to the features that encode the inferred target objects for mask prediction. Answer and mask supervision ties this guidance to the target objects.

On CaReVOS, Event-OPD reaches 60.16 $\mathcal{J}\&\mathcal{F}$, exceeding the strongest evaluated baseline by 2.21 points and its SFT initialization by 1.62 points. The latter gain shows that Event-OPD further improves the model after supervised target grounding.

In summary, we make the following contributions:

- We introduce CaReVOS, a task of causal reasoning over missing events, in which a model must explain an observed change through the participants of an interaction it never sees, spanning consequence, source, multi-solution, and counterfactual questions whose answers are grounded as full-video masks.

- We construct the CaReVOS dataset for this task, in which every question, answer, and mask trajectory is verified by human reviewers, comprising 1,432 videos, 6,386 reasoning expressions, and 373,828 object masks drawn from seven source datasets.
- We develop Event-OPD, a training framework that combines supervised fine-tuning with privileged on-policy distillation, in which guidance from a full-event teacher transfers event inference to a student that sees only the opening and closing segments, achieving 58.33 $\mathcal{J}$, 61.98 $\mathcal{F}$, and 60.16 $\mathcal{J}\&\mathcal{F}$ on the CaReVOS test split.

2 Related work

Referring and reasoning video object segmentation. Referring video object segmentation links language expressions to object masks across a video (Seo et al., 2020; Botach et al., 2022). ReferFormer (Wu et al., 2022) uses language-conditioned queries to associate objects over time, while MeViS (Ding et al., 2023) emphasizes motion expressions that distinguish objects through their actions and temporal relationships. Reasoning segmentation extends this formulation to implicit queries, for which identifying the target requires interpreting context and commonsense knowledge. LISA (Lai et al., 2024) connects language reasoning to pixel prediction in images. VISA (Yan et al., 2024) and VideoLISA (Bai et al., 2024) develop this capability for videos by combining multimodal language understanding with mask prediction and temporal object association. GLUS (Lin et al., 2025) combines global and local context, InstructSeg (Wei et al., 2025) unifies instructed image and video segmentation, and ReferDINO (Liang et al., 2025) strengthens visual grounding. CaReVOS instead asks models to infer event participants from observed state changes then segment them.

Causal and event reasoning in videos. AGQA (Grunde-McLaughlin et al., 2021) and MMR-V (Zhu et al., 2026) test compositional and multi-frame video reasoning, while NExT-QA (Xiao et al., 2021) focuses on causal and temporal action questions with textual answers. To study the physical causes of events, CLEVRER (Yi et al., 2020) examines collision-based explanation and counterfactuals, while DynSuperCLEVR (Wang et al., 2025) tests physical dynamics. LLCP (Chen et al., 2024) extends such reasoning to accident attribution and counterfactual prediction. Moving from textual answers to visual grounding, GroundMoRe (Deng et al., 2025) requires object masks for causal, sequential, and counterfactual questions in real videos. Other work asks about events beyond the observed frames: FeVOS (Lan et al., 2026) identifies objects through future interactions, while next-event prediction learns from future segments (Wang et al., 2026a). For missing events, TEMPURA (Cheng et al., 2026) reconstructs masked events and their causal explanations. CaReVOS connects missing-event inference to object grounding by asking models to select participants from the setup and outcome, then segment them throughout the video.

Policy optimization and on-policy distillation. Reward-based optimization trains reasoning models with feedback on their own outputs (Shao et al., 2024; Liu et al., 2026; Gong et al., 2026; Zhang et al., 2026c;a; Dai et al., 2026a). Knowledge distillation offers another training signal by comparing teacher and student output distributions (Gu et al., 2024; Ko et al., 2024). On-policy variants apply this signal to student-sampled continuations (Agarwal et al., 2024), so the teacher guides the model on its own predictions. The teacher can further receive privileged information (Vapnik & Vashist, 2009), such as reference evidence withheld from the student (Zhao et al., 2026; Penaloza et al., 2026). Video-OPSD (Wang et al., 2026c) and Clue-OPSD (Wang et al., 2026b) use selected visual evidence as the teacher's privileged input while the student reads the full video. In CaReVOS the teacher instead watches the complete event while the student sees only the opening and closing segments. The model must also ground the inferred answer as full-video object masks.

3 The CaReVOS benchmark

CaReVOS evaluates video reasoning segmentation when the central interaction is withheld from target inference. Given a reasoning expression and the opening and closing segments of a video, a model must identify the target object or object set and segment it throughout the full clip.

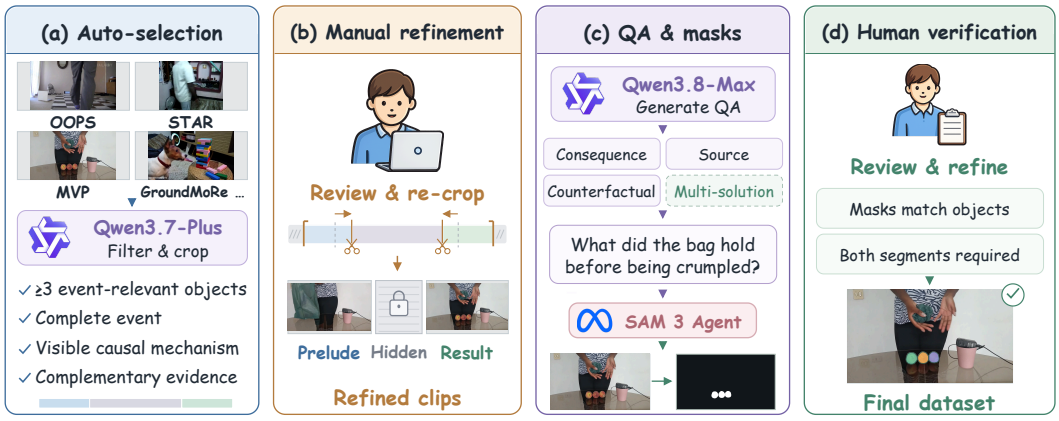

Figure 2: CaReVOS data construction. Automatic selection and manual refinement produce event clips. Human review then checks the generated expressions, answer objects, and mask trajectories.

3.1 Task definition

Given a video $V = (I_1, \ldots, I_T)$ and a reasoning expression q, the model identifies the objects that answer q from two visible segments. We denote their frame indices by $\mathcal{O} = \{1, \ldots, a\} \cup \{b, \ldots, T\}$, where $a + 1 < b$. The intervening frames contain the unobserved event relevant to the query. Target inference and mask prediction are written as

$$z = r(q, V_{\mathcal{O}}), \qquad \widehat{M}_{1:T} = s(V, z), \tag{1}$$

where z specifies the targets inferred solely from q and $V_{\mathcal{O}}$. The segmenter then localizes and tracks these objects throughout V. Frames outside $\mathcal{O}$ must not be used to alter the inferred target identity.

Each expression refers to a physical object or a complete set of objects sharing the queried role. For reference targets $\mathcal{A}_q$, the answer is their mask union $M_t = \bigcup_{k \in \mathcal{A}_q} M_t^k$. Evaluation is performed on all annotated frames, including frames with empty ground-truth masks when the targets are absent.

3.2 Dataset construction

Figure 2 summarizes the four-stage construction pipeline of CaReVOS. First, automatic selection identifies candidate events, and manual refinement then sets the observed and hidden segment boundaries so the opening and closing observations provide complementary evidence for inferring the event. With these boundaries in place, we generate reasoning expressions and annotate full-video object mask trajectories. Finally, reviewers check that each answer requires evidence from both observed segments and that the masks follow the intended objects through occlusions and interactions.

Stage 1: Automatic selection. We collect candidate videos from STAR (Wu et al., 2021), Oops! (Epstein et al., 2020), EPIC-KITCHENS (Damen et al., 2022), VOST (Tokmakov et al., 2023), NExT-QA (Xiao et al., 2021), GroundMoRe (Deng et al., 2025), and Minimal Video Pairs (Krojer et al., 2025). These sources cover everyday interactions, unexpected outcomes, and changes in object states. We use Qwen3.7-Plus (Qwen Team, 2026b) to screen and crop the videos, producing candidate clips that contain an initial scene, a complete interaction, and its outcome. Selection focuses on events with causal structure and multi-object interactions, where the roles of the event participants can be inferred by comparing the initial and final states. The prompt used for screening and cropping appears in Appendix E.

Stage 2: Manual refinement. Reviewers then adjust the opening, hidden, and closing segments so the first establishes the participants and the last reveals the outcome. They retain clips when the two observations provide complementary evidence about the answer. They revise or remove clips whose target is clear from either segment alone or whose explanation requires unavailable evidence.

