FaSTra: A FASTER YET STRONGER MULTI-TRAJECTORY LEARNING FRAMEWORK FOR REASONING VIDEO OBJECT SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Reasoning Video Object Segmentation (RVOS) aims to identify target objects from complex language descriptions and generate temporally consistent pixel-level masks across video frames. Existing methods either rely on fixed keyframes, making propagation vulnerable to temporal variations, or iteratively search for keyframes at the cost of increased inference overhead. We observe that trajectories initialized from different temporal frames exhibit complementary reliability across different temporal regions. Motivated by this observation, we propose **FaSTra**, a faster yet stronger multi-trajectory learning framework that reformulates keyframe selection as frame-wise trajectory selection. **FaSTra** employs a temporally diverse frame selector to sample a small set of keyframes and generate multiple candidate trajectories with a lightweight propagator, while a complementary trajectory router dynamically selects the most reliable trajectory at each frame. To enhance multi-trajectory learning, we construct a best-of-trajectories teacher and introduce tri-level propagation distillation as well as utility-aware route distillation to strengthen propagation and routing, respectively. We further introduce GRPO-based frame-impact optimization to align frame-level routing decisions with segmentation improvements. **FaSTra**-Qwen3-VL-4B outperforms Sa2VA-Qwen3-VL-4B by 12.9 and 6.5 $\mathcal{J}\&\mathcal{F}$ points on MeViS and Ref-DAVIS17, respectively, while improving frames per second (FPS) by 26.9-33.2% across videos of 60, 120, and 150 frames.

1 INTRODUCTION

Reasoning Video Object Segmentation (RVOS) aims to identify target objects from language descriptions involving implicit semantics, complex relations, or world knowledge, and generate temporally consistent pixel-level masks throughout the video (Gong et al., 2025). Unlike conventional video segmentation guided by explicit class labels or simple referring expressions, RVOS requires complex semantic reasoning, target localization, and long-term identity preservation (Shao et al., 2026). It therefore calls for a unified framework capable of vision-language understanding, spatial grounding, and temporal reasoning (Zhuge et al., 2026).

Existing methods (Hong & Do, 2026; Xu et al., 2026; Li et al., 2025) typically couple a multi-modal large language model (MLLM) for target reasoning with a video segmentor for mask prediction and propagation. As shown in Fig. 1, Sa2VA (Yuan et al., 2026) initializes SAM2 (Ravi et al., 2025) propagation from sampled keyframes, but fixed frames tend to be unreliable under occlusion, deformation, and target reappearance (Ding et al., 2025). In contrast, VideoSEG-O3 (Dai et al., 2026) improves robustness by handling challenging videos through iterative keyframe exploration, at the cost of repeated visual sampling and multi-turn reasoning overhead (Zeng et al., 2026). This contrast reveals a fundamental trade-off in balancing accuracy and efficiency.

We argue that this tension stems in part from treating keyframe selection and mask propagation as two decoupled stages (Xu & Yao, 2022). They mainly ask which keyframe should initialize propagation, while overlooking a broader perspective: given multiple complementary trajectories initialized from different frames, can the model dynamically select the most reliable trajectory at each frame without repeatedly searching for new keyframes?

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.

To answer this question, we analyze trajectories initialized from different keyframes and observe strong temporal complementarity in their propagation reliability. Moreover, keyframe accuracy does not necessarily reflect long-term propagation reliability (Heo et al., 2021). A locally accurate frame may suffer from rapid drift over time, whereas another frame may preserve target identity more reliably within a particular temporal interval (Hu et al., 2021). This finding suggests that video-level performance cannot be reduced to selecting the keyframe that appears optimal at the current frame. Instead, a small set of temporally diverse frames can generate complementary trajectories, enabling frame-wise selection of the most reliable trajectory.

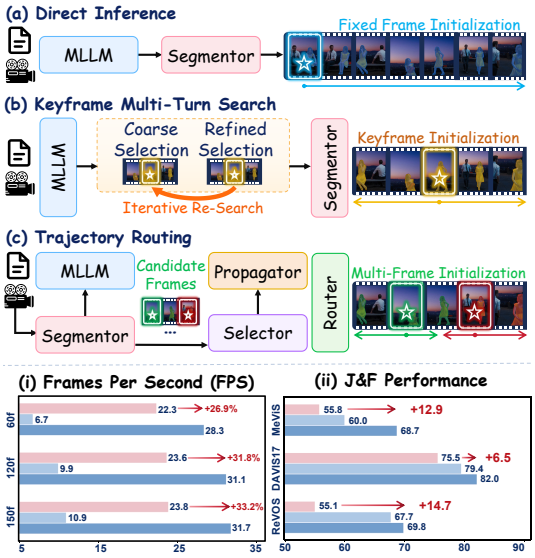

Figure 1: Comparison of RVOS keyframe propagation strategies and quantitative results of Sa2VA-4B, VideoSEG-O3-CoT-4B, and FaSTra-4B.

To this end, we propose a **Fa**ster yet **S**tronger multi-**Tra**jectory (`FaSTra`) framework that combines strong semantic reasoning with lightweight temporal processing. The computationally intensive MLLM first performs high-level semantic reasoning and target grounding. A temporally diverse frame (TDF) selector then selects a small set of diverse keyframes from the candidate predictions, which are propagated into full-video trajectories by a lightweight propagator (Zhou et al., 2025). Finally, a complementary trajectory (CT) router selects the most reliable trajectory at each frame and then stitch together to form the final video mask sequence. This design replaces expensive online keyframe search with lightweight frame-wise routing, improving robustness to occlusion, drift, and target reappearance while reducing inference overhead.

Although trajectory routing exploits temporal complementarity, conventional single-trajectory supervision cannot fully capture such cross-trajectory knowledge (Miles et al., 2023). Independent ground-truth supervision optimizes each trajectory in isolation, without exposing the potential gains from cross-trajectory complementarity or indicating which temporal states and boundary structures are worth preserving (Kim et al., 2026a). We therefore introduce a best-of-trajectories (BoT) Teacher to distill cross-trajectory complementary knowledge. During training, we generate multiple frame-initialized trajectories and construct a stitched BoT teacher by selecting the highest-quality trajectory at each frame according to ground-truth supervision. Tri-level propagation distillation then transfers boundary, identity, and temporal knowledge through soft masks, object representations, and propagation memories. Utility-aware route distillation further teaches the router to rank candidate trajectories by their frame-wise utility. Together, these objectives improve both propagation and routing, enabling the lightweight model to better exploit cross-trajectory complementarity.

Although route distillation provides the router with relative quality supervision, it still relies on fixed teacher signals. It encourages the router to imitate the teacher's preferences rather than directly optimizing the segmentation gains induced by its own routing decisions (Gong et al., 2026). This limitation becomes particularly pronounced when candidate trajectories have similar quality or when a routing choice causes substantial downstream performance degradation (Li et al., 2025; Gong et al., 2026). We therefore introduce GRPO-based frame-impact (FI) optimization. It directly optimizes frame-wise routing with segmentation gains as rewards and computes group-relative advantages over candidate trajectory choices at each frame. Unlike assigning one reward from overall video performance, FI-GRPO assigns action-specific credit based on frame-level segmentation outcomes. It thereby enables the router to learn the relative utility of choices, moving beyond teacher imitation toward direct optimization of routing-induced segmentation performance.

As shown in Fig. 1, `FaSTra` reformulates static keyframe selection as dynamic multi-trajectory routing. With lightweight propagation, cross-trajectory distillation, and frame-level optimization, it enables adaptive temporal reasoning with fewer MLLM invocations and lower inference overhead. Our main contributions are summarized as follows: ❶ We propose `FaSTra`, which exploits temporal complementarity among trajectories at a lower inference cost. ❷ We introduce two knowledge dis-

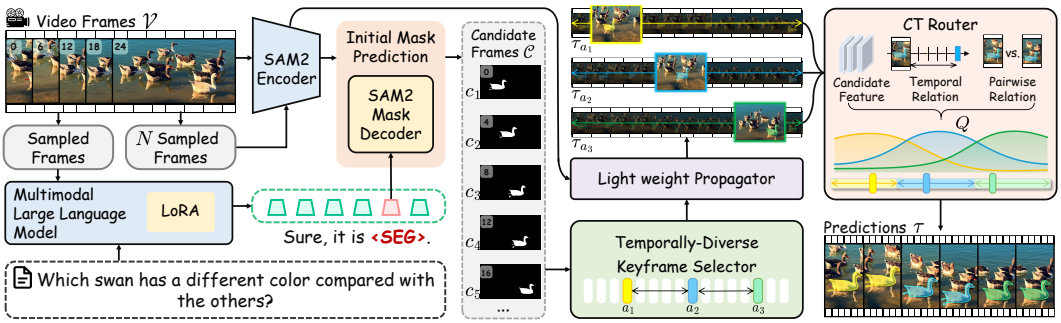

Figure 2: Overview of the proposed **FaSTra**.

tillation strategies based on a BoT teacher to jointly strengthen knowledge transfer to the lightweight propagator and trajectory router. ❸ We propose FI-GRPO optimization to align frame-level routing with segmentation impact, enabling more effective dynamic policies. Experiments show that **FaSTra** achieves both superior accuracy and faster inference, outperforming Sa2VA-4B by 12.9 and 6.5 $\mathcal{J}\&\mathcal{F}$ points on MeViS and Ref-DAVIS17, respectively, while improving FPS by 26.9-33.2%.