Statistic	Value
Video clips	1,432
Train / test clips	1,146 / 286
Expressions	6,386
Train / test expressions	5,109 / 1,277
Object tracks	4,376
Object masks	373,828
Mean duration	16.7 s
Median duration	14.2 s
Duration range	7.0–85.0 s
Words / expression	10.0
Expressions / clip	4.46
Multi-solution queries	387 (6.1%)
Annotation rate	5 fps

Table 1: CaReVOS statistics. Expression lengths and counts per clip are averages.

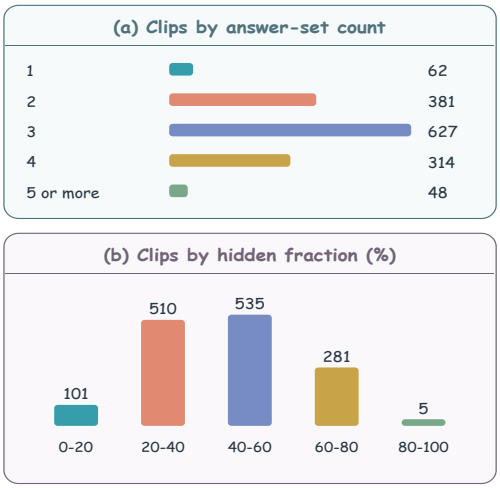

Figure 3: Distinct answer sets within each clip (a) and the fraction of its frames hidden (b).

Stage 3: Expression and mask annotation. We then use Qwen3.8-Max (Qwen Team, 2026c) to generate candidate questions and answers using the prompt in Appendix E. We group the questions into four categories: consequence, source, multi-solution, and counterfactual. Consequence questions identify objects affected by an event, whereas source questions trace an observed outcome to the objects that caused or enabled it. Multi-solution questions require identifying all valid targets among several candidates and may involve any of the other reasoning forms. Counterfactual questions ask whose outcome would change under a different action or condition. To ground these answers, we use Qwen3.8-Max with SAM 3 (Carion et al., 2026) to annotate the target masks.

Stage 4: Human verification. Reviewers first check each question against the complete event and the segment boundaries. They then verify that the answer can be inferred from the two observed segments together, but not from either segment alone. Based on these checks, they revise ambiguous wording, correct target-object assignments, and remove questions that reveal the answer or rely on event relations unsupported by the video evidence. For multi-object answers, reviewers also verify that the complete target set is included. With the targets confirmed, they inspect and correct the masks for object identity, spatial accuracy, and temporal consistency through occlusions and interactions. We annotate each clip at 5 fps, including empty masks when the target is absent.

3.3 Dataset statistics and splits

Table 1 summarizes CaReVOS, including its mean clip duration of 16.7 seconds and mean expression length of 10.0 words. Figure 3 complements these averages with the distributions of answer-set counts and hidden fractions. Most clips contain two to four distinct answer sets, so questions can test different target choices within the same scene. These choices depend on the queried role and the link between setup and outcome, with 44.3% of each clip hidden from the reasoner on average.

We split the dataset by original source video into training and test sets at an approximately 4:1 ratio, balancing source collections and question types. All clips derived from the same original video remain in one partition. Appendix A provides the source composition and split statistics.

4 Method

To address CaReVOS, we propose Event-OPD, a two-stage training framework shown in Figure 4. First, supervised fine-tuning (SFT) teaches the model to infer target objects from the opening and closing segments and segment them throughout the video. Privileged on-policy distillation then lets a frozen full-video teacher guide the student on its own sampled answers.

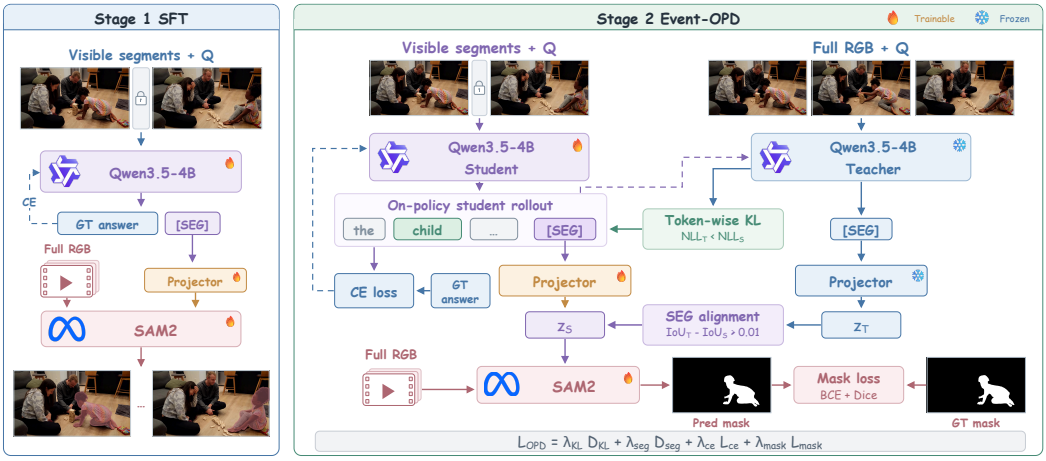

Figure 4: Overview of Event-OPD. SFT supervises answers from opening and closing segments and masks over the full video. A frozen full-video teacher then guides student-sampled answers through token-level KL and SEG alignment.

4.1 Segmentation Framework

We build on Sa2VA (Yuan et al., 2026), which connects a vision-language reasoner to a video segmenter through a segmentation token. We use Qwen3.5-4B (Qwen Team, 2026a) as the reasoner and SAM2 Hiera-L (Ravi et al., 2025) as the segmenter. Given $x = (q, V_{\mathcal{O}})$, the reasoner generates an answer y ending with the segmentation token. A learned projector maps this token's hidden state h_{SEG} to z, implementing the target representation defined in Section 3.1. The segmenter then predicts masks throughout the complete video:

$$y \sim \pi_\theta(\cdot \mid x), \qquad z = g_\phi(h_{\text{SEG}}), \qquad \widehat{M}_{1:T} = s_\psi(V, z). \tag{2}$$

Here, π_θ denotes the reasoner's autoregressive distribution, while g_ϕ and s_ψ denote the projector and video segmenter. Target inference relies only on the observed segments to determine the queried object set, while mask prediction uses the full video to localize and track the inferred targets over time. For multi-object answers, we supervise the union of their masks.

4.2 Supervised Fine-Tuning

Starting from the base vision-language model and a pretrained segmenter, we use supervised training to connect reference answers with their target masks. For a reference answer y^*, we minimize the mean token prediction loss

$$\mathcal{L}_{\text{text}} = -\frac{1}{|y^*|} \sum_{u=1}^{|y^*|} \log \pi_\theta(y^*_u \mid x, y^*_{<u}). \tag{3}$$

To supervise segmentation, we use the segmentation-token state produced while replaying y^* through the reasoner. This connects the reference answer to its target masks through the learned projector. We combine the text objective with binary cross-entropy (BCE) and Dice losses on the predicted masks, as in prior reasoning segmentation models (Lai et al., 2024; Bai et al., 2024):

$$\mathcal{L}_{\text{SFT}} = \mathcal{L}_{\text{text}} + \frac{1}{T} \sum_{t=1}^{T} \left[\lambda_{\text{BCE}} \mathcal{L}_{\text{BCE}}(\widehat{M}_t, M_t) + \lambda_{\text{Dice}} \mathcal{L}_{\text{Dice}}(\widehat{M}_t, M_t) \right]. \tag{4}$$

The text loss teaches the reasoner to generate the answer, while the mask losses supervise the corresponding objects in each frame. The resulting model provides the starting point for distillation. We provide the loss weights and optimization settings in Appendix B.

4.3 PRIVILEGED ON-POLICY DISTILLATION

SFT trains on reference answers, but at inference the model continues from its own predictions. We therefore distill on student-generated prefixes, allowing the teacher to guide the model under the contexts it encounters at inference (Agarwal et al., 2024).

We initialize the student from the SFT model and freeze a copy as its teacher. The teacher receives the complete video V, allowing it to use the interaction hidden from the student's input $V_{\mathcal{O}}$. For each response y sampled by the student, both models compute next-token distributions conditioned on its prefixes under their respective inputs, $x^+ = (q, V)$ and $x = (q, V_{\mathcal{O}})$ (Zhao et al., 2026; Penaloza et al., 2026). We combine token KL and SEG alignment with answer and mask supervision:

$$\mathcal{L}_{\text{OPD}} = \lambda_{\text{KL}}\mathcal{D}_{\text{KL}} + \lambda_{\text{seg}}\mathcal{D}_{\text{seg}} + \lambda_{\text{ce}}\,\mathcal{L}_{\text{ce}} + \lambda_{\text{mask}}\,\mathcal{L}_{\text{mask}}. \tag{5}$$

The first term is the token-level forward KL between the next-token distributions over the full vocabulary, with the teacher distribution detached:

$$\mathcal{D}_{\text{KL}} = \mathbb{E}_{y\sim\pi_\theta(\cdot|x)}\Big[\frac{1}{|y|}\sum_{u=1}^{|y|}\sum_{v}\pi_T(v \mid x^+, y_{<u})\log\frac{\pi_T(v \mid x^+, y_{<u})}{\pi_\theta(v \mid x, y_{<u})}\Big]. \tag{6}$$

To link the teacher's answer-level guidance to segmentation, we project the student and teacher states at the segmentation token of the same sampled response into the segmenter's prompt space, obtaining z_S and z_T. Their cosine distance, $\mathcal{D}_{\text{seg}} = 1 - \cos(z_S, z_T)$, computed with gradients stopped through z_T, encourages the student to form an object representation close to the teacher's.

To keep these updates tied to the annotated answer and targets, we also replay the reference answer y^* through the student reasoner. Its tokens supply the cross-entropy loss $\mathcal{L}_{\text{ce}}$, while its projected segmentation state prompts the segmenter for $\mathcal{L}_{\text{mask}}$, the weighted BCE and Dice losses in Eq. 4. We apply each distillation term only when the teacher performs better on its corresponding answer or mask measure. We provide the detailed activation conditions and loss weights in Appendix B.

5 EXPERIMENTS

5.1 IMPLEMENTATION DETAILS

We use Qwen3.5-4B (Qwen Team, 2026a) for reasoning and SAM2 Hiera-L (Ravi et al., 2025) for segmentation. We train the language LoRA (rank 128, scaling 256), projector, and SAM2 mask decoder, while freezing the visual encoder and temporal memory. We sample only the opening and closing frames at 1 fps for reasoning, whereas mask supervision covers every frame. We first train with SFT for 2 epochs on 4 A800 GPUs at a learning rate of 2×10^{-5}. We then run Event-OPD for 2 epochs with the same settings. We cap responses at 256 tokens during distillation and 2048 at test time. For comparison, we fine-tune six external baselines on the CaReVOS training split. We evaluate all methods on the CaReVOS test split using $\mathcal{J}$, $\mathcal{F}$, and their mean, averaging over frames and then questions following the DAVIS measures (Perazzi et al., 2016). For additional details on training and evaluation, please refer to Appendix B.

5.2 MAIN RESULTS

We compare Event-OPD with existing video segmentation methods in Table 2, under both zero-shot transfer and training on CaReVOS. Without CaReVOS training, VideoSEG-O3 and FeVOS-R1 lead at 47.58 and 46.21 $\mathcal{J}\&\mathcal{F}$. Yet their performance suggests that current reasoning segmentation models still struggle to infer the queried target from an event's setup and outcome alone.

Training on CaReVOS helps close this gap, improving all six external methods by 9.38 to 24.08 points. These consistent gains indicate that CaReVOS supervision strengthens their ability to reason about event roles and identify the corresponding objects from partial observations.

Against these improved baselines, Event-OPD achieves the best result at 60.16 $\mathcal{J}\&\mathcal{F}$, surpassing FeVOS-R1 by 2.21 points with gains in both $\mathcal{J}$ and $\mathcal{F}$. This result demonstrates the effectiveness of our training framework for jointly reasoning about causal events and localizing the corresponding objects. We provide further analysis by reasoning category and source dataset in Appendix C.

Table 2: CaReVOS test results. Best scores are bold. ‡ marks XMem++ adaptation.

Method	Backbone	$\mathcal{J}\uparrow$	$\mathcal{F}\uparrow$	$\mathcal{J}\&\mathcal{F}\uparrow$
Zero-shot				
SPARROW (Alansari et al., 2026)	VideoGPT-3.8B	17.15	22.33	19.74
ReferFormer (Wu et al., 2022)‡	ResNet-50	17.45	22.29	19.87
LMPM (Ding et al., 2023)‡	Swin-T	18.59	22.22	20.41
UniRef++ (Wu et al., 2023)‡	ResNet-50	20.36	25.44	22.90
GLUS (Lin et al., 2025)	LLaVA-7B	22.84	25.65	24.25
VideoLISA (Bai et al., 2024)	LLaVA-Phi-3.8B	22.01	27.28	24.64
ReVIOSa-4B (Jin et al., 2026)	InternVL2.5-4B	25.85	30.02	27.93
VRS-HQ (Gong et al., 2025)	Chat-UniVi-7B	27.71	31.50	29.60
Sa2VA (Yuan et al., 2026)	InternVL2.5-4B	28.71	32.59	30.65
InstructSeg (Wei et al., 2025)	Mipha-3B	28.76	33.09	30.92
MomentSeg (Dai et al., 2026b)	Qwen2.5-VL-3B	29.35	33.40	31.37
Sa2VA-i (Nekrasov et al., 2025)	InternVL2.5-4B	30.56	33.02	31.79
Veason-R1 (Gong et al., 2026)	Qwen2.5-VL-3B	29.68	35.27	32.47
FeVOS-R1 (Lan et al., 2026)	InternVL2.5-4B	43.40	49.03	46.21
VideoSEG-O3 (Dai et al., 2026a)	Qwen3-VL-4B	**44.98**	**50.18**	**47.58**
Trained on CaReVOS				
InstructSeg (Wei et al., 2025) (SFT)	Mipha-3B	40.87	45.33	43.10
GLUS (Lin et al., 2025) (SFT)	LLaVA-7B	42.43	47.56	45.00
Sa2VA (Yuan et al., 2026) (SFT)	InternVL2.5-4B	51.08	56.57	53.82
MomentSeg (Dai et al., 2026b) (SFT)	Qwen2.5-VL-3B	52.62	58.28	55.45
VideoSEG-O3 (Dai et al., 2026a) (SFT)	Qwen3-VL-4B	54.36	59.56	56.96
FeVOS-R1 (Lan et al., 2026) (SFT)	InternVL2.5-4B	55.04	60.86	57.95
Event-OPD (ours)	Qwen3.5-4B	**58.33**	**61.98**	**60.16**

Table 3: Training strategy ablation.

Strategy	$\mathcal{J}\uparrow$	$\mathcal{F}\uparrow$	$\mathcal{J}\&\mathcal{F}\uparrow$
SFT only	56.58	60.49	58.54
OPD + SFT	54.24	58.73	56.49
SFT + GRPO	54.91	59.06	56.98
GT-prefix KD	56.72	60.60	58.66
Event-OPD (ours)	**58.33**	**61.98**	**60.16**

Table 4: Distillation component ablation.

Variant	$\mathcal{J}\uparrow$	$\mathcal{F}\uparrow$	$\mathcal{J}\&\mathcal{F}\uparrow$
w/o $\mathcal{D}_{\mathrm{KL}}$	56.90	60.57	58.73
w/o $\mathcal{D}_{\mathrm{seg}}$	57.69	61.53	59.61
w/o $\mathcal{L}_{\mathrm{ce}}, \mathcal{L}_{\mathrm{mask}}$	56.18	59.98	58.08
w/o full video	57.36	61.24	59.30
Event-OPD (ours)	**58.33**	**61.98**	**60.16**

5.3 Ablations and analysis

We study the training strategies in Table 3 to examine how the two stages connect reasoning and segmentation. We first apply QA-only OPD before segmentation SFT, which reaches 56.49 $\mathcal{J}\&\mathcal{F}$, 2.05 points below SFT alone. This suggests a gap between reasoning and object grounding: QA-only distillation guides answer generation without directly supervising the object representation used for mask prediction. We therefore compare ways to improve the supervised model. GRPO reaches 56.98, while GT-prefix KD achieves 58.66, only 0.12 above the supervised baseline. The latter reuses annotated prefixes, leaving the teacher little opportunity to guide continuations arising from the student's own predictions. Event-OPD instead trains on these continuations and reaches 60.16. This supports first establishing the answer-to-mask connection, then refining the student's reasoning on its own continuations, where an early prediction can change the context for later decisions.