2 Related Work

Language-Driven Video Segmentation Tasks. MLLM-based language-driven segmentation has expanded from image-based reasoning segmentation to video settings (Yan et al., 2024; Bai et al., 2024). In RVOS, frame selection can directly affect downstream mask quality (Yan et al., 2024; Yuan et al., 2026). FindTrack (Cho et al., 2025) selects a reliable reference frame before propagation, whereas SOLA (Kim et al., 2024) generates multiple candidate trajectories and performs language-aligned video-level selection. However, a track propagated from a single frame may not remain reliable throughout the full video. We therefore retain multiple propagated trajectories and select the most reliable one for each temporal segment.

Knowledge Distillation Approaches. Knowledge distillation (KD) is widely used to endow lightweight models with the reasoning ability of multimodal large language models (MLLMs), with existing methods falling into logit-based (Shu et al., 2025), feature-based (Kim et al., 2026b), and relation-based (Cai et al., 2025) paradigms. Despite these advances, existing approaches mainly transfer static representations, semantic structures, or global cross-modal alignment, with limited consideration of jointly distilling temporal propagation and frame-wise trajectory selection. This motivates distillation mechanisms that explicitly preserve both propagation knowledge and trajectory-level decision knowledge for RVOS.

Reinforcement Learning (RL) Approaches. Group Relative Policy Optimization (GRPO) optimizes policies using relative rewards within sampled groups without an additional critic (Shao et al., 2024). In video reasoning segmentation, RL has been applied to keyframe selection and spatial grounding. Veason-R1 (Gong et al., 2026) optimizes keyframe localization and cross-frame consistency, ReVSeg (Li et al., 2025) decomposes reasoning into semantic interpretation, temporal evidence selection, and spatial grounding, and RCoT-Seg (Wen et al., 2026) explicitly models keyframe reliability for adaptive reselection. VideoSEG-O3 (Dai et al., 2026) further performs multi-turn temporal-spatial exploration with segmentation feedback. In contrast, we treat trajectories propagated from multiple frames as GRPO candidates and optimize how they are combined across temporal segments.

3 Methodology

Our proposed **FaSTra** framework is illustrated in Fig. 2. Given a video $\mathcal{V} = \{\boldsymbol{I}_t\}_{t=1}^{T}$ and a text query $\mathcal{E}$, **FaSTra** first employs an MLLM to jointly reason over the visual content and query semantics, yielding a language-conditioned <SEG> token with hidden embedding $\boldsymbol{e}$. We sample N candidate frames with temporal indices $\mathcal{C} = \{c_i\}_{i=1}^{N}$, where $c_i \in \{1, \ldots, T\}$. For each candidate frame c_i, the SAM2 (Ravi et al., 2025) mask decoder $\mathcal{F}_{\text{segmentor}}$ predicts the initial mask logits $\boldsymbol{L}_{c_i}$, an object-pointer embedding $\boldsymbol{p}_{c_i}$, and an object-presence score $o_{c_i} \in \mathbb{R}$. Based on these candidate representations, the TDF selector $\mathcal{F}_{\text{selector}}$ selects K temporally diverse keyframes, with their indices denoted by $\mathcal{A} = \{a_k\}_{k=1}^{K} \subseteq \mathcal{C}$. Each selected seed frame a_k independently initializes the

lightweight propagator $\mathcal{F}_{\text{propagator}}$ (Zhou et al., 2025), which performs bidirectional propagation over the entire video to produce a candidate trajectory $\boldsymbol{\tau}_{a_k} = \{\tau_{a_k,t}\}_{t=1}^{T}$. $\tau_{a_k,t}$ denotes the prediction from keyframe a_k at frame t, comprising mask logits $\boldsymbol{L}_{a_k,t}$, an object-pointer embedding $\boldsymbol{p}_{a_k,t}$, and an object-presence score $o_{a_k,t}$. Motivated by the complementary reliability of trajectories initialized at different temporal locations, the CT router $\mathcal{F}_{\text{router}}$ estimates their frame-wise quality scores, forming a quality matrix $\boldsymbol{Q} = [q_{a_k,t}] \in \mathbb{R}^{K\times T}$, where $q_{a_k,t}$ denotes the predicted quality of trajectory $\boldsymbol{\tau}_{a_k}$ at frame t. Based on these quality estimates, the router determines a frame-wise routing path $\boldsymbol{z} = \{z_t\}_{t=1}^{T}$, where $z_t \in \mathcal{A}$ specifies the source keyframe of the trajectory selected at frame t. The final trajectory is then obtained by stitching the selected frame-wise predictions as $\boldsymbol{\tau} = \{\tau_{z_t,t}\}_{t=1}^{T}$. In this way, **FaSTra** reformulates conventional keyframe-based propagation as a multi-trajectory routing problem, enabling frame-wise selection among complementary trajectories initialized from temporally diverse keyframes.

3.1 Temporally-Diverse Keyframe Selector

Trajectories initialized from different keyframes tend to exhibit complementary reliability across temporal regions (Kim et al., 2024). This motivates selecting multiple propagation keyframes to capture such temporal complementarity. However, top-K selection based solely on the initial object scores o_{c_i} is suboptimal, as frame-level confidence does not necessarily indicate long-term propagation quality, while high-confidence candidates can be temporally redundant and concentrated in easily visible regions (Guo et al., 2026). We therefore design a selector that jointly accounts for candidate utility and temporal diversity in selecting the K seed frames. For each candidate c_i, the selector encodes its mask quality, region geometry, temporal position, and cross-candidate consistency, and relates these cues, together with its object representation, to produce a raw utility score u_{c_i}. Then, we select keyframes sequentially. Starting from the raw utilities, the first keyframe is selected without a diversity penalty. The selected candidate is then removed from the support. After each selection, the score of every remaining candidate is reduced according to its temporal proximity to the newly selected keyframe.

3.2 Complementary Trajectory Router

A trajectory that is reliable in one temporal region may degrade under occlusion or target reappearance, while another initialized from a different keyframe may remain reliable (Cheng & Schwing, 2022; Li & Loy, 2018). We therefore estimate their reliability by frame and dynamically route among them over time. For a trajectory initialized from keyframe $a_k \in \mathcal{A}$ at frame t, the router produces a logit and its calibrated quality: $l_{a_k,t} = \mathcal{F}_{\text{router}}(\boldsymbol{e}, \boldsymbol{L}_{a_k,t}, \boldsymbol{p}_{a_k,t}, o_{a_k,t}, \rho_{a_k,t})$, $q_{a_k,t} = \sigma(l_{a_k,t})$, where $\rho_{a_k,t} = (t - a_k)/(T-1)$ is the normalized temporal offset from the propagation keyframe. Frame-wise greedy selection from $\boldsymbol{Q}$ may cause unstable switching between trajectories with similar scores. We therefore introduce a switching penalty and jointly decode a temporally coherent routing trajectory:

$$\boldsymbol{z} = \arg\max_{z_1,\ldots,z_T\in\mathcal{A}} \left[\sum_{t=1}^{T} q_{z_t,t} - \lambda_{\text{swi}} \sum_{t=2}^{T} \mathbb{I}[z_t \neq z_{t-1}]\right], \tag{1}$$

where $\mathbb{I}[\cdot]$ is the indicator function and λ_{swi} controls the switching penalty. The optimal path is efficiently obtained by Viterbi (Viterbi, 1967) decoding, and the final routed trajectory is constructed as $\boldsymbol{\tau} = \{\tau_{z_t,t}\}_{t=1}^{T}$, with corresponding mask logits $\boldsymbol{L}_t = \boldsymbol{L}_{z_t,t}$.

3.3 Best-of-Trajectories Teacher and Knowledge Distillation

Unlike conventional knowledge distillation (Gu et al., 2024), the BoT teacher is not a larger model (Tian et al., 2025), but a composite teacher constructed entirely from multi-trajectory complementarity and ground-truth trajectory selection. This design keeps teacher supervision in the same representation space as the student, enabling direct transfer of more reliable trajectory-selection behavior. Given the ground truth $\boldsymbol{Y}_t$, we evaluate each trajectory at each frame using the $\mathcal{J}\&\mathcal{F}$ metric and produce the teacher quality matrix $\boldsymbol{Q}^{\mathcal{T}} = [q^{\mathcal{T}}_{a_k,t}]$. The teacher route selects the highest-quality trajectory at each frame, yielding the BoT teacher trajectory $\boldsymbol{\tau}^{\mathcal{T}} = \{\tau_{z^{\mathcal{T}}_t,t}\}_{t=1}^{T}$ with mask logits $\boldsymbol{L}^{\mathcal{T}}_t = \boldsymbol{L}_{z^{\mathcal{T}}_t,t}$. Based on the BoT teacher, we introduce two complementary distillation strategies to transfer knowledge to the lightweight propagator and the router.