To test each training signal, we remove loss terms from Event-OPD in Table 4. Removing token KL lowers $\mathcal{J}\&\mathcal{F}$ to 58.73, suggesting that SEG alignment cannot replace the teacher's guidance on how an answer develops. Conversely, removing SEG cosine yields 59.61. While token KL guides

the answer, SEG alignment directly shapes the object representation passed to the segmenter. The drop suggests that aligning this representation helps translate the inferred target into accurate masks. Finally, removing the answer and mask supervision terms $\mathcal{L}_{\text{ce}}$ and $\mathcal{L}_{\text{mask}}$ lowers the score to 58.08, below SFT alone. The larger drop suggests that teacher agreement alone does not preserve the answer-to-mask connection. Direct supervision therefore keeps the answer and its spatial grounding tied to the same annotated target while distillation refines the model's event reasoning.

Beyond the loss design, we ask whether the teacher needs to observe the hidden event to provide useful guidance. Restricting it to the student's opening and closing frames gives 59.30, 0.86 below Event-OPD but 0.76 above SFT alone. Thus, distillation still helps when both models receive the same evidence. The further gain from the complete video suggests that observing the interaction helps the teacher guide how the student connects its setup and outcome. We provide additional ablations on teacher freezing, observed segments, and the choice of distillation loss in Appendix C.

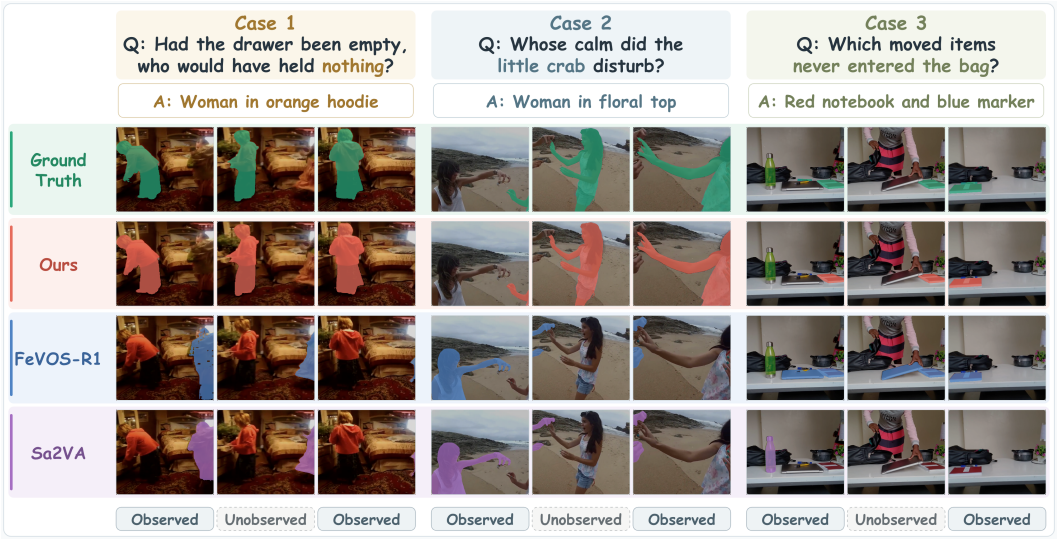

Figure 5: Qualitative comparison on CaReVOS.

5.4 Qualitative Results

We provide qualitative comparisons with FeVOS-R1 and Sa2VA in Figure 5 to examine how the methods select targets from event context. In the first case, both baselines follow the wrong person, while Event-OPD correctly selects the woman in the orange hoodie for the counterfactual query. A similar confusion appears in the second case: the baselines segment the person holding the crab, whereas Event-OPD segments the disturbed woman through the interval unseen by the reasoner.

The third case further requires selecting multiple objects. FeVOS-R1 includes an extra object and Sa2VA misses the required pair, while Event-OPD segments both the red notebook and blue marker. These differences illustrate the need to distinguish participants by their event roles and identify all objects that satisfy the question. Coherent masks alone cannot correct a mistaken target choice. We provide additional failure cases in Figure 8 and discuss the remaining challenges in Appendix C.

6 Conclusion

We introduced CaReVOS to study how models infer event participants from the setup and outcome of an unseen interaction, then segment them throughout the video. Since a coherent mask can still correspond to the wrong participant, this task requires both event reasoning and accurate localization. To improve these abilities together, we developed Event-OPD, which combines supervised training with guidance from a teacher that observes the full event. This guidance connects the student's own responses to the object representations used for segmentation. Our experiments show that this approach improves both target selection and mask accuracy under partial observation.

AI USE STATEMENT

We used generative AI tools to improve the clarity of our writing and assist with dataset construction. Section 3.2 describes how we used these tools for clip selection and annotation.

ETHICS STATEMENT

We built CaReVOS from publicly available video datasets for non-commercial research. As part of this process, the authors carried out all manual annotation and verification.

REPRODUCIBILITY STATEMENT

We describe CaReVOS in Section 3 and give the splits, evaluation protocol, and training settings in Appendices A and B. In addition, we plan to release code and annotations upon acceptance.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, 2024.

Mohamad Alansari, Naufal Suryanto, Divya Velayudhan, Sajid Javed, Naoufel Werghi, and Muzammal Naseer. SPARROW: Learning Spatial Precision and Temporal Referential Consistency in Pixel-Grounded Video MLLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17545–17556, 2026.

Zechen Bai, Tong He, Haiyang Mei, Pichao Wang, Ziteng Gao, Joya Chen, Lei Liu, Zheng Zhang, and Mike Zheng Shou. One Token to Seg Them All: Language Instructed Reasoning Segmentation in Videos. In *Advances in Neural Information Processing Systems*, volume 37, pp. 6833–6859, 2024.

Maksym Bekuzarov, Ariana Bermudez, Joon-Young Lee, and Hao Li. XMem++: Production-level Video Segmentation From Few Annotated Frames. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 635–644, 2023.

Adam Botach, Evgenii Zheltonozhskii, and Chaim Baskin. End-to-End Referring Video Object Segmentation with Multimodal Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4985–4995, 2022.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, RISHI HAZRA, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollar, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment Anything with Concepts. In *International Conference on Learning Representations*, 2026.

Guangyi Chen, Yuke Li, Xiao Liu, Zijian Li, Eman Al Suradi, Donglai Wei, and Kun Zhang. LLCP: Learning Latent Causal Processes for Reasoning-based Video Question Answer. In *International Conference on Learning Representations*, 2024.

Jen-Hao Cheng, Yi-Hao Peng, Huapeng Zhou, Vivian Wang, Huayu Wang, Hsiang-Wei Huang, Wenhao Chai, Hou-I Liu, Kuang-Ming Chen, Cheng-Yen Yang, Yi-Ling Chen, Vibhav Vineet, Qin Cai, and Jenq-Neng Hwang. TEMPURA: Temporal event masked prediction and understanding for reasoning in action. In *Third Conference on Language Modeling*, 2026.

Ming Dai, Sen Yang, Boqiang Duan, Boyuan Tong, Jiedong Zhuang, Wankou Yang, and Jingdong Wang. VideoSEG-O3: A Multi-turn Reinforcement Learning Framework for Reasoning Video Object Segmentation. In *International Conference on Machine Learning*, 2026a.

Ming Dai, Sen Yang, Boqiang Duan, Wankou Yang, and Jingdong Wang. MomentSeg: Moment-Centric Sampling for Enhanced Referring Video Object Segmentation. In *European Conference on Computer Vision*, pp. 413–431, 2026b.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling Egocentric Vision: Collection, Pipeline and Challenges for EPIC-KITCHENS-100. *International Journal of Computer Vision*, 130:33–55, 2022.

Andong Deng, Tongjia Chen, Shoubin Yu, Taojiannan Yang, Lincoln Spencer, Yapeng Tian, Ajmal Saeed Mian, Mohit Bansal, and Chen Chen. Motion-Grounded Video Reasoning: Understanding and Perceiving Motion at Pixel Level. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8625–8636, 2025.

Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Chen Change Loy. MeViS: A Large-scale Benchmark for Video Segmentation with Motion Expressions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2694–2703, 2023.

Dave Epstein, Boyuan Chen, and Carl Vondrick. Oops! Predicting Unintentional Action in Video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 919–929, 2020.

Sitong Gong, Yunzhi Zhuge, Lu Zhang, Zongxin Yang, Pingping Zhang, and Huchuan Lu. The Devil is in Temporal Token: High Quality Video Reasoning Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29183–29192, 2025.

Sitong Gong, Yunzhi Zhuge, Lu Zhang, Jiazuo Yu, Pingping Zhang, Xu Jia, and Huchuan Lu. Reinforcing Video Object Segmentation to Think before it Segments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3835–3844, 2026.

Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. AGQA: A Benchmark for Compositional Spatio-Temporal Reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11287–11297, 2021.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In *International Conference on Learning Representations*, 2024.

Woojeong Jin, Seongchan Kim, Jaeho Lee, and Seungryong Kim. InterRVOS: Interaction-Aware Referring Video Object Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10367–10376, 2026.

Shiu-Hong Kao, Yu-Wing Tai, and Chi-Keung Tang. Cot-rvs: Zero-shot chain-of-thought reasoning segmentation for videos. In *International Conference on Learning Representations*, 2026.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. Distillm: Towards streamlined distillation for large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 24872–24895, 2024.

Benno Krojer, Mojtaba Komeili, Candace Ross, Quentin Garrido, Koustuv Sinha, Nicolas Ballas, and Mahmoud Assran. A Shortcut-aware Video-QA Benchmark for Physical Understanding via Minimal Video Pairs. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. LISA: Reasoning Segmentation via Large Language Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9579–9589, 2024.

Kehan Lan, Kaining Ying, and Henghui Ding. FeVOS: Foresight Expression Video Object Segmentation. In *European Conference on Computer Vision*, pp. 342–359, 2026.

Tianming Liang, Kun-Yu Lin, Chaolei Tan, Jianguo Zhang, Wei-Shi Zheng, and Jian-Fang Hu. ReferDINO: Referring Video Object Segmentation with Visual Grounding Foundations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20009–20019, 2025.

Lang Lin, Xueyang Yu, Ziqi Pang, and Yu-Xiong Wang. GLUS: Global-Local Reasoning Unified into A Single Large Language Model for Video Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8658–8667, 2025.

Yuqi Liu, Tianyuan Qu, Zhisheng Zhong, Bohao Peng, Shu Liu, Bei Yu, and Jiaya Jia. VisionReasoner: Unified Reasoning-Integrated Visual Perception via Reinforcement Learning. In *International Conference on Learning Representations*, 2026.

Alexey Nekrasov, Ali Athar, Daan de Geus, Alexander Hermans, and Bastian Leibe. Sa2VA-i: Improving Sa2VA Results with Consistent Training and Inference. *arXiv preprint arXiv:2509.19082*, 2025. URL https://arxiv.org/abs/2509.19082.

Emiliano Penaloza, Dheeraj Vattikonda, Nicolas Gontier, Alexandre Lacoste, Laurent Charlin, and Massimo Caccia. Privileged information distillation for language models. In *International Conference on Machine Learning*, 2026.

Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 724–732, 2016.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026a. URL https://qwen.ai/blog?id=qwen3.5.

Qwen Team. Qwen3.7: The agent frontier, May 2026b. URL https://qwen.ai/blog?id=qwen3.7.

Qwen Team. Qwen3.8-max: A new bar for coding and cowork, August 2026c. URL https://qwen.ai/blog?id=qwen3.8.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. In *International Conference on Learning Representations*, 2025.

Seonguk Seo, Joon-Young Lee, and Bohyung Han. URVOS: Unified Referring Video Object Segmentation Network with a Large-Scale Benchmark. In *European Conference on Computer Vision*, pp. 208–223, 2020.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. URL https://arxiv.org/abs/2402.03300.

Pavel Tokmakov, Jie Li, and Adrien Gaidon. Breaking the "Object" in Video Object Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22836–22845, 2023.

Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural Networks*, 22(5-6):544–557, 2009.

Haonan Wang, Hongfu Liu, Xiangyan Liu, Chao Du, Kenji Kawaguchi, Ye Wang, and Tianyu Pang. Fostering Video Reasoning via Next-Event Prediction. In *International Conference on Learning Representations*, 2026a.

Kaishen Wang, Dongdi Zhao, Yijun Liang, Dingqiang Ye, Ruibo Chen, Heng Huang, and Di Fu. Where to look matters: On-policy self-distillation for long-video understanding, 2026b. URL https://arxiv.org/abs/2608.25356.

Xingrui Wang, Wufei Ma, Angtian Wang, Shuo Chen, Adam Kortylewski, and Alan Yuille. Compositional 4D Dynamic Scenes Understanding with Physics Priors for Video Question Answering. In *International Conference on Learning Representations*, 2025.

Ziyue Wang, Shiqi Huang, Weiwen Xu, Bihan Wen, and Xudong Jiang. Video-opsd: Exploiting privileged visual evidence for on-policy self-distillation in video large language models, 2026c. URL https://arxiv.org/abs/2608.27065.

Cong Wei, Yujie Zhong, Haoxian Tan, Yingsen Zeng, Yong Liu, Hongfa Wang, and Yujiu Yang. InstructSeg: Unifying Instructed Visual Segmentation with Multi-modal Large Language Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20193–20203, 2025.

Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. STAR: A Benchmark for Situated Reasoning in Real-World Videos. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021.

Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. Language as Queries for Referring Video Object Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4974–4984, 2022.

Jiannan Wu, Yi Jiang, Bin Yan, Huchuan Lu, Zehuan Yuan, and Ping Luo. UniRef++: Segment Every Reference Object in Spatial and Temporal Spaces. *arXiv preprint arXiv:2312.15715*, 2023. URL https://arxiv.org/abs/2312.15715.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. NExT-QA: Next Phase of Question-Answering to Explaining Temporal Actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9777–9786, 2021.

Cilin Yan, Haochen Wang, Shilin Yan, Xiaolong Jiang, Yao Hu, Guoliang Kang, Weidi Xie, and Efstratios Gavves. VISA: Reasoning Video Object Segmentation via Large Language Models. In *European Conference on Computer Vision*, pp. 98–115, 2024.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. CLEVRER: Collision Events for Video Representation and Reasoning. In *International Conference on Learning Representations*, 2020.

Haobo Yuan, Xiangtai Li, Tao Zhang, Yueyi Sun, Zilong Huang, Shilin Xu, Shunping Ji, Yunhai Tong, Lu Qi, Jiashi Feng, and Ming-Hsuan Yang. Sa2VA: Marrying SAM2 with MLLM for Dense Grounded Understanding of Images and Videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–14, 2026. doi: 10.1109/TPAMI.2026.3720379.

Congzhi Zhang, Zhibin Wang, Yinchao Ma, Jiawei Peng, Yihan Wang, Qiang Zhou, Jun Song, and Bo Zheng. Rewatch-r1: Boosting complex video reasoning in large vision-language models through agentic data synthesis. In *International Conference on Learning Representations*, 2026a.

Mingfang Zhang, Jingjing Pan, Ashutosh Kumar, Rajat Saini, Mustafa Erdogan, Hsuan-Kung Yang, Caixin Kang, Yifei Huang, Yoichi Sato, and Quan Kong. CaST-Bench: Benchmarking Causal Chain-Grounded Spatio-Temporal Reasoning for Video Question Answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 31856–31866, 2026b.

Xiaowen Zhang, Zhi Gao, Licheng Jiao, Lingling Li, and Qing Li. Stvg-r1: Incentivizing instance-level reasoning and grounding in videos via reinforcement learning. In *International Conference on Learning Representations*, 2026c.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. In *International Conference on Machine Learning*, 2026.

Kejian Zhu, Zhuoran Jin, Hongbang Yuan, Jiachun Li, Shangqing Tu, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. MMR-V: What's Left Unsaid? A Benchmark for Multimodal Deep Reasoning in Videos. In *International Conference on Learning Representations*, 2026.

A Dataset and annotation details

Source composition and data splits. Table 5 reports the retained clips and expressions from each source collection. Table 6 breaks down question categories and annotations by partition. Together, the splits contain 119,257 annotated frames, or 6.63 hours at 5 fps.

Table 5: Source composition of CaReVOS.

Source	Clips	Expressions
VOST (Tokmakov et al., 2023)	48	205
GroundMoRe (Deng et al., 2025)	74	352
EPIC-KITCHENS (Damen et al., 2022)	125	451
NExT-QA (Xiao et al., 2021)	137	586
Minimal Video Pairs (Krojer et al., 2025)	177	812
STAR (Wu et al., 2021)	409	1,827
Oops! (Epstein et al., 2020)	462	2,153
Total	1,432	6,386

Table 6: Question categories and annotations by split. Hidden percentages are averaged over clips.

Question type	Train	Test
Consequence	1,561	396
Source	1,671	404
Multi-solution	306	81
Counterfactual	1,571	396

Statistic	Train	Test
Object tracks	3,496	880
Annotated frames	95,547	23,710
Object masks	299,758	74,070
Hidden interval (%)	44.6	42.8

Composition by source and question type. Figure 6 shows how the question categories are distributed across source datasets. Although Oops! and STAR supply the most expressions, multi-solution questions account for only about 4% of each collection, compared with 19% in Minimal Video Pairs. Thus, the largest sources do not contribute equally to every form of reasoning.