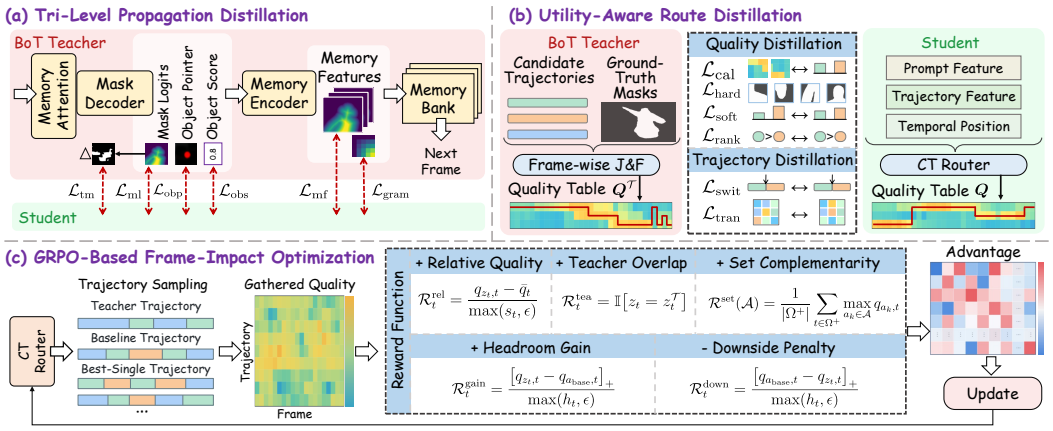

Figure 3: Overall framework of our proposed distillation and FI-GRPO strategies.

Tri-Level Propagation Distillation. Distilling only the final masks cannot fully capture the propagation behavior of the BoT teacher, as trajectory quality also depends on internal propagation states and their temporal evolution (Song & Zheng, 2026; Huang et al., 2026). We therefore distill knowledge at three levels: prediction-level logit distillation $\mathcal{L}_{\text{ml}}$, state-level feature distillation$\mathcal{L}_{\text{mf}}, \mathcal{L}_{\text{obp}}, \mathcal{L}_{\text{obs}}$, and temporal-level relation distillation $\mathcal{L}_{\text{tm}}$, $\mathcal{L}_{\text{gram}}$. Let $\boldsymbol{L}_t^{\mathcal{T},\mathcal{S}}$, $\boldsymbol{p}_t^{\mathcal{T},\mathcal{S}}$, $o_t^{\mathcal{T},\mathcal{S}}$, and $\boldsymbol{H}_t^{\mathcal{T},\mathcal{S}}$ denote the mask logits, object pointer, object-presence score, and memory representation of the teacher and student. The three distillation objectives are defined as follows:

$$\mathcal{L}_{\text{ml}} = \frac{1}{T}\sum_{t=1}^{T} \beta_t \left\| \boldsymbol{w}_t \odot \left(\sigma(\boldsymbol{L}_t^{\mathcal{S}}) - \sigma(\boldsymbol{L}_t^{\mathcal{T}})\right)\right\|_1,$$

$$\mathcal{L}_{\text{mf}} + \mathcal{L}_{\text{obp}} + \mathcal{L}_{\text{obs}} = \frac{1}{T}\sum_{t=1}^{T} \beta_t \Big[\mathcal{D}_H\left(\boldsymbol{H}_t^{\mathcal{S}}, \boldsymbol{H}_t^{\mathcal{T}}\right) + \mathcal{D}_p\left(\boldsymbol{p}_t^{\mathcal{S}}, \boldsymbol{p}_t^{\mathcal{T}}\right) + \mathcal{D}_o\left(o_t^{\mathcal{S}}, o_t^{\mathcal{T}}\right)\Big],$$

$$\mathcal{L}_{\text{tm}} = \frac{1}{T-1}\sum_{t=2}^{T} \left\|\Delta\boldsymbol{M}_t^{\mathcal{S}} - \Delta\boldsymbol{M}_t^{\mathcal{T}}\right\|_1, \mathcal{L}_{\text{gram}} = \frac{1}{T(T-1)} \left\|\text{OffDiag}\big(\boldsymbol{G}(\boldsymbol{H}^{\mathcal{S}}) - \boldsymbol{G}(\boldsymbol{H}^{\mathcal{T}})\big)\right\|_1, \tag{2}$$

where $\boldsymbol{M}_t^{\mathcal{T},\mathcal{S}} = \sigma(\boldsymbol{L}_t^{\mathcal{T},\mathcal{S}})$ and $\Delta\boldsymbol{M}_t = \boldsymbol{M}_t - \boldsymbol{M}_{t-1}$. β_t is the best-worst trajectory-quality gap, while $\boldsymbol{w}_t$ is a pixel-wise map emphasizing the teacher foreground and boundary. $\odot$ denotes element-wise multiplication, and $\|\cdot\|_1$ denotes the element-wise L_1 norm. $\mathcal{D}_H$ and $\mathcal{D}_p$ are cosine distances, whereas $\mathcal{D}_o$ is the absolute object-score difference. Finally, $\boldsymbol{G}(\cdot)$ is the temporal cosine-similarity Gram matrix and $\text{OffDiag}(\cdot)$ removes its diagonal.

Utility-Aware Route Distillation. We further introduce utility-aware route distillation to transfer the teacher's trajectory probabilities to the router through two-level supervision. For frame t, we form the student probability $\pi_{a_k,t} = \exp(l_{a_k,t}) / \sum_{j=1}^{K}\exp(l_{a_j,t})$ and $\boldsymbol{\pi}_t = \{\pi_{a_k,t}\}_{k=1}^{K}$. The teacher and student quality vectors are $\boldsymbol{q}_t^{\mathcal{T}} = \{q_{a_k,t}^{\mathcal{T}}\}_{k=1}^{K}$ and $\boldsymbol{q}_t = \{q_{a_k,t}\}_{k=1}^{K}$, respectively. Let Ω^+ denotes the object-present frames. Quality-level distillation comprises four complementary objectives, including calibration, hard assignment, soft-distribution matching, and pairwise supervision:

$$\mathcal{L}_{\text{cal}} = \frac{1}{K|\Omega^+|}\sum_{t\in\Omega^+}\sum_{k=1}^{K}\text{BCE}\big(q_{a_k,t}, q_{a_k,t}^{\mathcal{T}}\big), \quad \mathcal{L}_{\text{hard}} = -\frac{1}{|\Omega^+|}\sum_{t\in\Omega^+}\log\pi_{z_t^{\mathcal{T}},t},$$

$$\mathcal{L}_{\text{soft}} = \frac{1}{|\Omega^+|}\sum_{t\in\Omega^+}\text{KL}\big(\text{softmax}(\boldsymbol{q}_t^{\mathcal{T}}/\kappa)\,\big\|\,\boldsymbol{\pi}_t\big), \tag{3}$$

$$\mathcal{L}_{\text{rank}} = \frac{1}{|\Omega^+|\binom{K}{2}}\sum_{t\in\Omega^+}\sum_{i<j}\left|q_{a_i,t}^{\mathcal{T}} - q_{a_j,t}^{\mathcal{T}}\right| \cdot \text{softplus}\Big[-\text{sign}\Big(q_{a_i,t}^{\mathcal{T}} - q_{a_j,t}^{\mathcal{T}}\Big)\left(q_{a_i,t} - q_{a_j,t}\right)\Big].$$

BCE denotes binary cross-entropy and calibrates $q_{a_k,t}$ to the teacher's frame-wise $\mathcal{J}\&\mathcal{F}$ score. Hard-assignment distillation encourages the router to assign the highest preference to the teacher-

selected keyframe $z_t^{\mathcal{T}}$. KL denotes Kullback-Leibler divergence (Kullback & Leibler, 1951), and κ is the temperature for soft distribution matching. The function $\mathrm{softplus}(x) = \log(1+e^x)$ provides a smooth pairwise ranking penalty that decreases when the student quality difference agrees with the teacher preference and increases otherwise. Let $\varsigma_t^{\mathcal{T}} = \mathbb{I}[z_t^{\mathcal{T}} \neq z_{t-1}^{\mathcal{T}}]$ and $\varsigma_t = 1 - \sum_{k=1}^{K} \pi_{a_k,t-1}\pi_{a_k,t}$ denote the teacher and predicted switching probabilities. Trajectory-level distillation further aligns trajectory selection and switching:

$$\mathcal{L}_{\mathrm{swit}} = \frac{1}{T-1}\sum_{t=2}^{T} \left|\varsigma_t - \varsigma_t^{\mathcal{T}}\right|, \qquad \mathcal{L}_{\mathrm{tran}} = \frac{1}{T-1}\sum_{t=2}^{T} \mathrm{KL}\left(\boldsymbol{\xi}_t^{\mathcal{T}} \,\middle\|\, \boldsymbol{\pi}_{t-1} \otimes \boldsymbol{\pi}_t\right), \tag{4}$$

where $\boldsymbol{\xi}_t^{\mathcal{T}}$ is the one-hot joint distribution at $(z_{t-1}^{\mathcal{T}}, z_t^{\mathcal{T}})$, and $\otimes$ denotes the outer product.

3.4 GRPO-Based Frame-Impact Optimization

Distillation transfers the teacher's decisions but not their impact. Errors among similar trajectories are often negligible, whereas incorrect choices near occlusion recovery or identity drift can substantially degrade performance (Wen et al., 2026). Since these discrete decisions are non-differentiable, we optimize their impact through a reward-based RL strategy, termed FI-GRPO optimization.