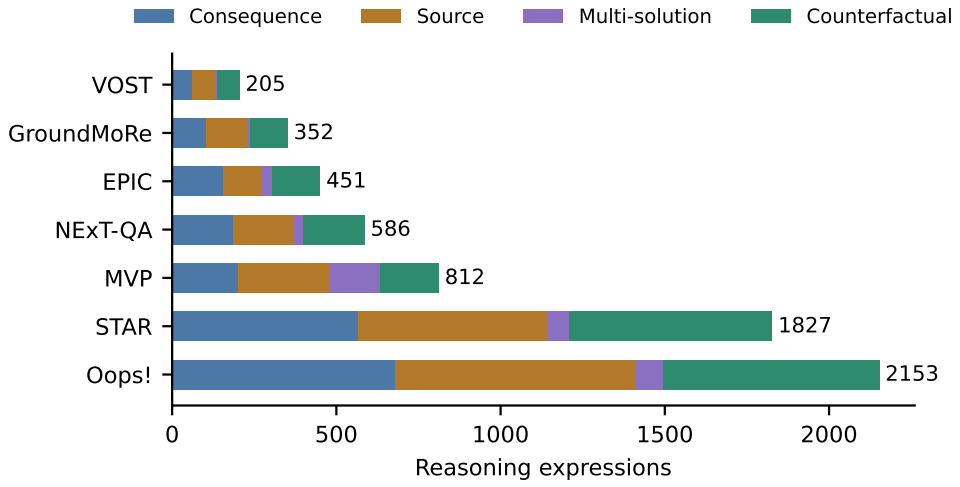

Figure 6: Distribution of reasoning expressions across source datasets.

Benchmark comparison. Table 7 compares CaReVOS with MeViS (Ding et al., 2023), ReVOS (Yan et al., 2024), GroundMoRe (Deng et al., 2025), and FeVOS (Lan et al., 2026). The first three provide frames of the event, while FeVOS provides only pre-event evidence. In contrast, CaReVOS asks models to connect observations before and after a hidden interaction, then evaluates masks throughout the video.

Table 7: Benchmark scale and temporal evaluation protocol.

Dataset	Videos	Queries	Masks (k)	Inference input	Mask output
MeViS	2,006	28,570	443	Full video	Full video
ReVOS	1,042	35,074	469	Full video	Full video
GroundMoRe	1,715	7,577	249	Full video	Event frames
FeVOS	968	14,525	37	Before event	Observed frames
CaReVOS	1,432	6,386	374	**Opening + closing**	**Full video**

Additional examples. We provide four additional examples in Figure 7. In A, the question names the dog's tug and asks who experiences its consequence. The dog pulls a toy beside the baby in the opening frames, while the closing frames show the baby on the floor, making the baby the target. B uses the same clip but reverses the inference: starting from the baby's fall, it asks whose pull caused it and selects the black dog. Thus, the question's causal direction changes the target even when the visual evidence is the same. An event question can also have several valid targets. In C, both the laptop and blue shirt move from the bag to the bed, so choosing only one misses part of the answer. While the first three examples ask about the actual event, D asks how a different condition would change its outcome. With an empty bottle, the woman would have stayed dry instead of getting wet.

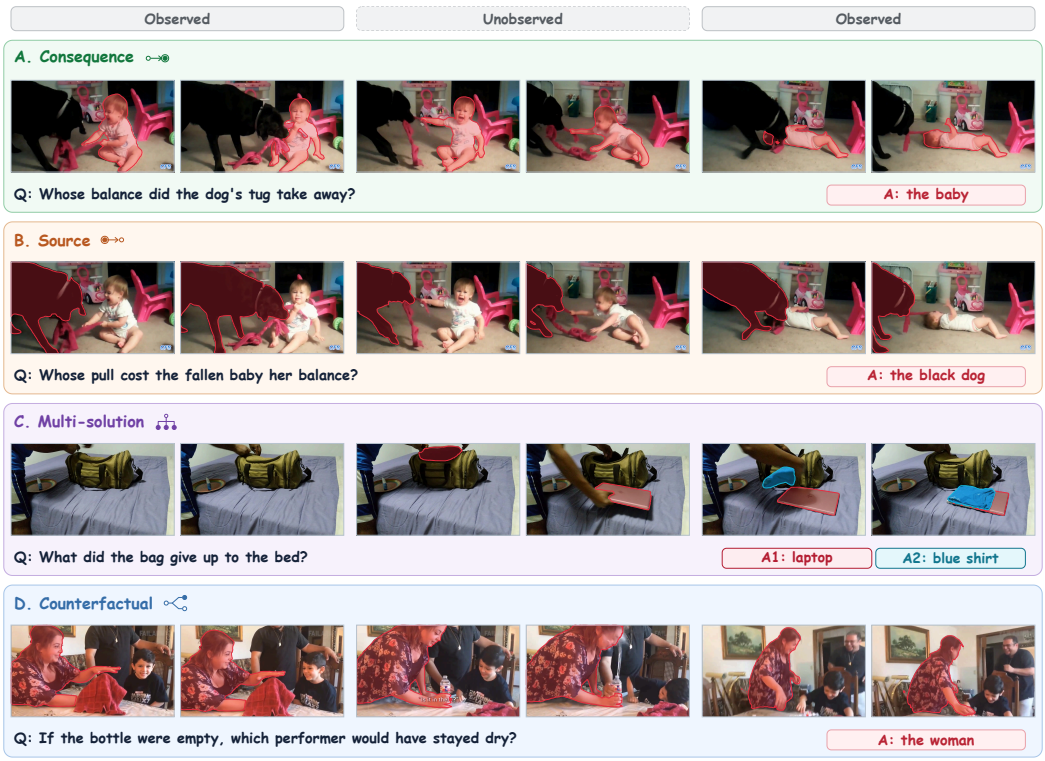

Figure 7: Additional examples of the four question categories with reference target masks.

B IMPLEMENTATION AND EVALUATION DETAILS

Input protocol. We annotate masks on a 5-fps timeline and sample the opening and closing segments independently for target reasoning. Within each segment, we take frames at 1 fps, include the first and last frames, and keep up to eight evenly spaced frames. For example, a 12-frame segment supplies frames 1, 6, 11, and 12. The reasoner receives only these frames and the question, so target inference excludes the hidden middle.

Once the target is chosen, the segmenter may use the full video to produce masks on the 5-fps timeline, including the middle. This separation follows the segmentation-token interface introduced by LISA (Lai et al., 2024) and extended to video by Sa2VA (Yuan et al., 2026). For baselines marked ‡ in Table 2, which cannot separate reasoning from full-video segmentation directly, we select a mask from their predictions on the observed frames and propagate it through the video with XMem++ (Bekuzarov et al., 2023). Here, hidden frames are used only to track the inferred targets.

Mask scoring. We score every frame of the complete clip, even when the target is absent. If a question has several valid targets, their union forms its reference mask. For a predicted mask M and reference mask G, we use the DAVIS region and contour measures (Perazzi et al., 2016):

$$\mathcal{J} = \frac{|M \cap G|}{|M \cup G|}, \qquad \mathcal{F} = \frac{2P_c R_c}{P_c + R_c}. \tag{7}$$

Here P_c and R_c are contour precision and recall. We average frame scores within each question and then across questions. The reported $\mathcal{J}\&\mathcal{F}$ is the mean of the resulting $\mathcal{J}$ and $\mathcal{F}$ scores.

Privileged on-policy distillation. We initialize the student from the SFT model and freeze a copy as the teacher. The student sees sampled opening and closing frames, whereas the teacher sees frames sampled across the complete video at 1 fps. To guide the student's own predictions, we first sample its answer y and replay the tokens through both models under their respective video inputs. At each prefix $y_{<u}$, the teacher supplies the next-token distribution for forward KL. If y contains a segmentation token, we also align the projected student and teacher states at that token through their cosine distance. Both distillation signals therefore guide the student on the answer it generated.

The GT-prefix KD baseline in Table 3 instead replays the annotated completion y^* for both distillation terms. Event-OPD uses the sampled y, while retaining separate student replays of y^* for answer cross-entropy and full-video mask supervision. Thus, distillation follows the student's own predictions, while supervision preserves the annotated answer and masks.