Existing RL-based RVOS methods mainly reward segmentation predictions while treating propagation as fixed post-processing (Zhong et al., 2026). Such video-level rewards provide only a global signal and cannot identify which trajectory decision contributes to the final gain (Li et al., 2025). We instead define frame-wise rewards over the quality matrix $\boldsymbol{Q}$, so that each routing decision receives feedback according to its actual impact. We define the best single-trajectory baseline as $a_{\mathrm{base}} = \arg\max_{a_k} \frac{1}{T}\sum_{t=1}^{T} q_{a_k,t}$. For each frame t, the available headroom is $h_t = \max_{a_k \in \mathcal{A}} q_{a_k,t} - q_{a_{\mathrm{base}},t}$, while $\bar{q}_t$ and s_t denote the mean and standard deviation of the trajectory qualities $\boldsymbol{q}_t$, respectively. The frame-impact reward is defined as follows:

$$\mathcal{R}_t^{\mathrm{route}}(z_t) = \underbrace{\frac{q_{z_t,t} - \bar{q}_t}{\max(s_t, \epsilon)}}_{\mathcal{R}_t^{\mathrm{rel}}} + \lambda_{\mathrm{g}} \underbrace{\frac{[q_{z_t,t} - q_{a_{\mathrm{base}},t}]_+}{\max(h_t, \epsilon)}}_{\mathcal{R}_t^{\mathrm{gain}}} - \lambda_{\mathrm{d}} \underbrace{\frac{[q_{a_{\mathrm{base}},t} - q_{z_t,t}]_+}{\max(h_t, \epsilon)}}_{\mathcal{R}_t^{\mathrm{down}}} + \lambda_{\mathrm{o}} \underbrace{\mathbb{I}[z_t = z_t^{\mathcal{T}}]}_{\mathcal{R}_t^{\mathrm{tea}}}, \tag{5}$$

where $[x]_+ = \max(x, 0)$ and $\epsilon > 0$ ensures numerical stability. $\mathcal{R}_t^{\mathrm{rel}}$ measures the relative quality of the selected trajectory, while $\mathcal{R}_t^{\mathrm{gain}}$ and $\mathcal{R}_t^{\mathrm{down}}$ measure its improvement and degradation relative to the best single-trajectory baseline. $\mathcal{R}_t^{\mathrm{tea}}$ provides additional guidance from the teacher trajectory assignment. Keyframe selection is optimized at the set level using the upper-bound quality provided by the selected trajectories:

$$\mathcal{R}^{\mathrm{set}}(\mathcal{A}) = \frac{1}{|\Omega^+|}\sum_{t \in \Omega^+} \max_{a_k \in \mathcal{A}} q_{a_k,t}. \tag{6}$$

The maximum favors complementary keyframe sets, since trajectories covering already well-represented temporal regions provide little additional reward. During optimization, we compare multiple routing assignments and obtain relative advantages. The router is then optimized with a clipped policy objective using the relative rewards of sampled keyframe sets. This FI-GRPO optimization assigns stronger credit to consequential decisions near occlusion, recovery, and identity drift, prevents easy keyframes from dominating training, and improves both trajectory switching and keyframe complementarity.

4 Experiments

4.1 Experimental Setups

All experiments are conducted on a single NVIDIA RTX PRO 6000 GPU. Training proceeds in three stages: supervised fine-tuning (SFT) following Sa2VA (Yuan et al., 2026), FI-GRPO, and FI-GRPO + KD. We evaluate `FaSTra` on four established benchmarks for referring segmentation: MeViS (Ding et al., 2023), Ref-YouTube-VOS (Seo et al., 2020), Ref-DAVIS17 (Khoreva et al., 2018), and Ref-SAV (Yuan et al., 2026), as well as ReVOS (Yan et al., 2024) for reasoning segmentation. We report region similarity $\mathcal{J}$, boundary accuracy $\mathcal{F}$, and $\mathcal{J}\&\mathcal{F}$. All ablation studies are conducted on the Ref-DAVIS17 dataset and discussed using $\mathcal{J}\&\mathcal{F}$. Detailed experimental settings and setup are provided in Appendix E.1.

Table 1: Main results on video object segmentation benchmarks. GLUS^{A} denotes the Additional-SFT variant. Sa2VA-i results follow the 8-frame setting (Nekrasov et al., 2025).

Method	MeViS			Ref-YouTube-VOS			Ref-DAVIS17			Ref-SAV		
	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
Specialized Methods												
FlowRVS-Wan2.1-T2V-1.3B (Jiang et al., 2026a)	51.1	47.6	54.6	69.6	67.1	72.1	73.3	68.4	78.2	–	–	–
DeRVOS (Cheng et al., 2026)	51.8	48.1	55.4	70.0	68.0	71.9	70.9	68.3	73.5	–	–	–
Training-free / Zero-shot Methods												
SDAM-Qwen2.5-VL (Zhu et al., 2026)	48.6	45.5	51.7	65.3	63.4	67.1	76.0	73.2	78.8	–	–	–
Refer-Agent-Ovis2.5-9B (Jiang et al., 2026b)	54.7	51.6	57.7	71.3	69.0	73.6	–	–	–	–	–	–
CoT-RVS-GPT-4o (Kao et al., 2026)	52.2	48.7	55.7	–	–	–	79.1	75.3	82.9	–	–	–
1–3B MLLM-based Methods												
Sa2VA-Qwen3-VL-2B (Yuan et al., 2026)	48.8	44.7	53.0	56.3	54.7	57.9	67.1	62.8	71.4	39.1	37.6	40.5
Sa2VA-i-InternVL2.5-1B (Nekrasov et al., 2025)	52.6	49.9	55.3	70.3	68.4	72.2	73.6	70.1	77.1	–	–	–
InstructSeg-Mipha-3B (Wei et al., 2025b)	–	–	–	67.5	65.4	69.5	71.1	67.3	74.9	–	–	–
VideoSEG-O3-Qwen3-VL-2B (Dai et al., 2026)	55.6	52.5	58.6	70.5	68.3	**72.7**	80.0	76.0	84.0	62.9	61.8	64.0
HyperSeg-Qwen2.5-VL-3B (Wei et al., 2025a)	–	–	–	–	–	68.5	–	–	71.2	–	–	–
`FaSTra`-Qwen3-VL-2B	**59.4**	**55.7**	**63.1**	**70.7**	**68.7**	**72.7**	**82.0**	**78.0**	**86.0**	**64.4**	**63.1**	**65.7**
4–8B MLLM-based Methods												
RViSeg-3.8B (Shao et al., 2026)	44.1	41.8	46.3	62.8	61.5	64.2	61.2	59.3	64.4	–	–	–
Sa2VA-InternVL2.5-4B (Yuan et al., 2026)	46.4	43.3	49.5	71.3	69.1	73.5	73.7	69.6	77.8	–	–	–
Sa2VA-Qwen3-VL-4B (Yuan et al., 2026)	55.8	51.4	60.2	67.0	65.3	68.7	75.5	71.1	79.9	42.2	40.9	43.5
Sa2VA-i-InternVL2.5-4B (Nekrasov et al., 2025)	56.6	53.9	59.3	73.2	71.0	75.4	78.6	75.1	82.1	–	–	–
GLUS^{A}-LLaVA-7B (Lin et al., 2025)	51.3	48.5	54.2	67.3	65.5	69.0	–	–	–	–	–	–
UniPixel-Qwen2.5-VL-7B (Liu et al., 2025)	55.8	53.2	58.3	71.0	69.5	72.4	76.4	72.7	80.1	69.0	68.5	69.6
VISA-Chat-UniVi-7B (Yan et al., 2024)	43.5	40.7	46.3	61.5	59.8	63.2	69.4	66.3	72.5	11.8	13.2	11.3
RViSeg-7B (Shao et al., 2026)	45.8	43.7	48.0	62.8	61.7	64.2	62.3	59.8	64.8	–	–	–
ViLLa-6B (Zheng et al., 2025)	49.4	46.5	52.3	67.5	64.6	70.4	74.3	70.6	78.0	–	–	–
CoT-RVS-online-7B (Kao et al., 2026)	45.9	42.7	49.1	–	–	–	73.9	70.4	77.5	–	–	–
Sa2VA-InternVL2.5-8B (Yuan et al., 2026)	51.5	–	–	72.3	–	–	75.9	–	–	–	–	–
Sa2VA-i-InternVL2.5-8B (Nekrasov et al., 2025)	59.5	56.6	62.4	73.9	71.7	76.3	79.1	75.6	82.7	–	–	–
VideoSEG-O3-Qwen3-VL-4B (Dai et al., 2026)	60.0	57.0	63.0	74.1	71.5	**76.6**	79.4	75.5	83.2	65.5	64.3	66.7
VIRST-VideoChat-Flash-Qwen2-7B (Hong & Do, 2026)	62.9	60.4	65.4	74.2	72.2	76.1	79.5	75.9	83.1	–	–	–
RCoT-Seg-Qwen2.5-VL-7B (Wen et al., 2026)	54.8	51.9	57.7	73.3	71.5	75.1	78.6	75.0	82.2	–	–	–
`FaSTra`-Qwen3-VL-4B	**68.7**	**65.4**	**72.0**	**74.6**	**72.5**	**76.6**	**82.0**	**78.5**	**85.6**	**72.2**	**71.1**	**73.3**

4.2 Comparisons with State-of-the-art Methods

As presented in Table 1, `FaSTra` achieves the highest $\mathcal{J}\&\mathcal{F}$ scores across all four benchmarks within both model-size groups, demonstrating consistent gains across different model capacities. Notably, under the same Qwen3-VL-2B backbone, our 2B model outperforms VideoSEG-O3 (Dai et al., 2026) by 3.8, 0.2, 2.0, and 1.5 points on MeViS, Ref-YouTube-VOS, Ref-DAVIS17, and Ref-SAV, respectively. The consistent gains support our central observation that propagation reliability varies over time and that trajectories initialized from different keyframes provide complementary information that can be dynamically exploited at each frame. This advantage further extends to the more challenging ReVOS benchmark in Table 2, where our 4B model achieves an overall $\mathcal{J}\&\mathcal{F}$ of 69.8, surpassing the best-performing SFT-based and RL-based baselines by 1.4 and 2.1 points, respectively. Together, these results demonstrate that exploiting cross-trajectory temporal complementarity through lightweight dynamic routing is effective for both referring segmentation and more complex reasoning-driven segmentation. More experimental comparisons are provided in Appendix E.2.