Teacher-advantage conditions. We use the annotated answer and masks to decide when each teacher signal contributes. Both models score the reference completion y^*, and the KL term is used only when the teacher's mean token NLL is lower than the student's. For the segmentation term, we decode the student and teacher states from the sampled y over the full video. We use their cosine loss only when the teacher's mean frame IoU exceeds the student's by more than 0.01. Inactive distillation terms contribute zero, while the answer and mask losses remain active.

The active terms use the following weights:

$$\mathcal{L}_{\mathrm{OPD}} = 2\mathcal{D}_{\mathrm{KL}} + 0.5\mathcal{D}_{\mathrm{seg}} + 0.1\mathcal{L}_{\mathrm{ce}} + 0.1\mathcal{L}_{\mathrm{mask}}. \tag{8}$$

Table 8 gives the remaining training settings.

Table 8: Additional settings for SFT and Event-OPD.

Setting	SFT	Event-OPD
Batch size	8	8
Loss terms	$\mathcal{L}_{\mathrm{text}}$ / $\mathcal{L}_{\mathrm{BCE}}$ / $\mathcal{L}_{\mathrm{Dice}}$	$\mathcal{D}_{\mathrm{KL}}$ / $\mathcal{D}_{\mathrm{seg}}$ / $\mathcal{L}_{\mathrm{ce}}$ / $\mathcal{L}_{\mathrm{mask}}$
Loss weights	1 / 2 / 0.5	2 / 0.5 / 0.1 / 0.1
Optimizer	AdamW	AdamW
Precision	BF16	BF16

Model prompts. We ask the student reasoner to infer the missing event from its setup and outcome and identify the objects that answer the question. To guide this inference, we ask the privileged teacher to use evidence from the complete event.

Student system prompt

You are the student reasoner in a video reasoning and segmentation task. I will give you a question and frames sampled from the opening and closing segments of a video.

The event between those segments is hidden from you. Use the visible setup and outcome to reason about what happened and identify which objects answer the question. Briefly describe the targets and end with [SEG] so they can be segmented throughout the video.

Privileged teacher system prompt

You are the privileged teacher in a video reasoning and segmentation task. I will give you a question and frames sampled across the video, including the event hidden from the student.

The student sees only the opening and closing segments. Infer the event from the additional frames and identify which objects answer the question. Use this evidence to guide the student's reasoning, then describe the targets and end with [SEG] for segmentation.

C Additional results and failure cases

Table 9: Test $\mathcal{J}\&\mathcal{F}$ by question category.

Model	Consequence	Source	Multi-solution	Counterfactual
VideoSEG-O3 (SFT)	56.56	57.83	39.98	59.94
FeVOS-R1 (SFT)	56.80	**59.73**	44.70	59.98
Sa2VA (SFT)	52.70	56.57	40.62	54.85
MomentSeg (SFT)	53.05	56.77	45.31	58.59
Event-OPD (ours)	**60.86**	59.43	**49.28**	**62.42**

We examine how Event-OPD handles the four question categories in Table 9. Compared with FeVOS-R1 (SFT), it gains 4.06 points on consequence, 4.58 on multi-solution, and 2.44 on counterfactual questions. The gain on multi-solution is particularly relevant because these questions require identifying all valid participants, so finding one plausible target is insufficient. Even so, this remains the lowest-scoring category for every listed model, indicating that grounding the complete set of targets remains difficult. Source questions show a different pattern: FeVOS-R1 retains a 0.30 point lead. The benefits of Event-OPD are therefore more pronounced for tracing consequences and considering multiple targets than for identifying the cause of an event.

Table 10: Test $\mathcal{J}\&\mathcal{F}$ by source dataset.

Model	STAR	Oops!	EPIC	VOST	NExT-QA	GroundMoRe	MVP
VideoSEG-O3 (SFT)	56.28	60.85	44.59	49.60	56.87	59.45	55.18
FeVOS-R1 (SFT)	59.11	60.13	46.50	**55.48**	60.07	58.29	54.20
Sa2VA (SFT)	51.88	57.78	41.77	43.05	58.02	67.06	47.46
MomentSeg (SFT)	54.70	58.20	44.14	47.06	58.57	57.85	54.26
Event-OPD (ours)	**59.94**	**62.05**	**47.03**	52.20	**63.28**	**68.70**	**58.19**

We further compare results by source in Table 10. Event-OPD leads on six of seven, including GroundMoRe, where it exceeds Sa2VA (SFT), the strongest baseline, by 1.64 points. However, FeVOS-R1 (SFT), which builds on a model trained for foresight reasoning (Lan et al., 2026), retains a 3.28 point lead on VOST. EPIC-KITCHENS remains difficult for all listed models: every score is below 50, and Event-OPD leads FeVOS-R1 by 0.53 points.

To examine the errors behind these scores, we provide two failure cases in Figure 8. The first concerns incomplete masks and extra regions, while the second shows a wrong participant being selected and tracked.

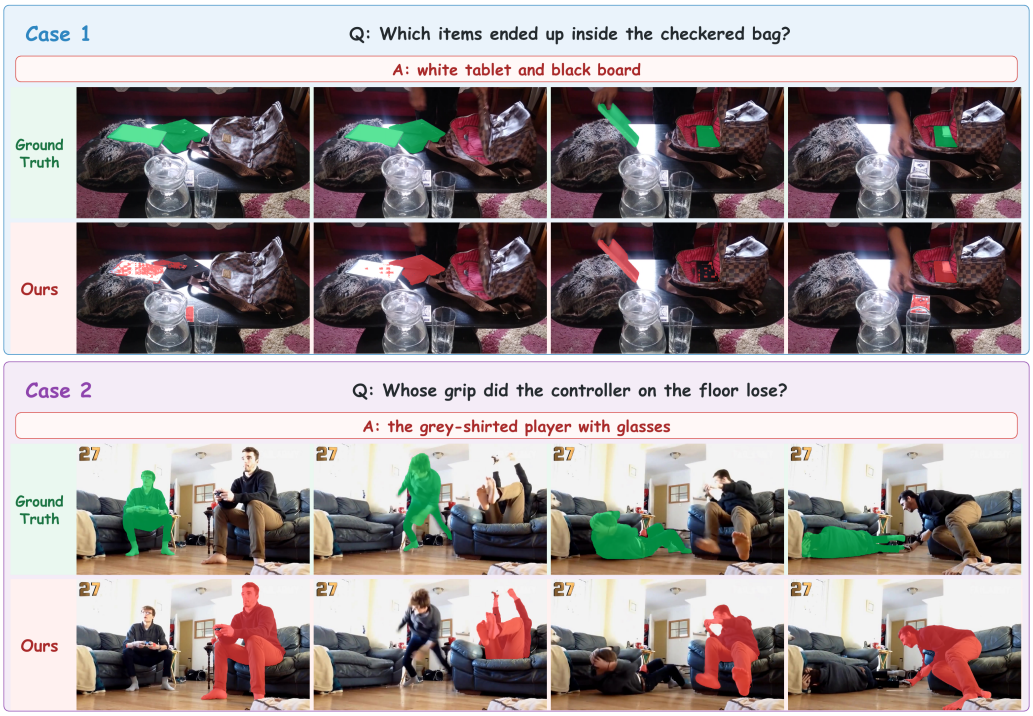

Figure 8: Failure cases of Event-OPD. Each case shows the question, reference answer, and four sampled frames. Green marks ground-truth masks, and red marks our model's masks.

In Case 1, the question asks which items end up inside the checkered bag. The reference contains both the white tablet and the black board. Our masks miss parts of these items and extend into nearby regions as the items are handled beside the bag. A correct multi-solution answer thus requires both finding all valid targets and separating their masks from nearby objects.

In contrast, Case 2 concerns the choice of participant. The question asks whose grip the controller on the floor lost, and the answer is the grey-shirted player with glasses, who falls during the interaction. Instead, our prediction follows the black-shirted player. Its masks stay on that person across the displayed frames, so the main error is assigning the controller's loss to the wrong participant rather than losing track of a correctly chosen one.

Teacher update ablation. To examine whether the teacher should be updated during distillation, we replace the frozen teacher with one that uses the student's current weights to read the full video (Table 11). This lowers $\mathcal{J}\&\mathcal{F}$ from 60.16 to 58.82. A moving teacher changes its guidance as the student learns, whereas a frozen teacher provides a fixed reference. The result suggests that retaining this reference is more helpful for learning from the additional event evidence.