4.3 Ablation Study

Component-wise ablations. As shown in Table 3, dynamic keyframe selection consistently outperforms uniform sampling across different routing strategies, highlighting the importance of informative keyframe selection. However, nearest-keyframe assignment provides only marginal gains

Table 2: Main results on reasoning video object segmentation. We include recent SFT-based, training-free, and reinforcement-learning-based methods on ReVOS (Yan et al., 2024).

Method	Referring			Reasoning			Overall		
	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
SFT-based Methods									
Sa2VA-i-InternVL2.5-1B (Nekrasov et al., 2025)	58.6	56.4	60.7	48.3	45.9	50.7	53.4	51.2	55.7
HyperSeg-Mipha-3B (Wei et al., 2025a)	58.5	56.0	60.9	53.0	50.2	55.8	55.7	53.1	58.4
InstructSeg-Mipha-3B (Wei et al., 2025b)	57.0	54.8	59.2	51.9	49.2	54.7	54.5	52.0	56.9
Sa2VA-InternVL2.5-4B (Yuan et al., 2026)	63.0	60.5	65.5	56.2	53.2	59.1	59.6	56.8	62.3
Sa2VA-i-InternVL2.5-4B (Nekrasov et al., 2025)	66.5	64.3	68.8	60.9	58.2	63.6	63.7	61.2	66.1
UniPixel-Qwen2.5-VL-7B (Liu et al., 2025)	65.8	63.9	67.8	61.5	59.4	63.7	63.7	61.7	65.7
VRS-HQ-Chat-UniVi-7B (Gong et al., 2025)	62.1	59.8	64.5	56.1	53.5	58.7	59.1	56.6	61.6
Sa2VA-i-InternVL2.5-8B (Nekrasov et al., 2025)	68.7	66.3	71.1	63.1	60.3	65.8	65.9	63.3	68.5
VRS-HQ-Chat-UniVi-13B (Gong et al., 2025)	63.3	61.1	65.5	56.8	54.1	59.4	60.0	57.6	62.5
VIRST-VideoChat-Flash-Qwen2-7B (Hong & Do, 2026)	70.8	68.8	72.8	66.1	63.9	68.3	68.4	66.3	70.6
RViSeg-3.8B (Shao et al., 2026)	52.6	51.0	54.2	47.4	45.8	49.0	50.3	49.3	51.2
RViSeg-7B (Shao et al., 2026)	52.2	50.8	53.5	48.2	46.5	49.9	50.6	49.5	51.7
RGA3-7B (Wang et al., 2025)	60.5	58.7	62.3	55.4	53.1	57.7	58.0	55.9	60.0
ViLLa-6B (Zheng et al., 2025)	–	–	–	–	–	–	57.0	54.9	59.1
Training-free / Zero-shot Methods									
SDAM-Qwen2.5-VL (Zhu et al., 2026)	61.5	59.0	63.9	54.6	51.8	57.5	58.0	55.4	60.7
CoT-RVS-GPT-4o (Kao et al., 2026)	–	–	–	–	–	–	55.9	52.8	59.0
Refer-Agent-Ovis2.5-9B (Jiang et al., 2026b)	65.0	62.3	67.7	57.7	54.3	61.0	61.3	58.3	64.3
RL-based Methods									
Veason-R1-Qwen2.5-VL-7B (Gong et al., 2026)	63.6	60.7	66.5	59.0	55.8	62.2	61.3	58.2	64.4
VideoSeg-R1-Qwen2.5-VL-7B (Xu et al., 2026)	–	–	–	–	–	–	61.1	58.2	64.0
VideoSEG-O3-Qwen3-VL-2B (Dai et al., 2026)	67.5	65.0	70.0	62.0	59.0	65.0	64.8	62.0	67.5
VideoSEG-O3-Qwen3-VL-4B (Dai et al., 2026)	70.3	67.7	72.9	65.1	62.4	67.9	67.7	65.0	70.4
VideoSeg-R1-Qwen2.5-VL-3B (Xu et al., 2026)	–	–	–	–	–	–	58.6	56.4	60.8
RCoT-Seg-Qwen2.5-VL-3B (Wen et al., 2026)	64.3	61.6	66.9	58.2	54.9	61.6	61.2	58.2	64.3
RCoT-Seg-Qwen2.5-VL-7B (Wen et al., 2026)	65.9	63.2	68.7	59.7	56.6	62.8	62.8	59.9	65.7
FaSTra-Qwen3-VL-2B	67.6	65.5	69.6	62.1	58.6	65.6	64.8	62.1	67.6
FaSTra-Qwen3-VL-4B	**72.7**	**70.3**	**75.1**	**67.0**	**64.3**	**69.7**	**69.8**	**67.3**	**72.4**

Table 3: Ablation study on TDF selector and CT router, where "Nearest" denotes directly assigning each frame the propagation result from its temporally closest keyframe.

Configuration		Segmentation		
Selector	Router	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
Baseline Model: Sa2VA (Yuan et al., 2026) w/ single keyframe, w/o router		73.7	81.0	77.4
Uniformly Sampled Keyframes	Nearest	68.7	76.0	72.4
	CT Router w/o Switching Penalty	71.9	79.1	75.5
	CT Router w/ Switching Penalty	72.4	79.9	76.1
Dynamic Keyframes w/o Time Penalty	Nearest	74.1	81.4	77.7
	CT Router w/o Switching Penalty	**75.0**	81.8	**78.4**
	CT Router w/ Switching Penalty	74.8	81.8	78.3
Dynamic Keyframes w/ Time Penalty	Nearest	73.7	81.0	77.4
	CT Router w/o Switching Penalty	74.7	81.8	78.2
	CT Router w/ Switching Penalty	74.8	**82.0**	**78.4**

Table 4: Ablation study on training stage.

Variant	Training Methods			Segmentation		
	SFT	FI-GRPO	KD	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
Baseline				73.7	81.0	77.4
Stage I	✓			75.2	82.5	78.8
Stage II	✓	✓		76.4	83.4	79.9
Stage III	✓	✓	✓	**78.5**	**85.6**	**82.0**

Table 5: Ablation study on distillation components.

Distillation Loss			Segmentation		
Quality	Traj.	Tri-level	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
✓			74.6	81.7	78.2
✓	✓		75.1	82.3	78.7
✓	✓	✓	**78.5**	**85.6**	**82.0**

over the single-keyframe baseline, indicating that multiple trajectories alone cannot fully exploit their temporal complementarity. In contrast, the CT router dynamically selects trajectories according to frame-wise reliability, improving $\mathcal{J}\&\mathcal{F}$ from 77.4 to 78.4 with temporally diverse keyframes. Although removing the switching penalty yields comparable accuracy, we retain it to discourage unnecessary trajectory switching and promote temporally coherent routing.

Table 6: Ablation study on the candidate frame segmentor and propagator. GFLOPs and FPS are measured per frame.

Segmentation Model		Segmentation			Efficiency	
Frame Segmentor	Propagator	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	GFLOPs	FPS
EdgeTAM	EdgeTAM	1.4	1.8	1.6	**351.9**	46.4
SAM2	SAM2	70.0	79.5	74.7	3813.1	15.8
EdgeTAM	SAM2	31.8	39.8	35.8	3589.7	25.5
SAM2	EdgeTAM	**74.8**	**82.0**	**78.4**	580.5	43.4

Table 7: Ablation study on RL optimization components. All evaluations are conducted without post-processing to maintain consistency.

Variant	Reward Term					Segmentation		
	$\mathcal{R}^{\text{rel}}$	$\mathcal{R}^{\text{gain}}$	$\mathcal{R}^{\text{down}}$	$\mathcal{R}^{\text{tea}}$	$\mathcal{R}^{\text{set}}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
Quality only	✓	–	–	–	–	77.3	84.2	80.8
+ Headroom Gain	✓	✓	–	–	–	77.4	84.5	80.9
+ Downside	✓	✓	✓	–	–	77.5	84.5	81.0
+ Teacher Overlap	✓	✓	✓	✓	–	77.6	84.6	81.1
Joint reward	✓	✓	✓	✓	✓	78.1	85.3	81.7



Figure 4: Qualitative comparisons between **FaSTra** and representative RVOS networks.

Table 6 further examines the trade-off between initial mask quality and propagation efficiency. Using SAM2 (Ravi et al., 2025) for frame segmentation and EdgeTAM (Zhou et al., 2025) for propagation achieves the best performance, outperforming the all-SAM2 configuration by 3.7 points while increasing throughput from 15.8 to 43.4 FPS. In contrast, replacing SAM2 initialization with EdgeTAM causes a substantial performance drop even with SAM2 propagation, indicating that reliable initial masks are more critical than a heavyweight propagator. These results support our asymmetric design of strong frame-level segmentation followed by lightweight temporal propagation.