Observed-segment ablation. We then examine why CaReVOS provides observations on both sides of the hidden interaction by giving the reasoner either segment alone (Table 12). The opening alone reaches 54.71 $\mathcal{J}\&\mathcal{F}$, compared with 49.08 for the closing alone, suggesting that the initial scene provides useful information about potential participants that is harder to recover from the outcome. Yet combining both segments raises the score to 60.16. The outcome therefore adds evidence that helps distinguish the participants' roles, even though it is less informative on its own. Together, these results support the use of the setup and outcome as complementary evidence for inferring the missing interaction.

Table 11: Ablation on teacher freezing.

Teacher	$\mathcal{J}\uparrow$	$\mathcal{F}\uparrow$	$\mathcal{J}\&\mathcal{F}\uparrow$
Moving	56.82	60.82	58.82
Frozen (ours)	**58.33**	**61.98**	**60.16**

Table 12: Ablation on observed segments.

Reasoner input	$\mathcal{J}\uparrow$	$\mathcal{F}\uparrow$	$\mathcal{J}\&\mathcal{F}\uparrow$
Opening only	52.86	56.55	54.71
Closing only	47.06	51.10	49.08
Opening + closing	**58.33**	**61.98**	**60.16**

Distillation loss ablation. Finally, we assess the choice of token distillation loss by replacing forward KL with Jensen–Shannon divergence (JSD) in Table 13. The score falls from 60.16 to 59.39 $\mathcal{J}\&\mathcal{F}$. JSD nevertheless remains above the SFT baseline of 58.54, suggesting that the benefit of the distillation framework is not limited to forward KL.

Table 13: Distillation loss ablation.

Loss	$\mathcal{J}\uparrow$	$\mathcal{F}\uparrow$	$\mathcal{J}\&\mathcal{F}\uparrow$
JSD	57.56	61.22	59.39
Forward KL (ours)	**58.33**	**61.98**	**60.16**

D LIMITATIONS

Although CaReVOS draws on seven source datasets, it remains limited in scale and focuses on short clips. This limits the range of interactions covered and leaves longer sequences of related events underexplored. Such sequences may require models to connect several causes and outcomes before identifying the relevant participants. Collecting more and longer videos would allow us to evaluate this broader form of event reasoning.

For model training, GRPO did not improve on SFT in our experiments, so more effective use of RL remains an open direction. A useful reward should distinguish the relevant event participants and assess how well they are segmented. However, mask accuracy alone does not reveal whether the model has correctly inferred the missing event. Rewards that also assess the event explanation, together with other optimization strategies, may help improve reasoning and segmentation jointly.

E ANNOTATION PROMPTS

We use Qwen3.7-Plus to assist with clip screening and Qwen3.8-Max to assist with question generation, then manually review and correct the outputs after each stage. The prompt excerpts below omit illustrative examples and other non-essential instructions for readability.

Stage 1 screening and cropping prompt, Qwen3.7-Plus

Task: Video Filtering & Cropping for First/Last−Segment Reasoning Segmentation

You curate clips for a reasoning−segmentation benchmark. The downstream model will see ONLY a short opening segment (PRELUDE) and a short closing segment (RESULT); the MIDDLE is hidden. From those two glimpses it must reason out which object to segment. Watch the video, find event segments worth cropping, and return them as JSON.
BE STRICT: a rejected good video costs little; a bad clip poisons the benchmark.

What counts as an OBJECT
A countable object must be ALL of:
− a discrete, bounded THING with a closed contour (person, animal, vehicle, tool, ball, chair , door, box). NEVER a scene surface or region: floor, ground, grass, wall, ceiling, road, water, sky, background crowd. If your story needs ”the floor/ground/wall”, drop it.
− nameable with a SPECIFIC noun phrase (”red toy car”, ”man in blue shirt”);
− visible and segmentable at a glance, and TRACKABLE through occlusion and crossings.
− EVENT−RELEVANT: it is the object that changes, causes, receives, carries, supports, controls, or enables the change; OR it is a clearly segmentable alternative whose presence makes identifying the manipulated target from the endpoints non−trivial. A random background prop does not count, but a distinct neighboring item may count when the viewer must distinguish it from the target to answer the downstream question.
Rules
1. OBJECT COUNT: >=3 qualifying objects, each relevant to the event or to resolving the target identity: the object that CHANGES, the object(s) that could have caused or received the change, and other independently segmentable participants a question could ask about. People count. Do not pad with scene surfaces, witnesses, or irrelevant background items; reject the clip instead. The count is a floor to be met by genuine participants or identity− relevant alternatives, never a quota to be filled. If fewer than three survive, return no clip.
6. TWO−SEGMENT NEED: the prelude and the result must carry COMPLEMENTARY evidence. If either visible segment ALONE settles what happened and to which object, drop the clip.
9. ENDPOINT CONTRAST (decisive): a reader who sees ONLY the prelude and ONLY the result must be able to tell that something changed and to name what it was. Something must have appeared, vanished, moved to a clearly different place, opened, closed, broken, emptied, filled, or changed colour/shape/state. It is NOT enough that a real event happened in the hidden middle. Step A: write one short phrase naming a difference between the last prelude moment and the first result moment, using only those two views, for example ”the bag is now empty and folded” or ”the glass has moved from counter to dishwasher”. Step B: if you wrote a phrase in step A, rule 9 PASSES; if you could write no phrase at all, rule 9 fails.
Output: return ONLY a JSON array, one object per valid clip (or [] for none).

```
[
  {
    ”clip_id”: ”sourcename_start_end”, ”source_video”: ”original filename or path”,
    ”timestamp”: {”start”: ”MM:SS.mmm”, ”end”: ”MM:SS.mmm”},
    ”segments”: {”prelude_end”: ”MM:SS.mmm”, ”result_start”: ”MM:SS.mmm”},
    ”objects”: [
      {”name”: ”specific subject”, ”role”: ”changing_object | cause_candidate | instrument | carrier | support | enabler | initiator | alternative_target”, ”segmentable”: true}
    ],
    ”core_event”: ”one sentence”,
    ”reason”: ”the visible mechanism and what each segment adds”,
    ”target_object”: ”the object the model must segment”
  }
]
```

Stage 3 question and answer prompt, Qwen3.8-Max

You receive frames sampled from the opening and closing segments, together with the full short video. In one call, return final reasoning QA pairs and their segmentation targets. Use the full video to verify the hidden event and answer identity.

Task boundary

The downstream model sees only the sampled opening and closing frames and the question. A valid question asks it to infer the missing event and select the objects that answer the question.
Content grounding

Ask only the unknown, but state the video−specific event relation being asked about.
Rejection gates

Inspect both observed segments, verify the full video, and delete any candidate failing one gate:

1. ∗∗Content gate:∗∗ reject a question whose only content is a generic role plus ‘incident‘, ‘ event‘, ‘outcome‘, ‘situation‘, ‘mishap‘, or ‘action‘.
2. ∗∗Target−reference gate:∗∗ reject colour, clothing, size, location, direction, pose, or endpoint−state details that visually point to the answer.
3. ∗∗Single−segment gate:∗∗ reject when either observed segment alone selects the answer with comparable confidence.
4. ∗∗Reasoning−role gate:∗∗ reject when the question asks only for a raw visible fact rather than an event relation, hidden transition, or counterfactual.
5. ∗∗Grounding gate:∗∗ reject when the answer is not a discrete segmentable instance visible in at least one observed segment.
6. ∗∗Cross−video gate:∗∗ reject when the unchanged question would fit many unrelated clips equally well.
7. ∗∗Naturalness gate:∗∗ reject template−like meta−language instead of a normal question.
8. ∗∗One−core gate:∗∗ reject a question containing a second clause or chained consequence.
Surface contract

− One natural English question of 5−14 words, ending in one ‘?‘, with one event relation and one unknown and no narrated evidence or explanation.
− No annotation language such as segment, mask, identify, locate, or endpoint.
− Do not reveal the target’s colour, clothing, size, exact position, or pose.
Output

Return exactly one JSON object and no markdown:

```
{
  ”video”: ”filename”,
  ”status”: ”ok|insufficient|reject”,
  ”status_reason”: ”short reason”,
  ”objects”: [{”id”: ”o1”, ”name”: ”private annotation name”,
    ”category_noun”: ”person|animal|vehicle|object”, ”segmentable”: true}],
  ”qa_pairs”: [{”qid”: ”q01”, ”question_intent”: ”affected|risk|cause|responsibility|response|
enabler|necessary_condition|shared_role|counterfactual”,
    ”question”: ”content−grounded reasoning question”,
    ”answer_text”: ”specific target description”, ”target_ids”: [”o1”]}]
}
```

Set ‘status‘ from the final count: ‘ok‘ for 3−4, ‘insufficient‘ for 1−2, ‘reject‘ for zero.