Training approach ablations. Table 4 shows that all three training stages contribute to performance improvement, with $\mathcal{J}\&\mathcal{F}$ progressively increasing from 77.4 to 82.0. Table 5 further validates the distillation design. The optimal performance is achieved when combining quality, trajectory, and trilevel distillation. Similarly, the ablations in Table 7 show that all RL components contribute to more effective routing optimization. These results demonstrate that our training strategy progressively strengthens keyframe selection quality and dynamic routing capability.

4.4 Qualitative Study

As shown in Fig. 4, **FaSTra** remains robust in challenging scenes involving blurred video, close interaction, occlusion, and substantial pose changes. Sa2VA (Yuan et al., 2026) and VideoSEG-O3 (Dai et al., 2026) frequently confuse the target with the interacting player or fail to maintain a stable mask, whereas **FaSTra** consistently grounds and tracks the correct target across the sequence. Notably, the different mask colors indicate predictions selected from different trajectories, illustrating how the CT router dynamically exploits their temporal complementarity to preserve target identity. A remaining limitation is imperfect boundary delineation around limbs and heavily overlapping regions, which motivates future work on finer boundary-aware modeling. More qualitative results are provided in Appendix F.

5 Conclusion and Future Work

We proposed **FaSTra** framework, which reformulates keyframe selection as frame-wise routing over complementary trajectories. Temporally diverse keyframes and lightweight propagation enable efficient modeling, while BoT distillation and FI-GRPO strengthen propagation and align routing with frame-level segmentation gains. Experiments demonstrate strong accuracy with favorable efficiency. Future work will explore adaptive trajectory budgets and more robust routing under prolonged occlusion. Additional architecture, training, and evaluation details are provided in the supplement.

AI USE STATEMENT

Generative AI tools were used solely for language polishing and improving the clarity and fluency of the manuscript. They were not used to generate research ideas, design the methodology, conduct experiments, analyze results, or draw scientific conclusions. All technical content was produced and verified by the authors.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide detailed descriptions of the training pipeline, optimization hyperparameters, trainable parameter subsets, sampling strategy, and inference configuration. Specifically, we report the three-stage training procedure, optimizer settings, learning rates, update counts, gradient handling, FI-GRPO sampling protocol, propagation and route distillation settings, as well as the frozen and trainable modules. We also specify the inference-time frame sampling, keyframe selection, trajectory routing, switching penalty, numerical precision, visual-token range, and mask threshold. Together with the model architecture and benchmark settings reported in the main paper and supplementary material, these details provide sufficient information to reproduce our training and evaluation pipeline.

REFERENCES

Zechen Bai, Tong He, Haiyang Mei, Pichao Wang, Ziteng Gao, Joya Chen, Lei Liu, Zheng Zhang, and Mike Z Shou. One token to seg them all: Language instructed reasoning segmentation in videos. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 6833–6859, 2024.

Yuxuan Cai, Jiangning Zhang, Haoyang He, Xinwei He, Ao Tong, Zhenye Gan, Chengjie Wang, Zhucun Xue, Yong Liu, and Xiang Bai. LLaVA-KD: A Framework of Distilling Multimodal Large Language Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 239–249. IEEE, 2025.

Ho Kei Cheng and Alexander G Schwing. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *European conference on computer vision*, pp. 640–658. Springer, 2022.

Wenxuan Cheng, Ming Dai, Huimin Lu, and Wankou Yang. DeRVOS: Decoupling consistent trajectory generation and multimodal understanding for referring video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24651–24662, 2026.

Suhwan Cho, Seunghoon Lee, Minhyeok Lee, Jungho Lee, and Sangyoun Lee. Find first, track next: Decoupling identification and propagation in referring video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 3873–3883, 2025.

Claudia Cuttano, Gabriele Trivigno, Gabriele Rosi, Carlo Masone, and Giuseppe Averta. SAM-WISE: Infusing Wisdom in SAM2 for Text-Driven Video Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3395–3405. IEEE, 2025.

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. *Computing Research Repository (CoRR)*, 2026.

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 (ICCV)*, pp. 2694–2703, 2023.

Shuangrui Ding, Rui Qian, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Yuwei Guo, Dahua Lin, and Jiaqi Wang. SAM2Long: Enhancing SAM 2 for Long Video Segmentation with a Training-Free Memory Tree. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13614–13624. IEEE, 2025.

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 (CVPR)*, pp. 29183–29192. IEEE, 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 (CVPR)*, pp. 3835–3844, 2026.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge Distillation of Large Language Models. In *International Conference on Learning Representations (ICLR)*, volume 2024, pp. 32694–32717, 2024.

Weiyu Guo, Ziyang Chen, Shaoguang Wang, Jianxiang He, Yijie Xu, Jinhui Ye, Ying Sun, and Hui Xiong. Logic-in-frames: Dynamic keyframe search via visual semantic-logical verification for long video understanding. *Advances in Neural Information Processing Systems*, 38:124389–124422, 2026.

Yuk Heo, Yeong Jun Koh, and Chang-Su Kim. Guided interactive video object segmentation using reliability-based attention maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7318–7326. IEEE, 2021.

Jihwan Hong and Jaeyoung Do. VIRST: Video-Instructed Reasoning Assistant for SpatioTemporal Segmentation. *Computing Research Repository (CoRR)*, 2026.

Li Hu, Peng Zhang, Bang Zhang, Pan Pan, Yinghui Xu, and Rong Jin. Learning position and target consistency for memory-based video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4142–4152. IEEE, 2021.

Xijie Huang, Chengming Xu, Donghao Luo, Xiaobin Hu, Peng Tang, Xu Peng, Jiangning Zhang, Chengjie Wang, and Yanwei Fu. Ffp-300k: Scaling first-frame propagation for generalizable video editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23172–23181, 2026.

Dengyang Jiang, Liuzhuozheng Li, Sizhe Dang, Chengzu Li, Harry Yang, Guang Dai, Mengmeng Wang, Jingdong Wang, et al. Deforming videos to masks: Flow matching for referring video segmentation. In *International Conference on Learning Representations (ICLR)*, volume 2026, pp. 65586–65601, 2026a.

Haichao Jiang, Tianming Liang, Wei-Shi Zheng, and Jian-Fang Hu. Refer-Agent: A Collaborative Multi-Agent System with Reasoning and Reflection for Referring Video Object Segmentation. *Computing Research Repository (CoRR)*, 2026b.

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 (ICLR)*, volume 2026, pp. 58807–58830, 2026.

Anna Khoreva, Anna Rohrbach, and Bernt Schiele. Video object segmentation with language referring expressions. In *Asian Conference on Computer Vision (ACCV)*, volume 11364 of *Lecture Notes in Computer Science*, pp. 123–141. Springer, 2018. doi: 10.1007/978-3-030-20870-7_8.

Jihun Kim, Hoyong Kwon, Hyeokjun Kweon, and Kuk-Jin Yoon. Bootstrapping video semantic segmentation model via distillation-assisted test-time adaptation. *Computing Research Repository (CoRR)*, 2026a.

Jiwan Kim, Kibum Kim, Sangwoo Seo, and Chanyoung Park. CompoDistill: Attention Distillation for Compositional Reasoning in Multimodal LLMs. In *International Conference on Learning Representations (ICLR)*, volume 2026, pp. 71814–71844, 2026b.

Seongchan Kim, Woojeong Jin, Sangbeom Lim, Heeji Yoon, Hyunwook Choi, and Seungryong Kim. Referring video object segmentation via language-aligned track selection. *Computing Research Repository (CoRR)*, 2024.

Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

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 (CVPR)*, pp. 9579–9589, 2024.

Xiaoxiao Li and Chen Change Loy. Video object segmentation with joint re-identification and attention-aware mask propagation. In *European Conference on Computer Vision*, pp. 93–110. Springer, 2018.

Yifan Li, Yingda Yin, Lingting Zhu, Weikai Chen, Shengju Qian, Xin Wang, and Yanwei Fu. ReVSeg: Incentivizing the reasoning chain for video segmentation with reinforcement learning. *Computing Research Repository (CoRR)*, 2025.

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 (ICCV)*, pp. 20009–20019. IEEE, 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 (CVPR)*, pp. 8658–8667. IEEE, 2025.

Ye Liu, Zongyang Ma, Junfu Pu, Zhongang Qi, Yang Wu, Ying Shan, and Chang Chen. UniPixel: Unified object referring and segmentation for pixel-level visual reasoning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025.

Roy Miles, Mehmet Kerim Yucel, Bruno Manganelli, and Albert Saa-Garriga. MobileVOS: Real-Time Video Object Segmentation Contrastive Learning Meets Knowledge Distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10480–10490. IEEE, 2023.

Shehan Munasinghe, Hanan Gani, Wenqi Zhu, Jiale Cao, Eric Xing, Fahad Shahbaz Khan, and Salman Khan. VideoGLaMM: A Large Multimodal Model for Pixel-Level Visual Grounding in Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19036–19046. IEEE, 2025.

Alexey Nekrasov, Ali Athar, Daan de Geus, Alexander Hermans, and Bastian Leibe. Sa2VA-i: Improving Sa2VA results with consistent training and inference. *Computing Research Repository (CoRR)*, 2025.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. SAM 2: Segment Anything in Images and Videos. In *International Conference on Learning Representations (ICLR)*, volume 2025, pp. 28085–28128, 2025.

Fu Rong, Meng Lan, Qian Zhang, and Lefei Zhang. MPG-SAM 2: Adapting SAM 2 with mask priors and global context for referring video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 23979–23989. IEEE, 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 (ECCV)*, volume 12360 of *Lecture Notes in Computer Science*, pp. 208–223. Springer, 2020. doi: 10.1007/978-3-030-58555-6_13.

Yanyan Shao, Shuting He, Gengze Zhou, Qi Ye, Xiufang Shi, Jiming Chen, and Qi Wu. Towards reasoning-centric video object segmentation via multi-modal large language models. *IEEE Transactions on Image Processing*, pp. 8206–8218, 2026.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *Computing Research Repository (CoRR)*, 2024.

Fangxun Shu, Yue Liao, Lei Zhang, Le Zhuo, Chenning Xu, Guanghao Zhang, Haonan Shi, Weilong Dai, Zhelun Yu, Wanggui He, et al. LLaVA-MoD: Making LLaVA Tiny via MoE-Knowledge Distillation. In *International Conference on Learning Representations (ICLR)*, volume 2025, pp. 9386–9404, 2025.

Mingyang Song and Mao Zheng. A survey of on-policy distillation for large language models. *Computing Research Repository (CoRR)*, 2026.

Yijun Tian, Yikun Han, Xiusi Chen, Wei Wang, and Nitesh V Chawla. Beyond answers: Transferring reasoning capabilities to smaller llms using multi-teacher knowledge distillation. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining (WSDM)*, pp. 251–260, 2025.

Andrew Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE transactions on Information Theory*, 13(2):260–269, 1967.

Haochen Wang, Qirui Chen, Cilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, Weidi Xie, and Stratis Gavves. Object-centric video question answering with visual grounding and referring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22274–22284, 2025.

Cong Wei, Yujie Zhong, Haoxian Tan, Yong Liu, Jie Hu, Dengjie Li, Zheng Zhao, and Yujiu Yang. HyperSeg: Hybrid Segmentation Assistant with Fine-grained Visual Perceiver. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8931–8941. IEEE, 2025a.

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 (ICCV)*, pp. 20193–20203. IEEE, 2025b.

Junwei Wen, Deshui Miao, Guangming Lu, Xin Li, and Wenjie Pei. RCoT-Seg: Reinforced chain-of-thought for video reasoning and segmentation. *Computing Research Repository (CoRR)*, 2026.

Kai Xu and Angela Yao. Accelerating video object segmentation with compressed video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1332–1341. IEEE, 2022.

Zishan Xu, Yifu Guo, Yuquan Lu, Fengyu Yang, Junxin Li, and Lihua Cai. VideoSeg-R1: Reasoning Video Object Segmentation via Reinforcement Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 40, pp. 11496–11504, 2026.

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 (ECCV)*, pp. 78–97, 2024.

Haobo Yuan, Xiangtai Li, Tao Zhang, Yueyi Sun, Zilong Huang, Shilin Xu, Shunping Ji, Yunhai Tong, Lu Qi, Jiashi Feng, et al. Sa2VA: Marrying SAM2 with MLLM for dense grounded understanding of images and videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026. doi: 10.1109/TPAMI.2026.3720379.

Xiangyu Zeng, Zhiqiu Zhang, Yuhan Zhu, Xinhao Li, Zikang Wang, Changlian Ma, Qingyu Zhang, Zizheng Huang, Kun Ouyang, Tianxiang Jiang, et al. Video-o3: Native interleaved clue seeking for long video multi-hop reasoning. *Computing Research Repository (CoRR)*, 2026.

Rongkun Zheng, Lu Qi, Xi Chen, Yi Wang, Kun Wang, and Hengshuang Zhao. ViLLa: Video reasoning segmentation with large language model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 23667–23677, 2025.

Hao Zhong, Muzhi Zhu, Zongze Du, Zheng Huang, Canyu Zhao, Mingyu Liu, Wen Wang, Hao Chen, and Chunhua Shen. Omni-r1: Reinforcement learning for omnimodal reasoning via two-system collaboration. *Advances in Neural Information Processing Systems*, 38:45266–45297, 2026.

Chong Zhou, Chenchen Zhu, Yunyang Xiong, Saksham Suri, Fanyi Xiao, Lemeng Wu, Raghuraman Krishnamoorthi, Bo Dai, Chen Change Loy, Vikas Chandra, et al. EdgeTAM: On-Device Track Anything Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13832–13842. IEEE, 2025.

Jiawen Zhu, Zhi-Qi Cheng, Jun-Yan He, Chenyang Li, Bin Luo, Huchuan Lu, Yifeng Geng, and Xuansong Xie. Tracking with human-intent reasoning. *Computing Research Repository (CoRR)*, 2023.

Zhengtong Zhu, Jiaqing Fan, Zhixuan Liu, and Fanzhang Li. Training-free spatio-temporal decoupled reasoning video segmentation with adaptive object memory. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 40, pp. 14022–14030, 2026.

Yunzhi Zhuge, Sitong Gong, Lu Zhang, Qi Xu, Wenda Zhao, Jin Zhan, and Huchuan Lu. Context-Infused Trajectories: Enhancing context and frame consistency in reasoning video object segmentation. *IEEE Transactions on Image Processing*, 2026.

APPENDIX

CONTENTS

A Symbol Definitions **15**

B Network Architecture **17**
B.1 Temporally-Diverse Keyframe Selector 17
B.2 RL Training Objectives 17

C Overall Loss Function **18**

D Dataset Information **18**

E Additional Quantitative Experimental Results **20**
E.1 Implementation Details 20
E.2 Additional State-of-the-Art Comparisons 20
E.3 Additional Comparisons of Model Architectures and Inference Efficiency 24
E.4 Additional Analysis of Distillation Loss Convergence 26
E.5 Additional Analysis of FI-GRPO Reward Dynamics 26

F Additional Qualitative Experimental Results **26**
F.1 Prediction Results 26
F.2 Failure Cases and Limitation 29

This appendix provides additional details and analyses to complement the main paper. We first summarize the notation used throughout the paper, followed by detailed descriptions of the network architecture and training objectives. We then provide dataset information and implementation details, including the training and inference configurations. Additional quantitative experiments further compare **FaSTra** with existing methods across different backbones and model scales, and analyze its architecture and inference efficiency. Finally, we present additional analyses of the distillation and FI-GRPO optimization dynamics, together with qualitative results and failure cases, to provide further insight into the behavior and limitations of **FaSTra**.

A SYMBOL DEFINITIONS

In this section, Table 8 provides a comprehensive summary of the notations and their corresponding physical or semantic meanings used in the **FaSTra** framework.

Table 8: Key notations used throughout the **FaSTra** framework.

Notations	Descriptions
General Notations and Framework Overview	
T, N, K	Number of video frames, sampled candidate frames, and selected keyframes, respectively.
t, c_i, a_k	Current frame index, index of the i-th candidate frame, and index of the k-th selected keyframe.
$\mathcal{V} = \{\boldsymbol{I}_t\}_{t=1}^{T}$	Input video and its frame sequence.
$\mathcal{E}$	Natural-language query specifying the target object.
$\boldsymbol{e}$	Hidden embedding of the language-conditioned <SEG> token produced by the MLLM.
$\boldsymbol{Y}_t$	Ground-truth target mask at frame t.
$\mathcal{J}, \mathcal{F}$	Region similarity and contour accuracy; $\mathcal{J}\&\mathcal{F}$ is their mean.
$\mathcal{F}_{\text{segmentor}}$	SAM2 mask decoder used to predict an initial mask and object representations at a candidate frame.
$\mathcal{F}_{\text{selector}}$	Temporally-Diverse Keyframe selector that chooses complementary propagation seeds.
$\mathcal{F}_{\text{propagator}}$	Lightweight bidirectional video propagator that generates a full-video trajectory from each selected keyframe.
$\mathcal{F}_{\text{router}}$	Complementary Trajectory router that estimates frame-wise trajectory reliability.
Temporally-Diverse Keyframe Selector	
$\mathcal{C} = \{c_i\}_{i=1}^{N}$	Set of candidate-frame indices sampled from the video.

Continued on the next page

Table 8 continued

Notations	Descriptions
$\mathcal{A} = \{a_k\}_{k=1}^{K}$	Set of temporally diverse keyframe indices selected from $\mathcal{C}$.
$\boldsymbol{L}_{c_i}, \boldsymbol{p}_{c_i}, o_{c_i}$	Initial mask logits, object-pointer embedding, and object-presence score predicted at candidate frame c_i.
$u_c, s_c^{(k)}$	Raw utility of candidate c and its diversity-adjusted score at selection step k.
$\mathcal{A}_k$	Set of the first k selected keyframes.
$P_\theta(a_k = c \mid \mathcal{A}_{k-1})$	Selector policy for choosing candidate c as the k-th keyframe.
θ	Learnable parameters of the keyframe selector.
$\lambda_{\mathrm{div}}, \sigma_{\mathrm{div}}$	Strength and temporal range of the keyframe-diversity penalty.
Complementary Trajectory Router	
$\boldsymbol{\tau}_{a_k} = \{\tau_{a_k,t}\}_{t=1}^{T}$	Candidate trajectory propagated bidirectionally from keyframe a_k.
$\boldsymbol{L}_{a_k,t}$	Mask logits at frame t from the trajectory initialized at keyframe a_k.
$\boldsymbol{p}_{a_k,t}, o_{a_k,t}$	Object-pointer embedding and object-presence score at frame t for the trajectory initialized at a_k.
$l_{a_k,t}, q_{a_k,t}$	Router logit and calibrated quality score for trajectory a_k at frame t, where $q_{a_k,t} = \sigma(l_{a_k,t})$.
$\boldsymbol{Q} = [q_{a_k,t}]$	Frame-wise trajectory-quality matrix in $\mathbb{R}^{K\times T}$.
$\rho_{a_k,t}$	Normalized temporal offset $(t - a_k)/(T-1)$ between frame t and keyframe a_k.
$\boldsymbol{z} = \{z_t\}_{t=1}^{T}$	Routing path; $z_t \in \mathcal{A}$ identifies the trajectory selected at frame t.
$\boldsymbol{\tau} = \{\tau_{z_t,t}\}_{t=1}^{T}$	Final output trajectory stitched according to the routing path.
λ_{swi}	Penalty controlling the frequency of trajectory switches during Viterbi decoding.
Best-of-Trajectories Teacher and Knowledge Distillation	
$\mathcal{T}, \mathcal{S}$	Superscripts identifying teacher and student quantities, respectively.
$\boldsymbol{Q}^{\mathcal{T}}, z_t^{\mathcal{T}}$	Ground-truth-evaluated teacher quality matrix and its best trajectory assignment at frame t.
$\boldsymbol{\tau}^{\mathcal{T}}, \boldsymbol{L}_t^{\mathcal{T}}$	Stitched Best-of-Trajectories teacher trajectory and its mask logits at frame t.
$\boldsymbol{M}_t^{\mathcal{T},\mathcal{S}}$	Teacher or student soft mask, defined by applying the sigmoid function to $\boldsymbol{L}_t^{\mathcal{T},\mathcal{S}}$.
$\boldsymbol{H}_t^{\mathcal{T},\mathcal{S}}$	Teacher or student propagation-memory representation at frame t.
$\beta_t, \boldsymbol{w}_t$	Frame weight derived from the best–worst trajectory gap and pixel-wise weight emphasizing teacher foreground and boundaries.
$\Delta\boldsymbol{M}_t$	Temporal soft-mask difference $\boldsymbol{M}_t - \boldsymbol{M}_{t-1}$.
$\mathcal{D}_H, \mathcal{D}_p, \mathcal{D}_o$	Distances for memory features, object pointers, and object-presence scores.
$\boldsymbol{G}(\boldsymbol{H}), \mathrm{OffDiag}(\cdot)$	Temporal cosine-similarity Gram matrix and the operator that removes its diagonal.
$\mathcal{L}_{\mathrm{ml}}$	Prediction-level mask-logit distillation loss.
$\mathcal{L}_{\mathrm{mf}}, \mathcal{L}_{\mathrm{obp}}, \mathcal{L}_{\mathrm{obs}}$	State-level losses for memory features, object pointers, and object-presence scores.
$\mathcal{L}_{\mathrm{tm}}, \mathcal{L}_{\mathrm{gram}}$	Temporal mask-difference and temporal-relation distillation losses.
$\pi_{a_k,t}, \boldsymbol{\pi}_t$	Student routing probability for trajectory a_k and the probability vector over all trajectories at frame t.
$\boldsymbol{q}_t^{\mathcal{T}}, \boldsymbol{q}_t$	Teacher and student trajectory-quality vectors at frame t.
Ω^+	Set of frames in which the target object is present.
κ	Temperature used for teacher soft-distribution matching.
$\mathcal{L}_{\mathrm{cal}}, \mathcal{L}_{\mathrm{hard}}, \mathcal{L}_{\mathrm{soft}}, \mathcal{L}_{\mathrm{rank}}$	Quality calibration, hard-assignment, soft-distribution, and pairwise-ranking losses.
$\varsigma_t^{\mathcal{T}}, \varsigma_t$	Teacher switching indicator and predicted switching probability at frame t.
$\boldsymbol{\xi}_t^{\mathcal{T}}$	One-hot teacher transition distribution for two consecutive routing assignments.
$\otimes$	Outer product used to form the predicted transition distribution.
$\mathcal{L}_{\mathrm{swit}}, \mathcal{L}_{\mathrm{tran}}$	Switching-probability and transition-distribution distillation losses.
GRPO-Based Frame-Impact Optimization	
a_{base}	Keyframe whose single propagated trajectory has the highest mean quality over the video.
$\bar{q}_t, s_t, h_t$	Mean quality, quality standard deviation, and improvement headroom across trajectories at frame t.
$\mathcal{R}_t^{\mathrm{rel}}, \mathcal{R}_t^{\mathrm{gain}}, \mathcal{R}_t^{\mathrm{down}}, \mathcal{R}_t^{\mathrm{tea}}$	Relative-quality, headroom-gain, downside, and teacher-overlap reward terms.
$\lambda_{\mathrm{g}}, \lambda_{\mathrm{d}}, \lambda_{\mathrm{o}}$	Weights of the gain, downside, and teacher-overlap reward terms.
ϵ	Small positive constant used for numerical stability in reward normalization.
$G, \boldsymbol{z}^g$	Number of sampled routing assignments and the g-th sampled routing path.

Continued on the next page

Table 8 continued

Notations	Descriptions
$A_{g,t}, \varrho_{g,t}$	Group-relative advantage and policy probability ratio for sample g at frame t.
ε	Clipping radius in the GRPO surrogate objective.
$\mathcal{J}_{\text{route}}, \mathcal{L}_{\text{router}}$	Clipped routing objective and total router-training loss.
$\phi, \boldsymbol{\pi}_t^{\text{old}}, \boldsymbol{\pi}_{\text{ref}}$	Router parameters, previous-policy distribution, and reference-policy distribution.
$\lambda_H, \lambda_{\text{KL}}$	Weights of policy entropy and KL regularization.
$\mathcal{R}^{\text{set}}(\mathcal{A})$	Set-level reward measuring the upper-bound quality supplied by selected keyframes.
$\mathcal{J}_{\text{key}}, \bar{\mathcal{R}}$	Keyframe-selection objective and within-video reward baseline.

B Network Architecture

B.1 Temporally-Diverse Keyframe Selector

Trajectories initialized from different keyframes tend to exhibit complementary reliability across temporal regions. This motivates selecting multiple propagation seeds to capture such temporal complementarity. However, top-K selection based solely on the initial object scores o_{c_i} is suboptimal: frame-level confidence does not necessarily indicate long-term propagation quality, while high-confidence candidates can be temporally redundant and concentrated in easily visible regions. These observations motivate a selector that jointly accounts for candidate utility and temporal diversity in selecting the K seed frames.

For each candidate c_i, the selector encodes its mask quality, region geometry, temporal position, and cross-candidate consistency, and relates these cues together with its object representation to the query semantics. The resulting query-conditioned correction is bounded and combined with the normalized object-presence confidence to produce a raw utility score u_{c_i}. Then, we select keyframes sequentially. Starting from the raw utilities, the first keyframe is selected without a diversity penalty. After each selection, the score of every remaining candidate is reduced according to its temporal proximity to the newly selected keyframe:

$$\begin{aligned} s_c^{(1)} &= u_c, \qquad \mathcal{A}_0 = \varnothing, \\ s_c^{(k)} &= s_c^{(k-1)} - \frac{2\lambda_{\text{div}}}{K-1} \exp\!\left(-\frac{|c - a_{k-1}|}{\sigma_{\text{div}}(T-1)}\right), \quad k = 2, \dots, K, \\ P_\theta(a_k = c \mid \mathcal{A}_{k-1}) &= \frac{\exp\left(s_c^{(k)}\right)}{\sum_{\tilde{c} \in \mathcal{C} \setminus \mathcal{A}_{k-1}} \exp\left(s_{\tilde{c}}^{(k)}\right)}, \quad c \in \mathcal{C} \setminus \mathcal{A}_{k-1}, \end{aligned} \tag{7}$$

where $\mathcal{A}_{k-1} = \{a_1, \dots, a_{k-1}\}$. Thus, after selecting a_1, its distance penalty is subtracted before selecting a_2; after selecting a_2, a second penalty is subtracted before selecting a_3, and so forth. During training, each a_k is sampled from Eq. 7; during inference, it is the highest-scoring remaining candidate. The selected candidate is then removed from the support. The parameters λ_{div} and σ_{div} control the penalty strength and temporal range, respectively.

B.2 RL Training Objectives

Routing is optimized at the trajectory level. We sample G routing assignments $\{\boldsymbol{z}^g\}_{g=1}^G$ per video and always inject the teacher trajectory assignment $\boldsymbol{z}^{\mathcal{T}}$, a baseline that always uses the best single trajectory, and the current greedy trajectory, so the gain is measured against not routing; rewards are standardized within the group on every frame and the objective is clipped with range ε:

$$\begin{aligned} A_{g,t} &= \frac{\mathcal{R}_t^{\text{route}}(z_t^g) - \text{mean}_{g'} \mathcal{R}_t^{\text{route}}(z_t^{g'})}{\text{std}_{g'} \mathcal{R}_t^{\text{route}}(z_t^{g'})}, \qquad \varrho_{g,t} = \pi_{z_t^g,t} / \pi^{\text{old}}_{z_t^g,t}, \\ \mathcal{J}_{\text{route}}(\phi) &= \mathbb{E}_{g,t}\Big[\min\left(\varrho_{g,t} A_{g,t},\ \text{clip}(\varrho_{g,t}, 1-\varepsilon, 1+\varepsilon) A_{g,t}\right)\Big], \end{aligned} \tag{8}$$

$$\mathcal{L}_{\text{router}} = -\mathcal{J}_{\text{route}} - \lambda_H \mathcal{H}[\pi_\phi] + \mathcal{L}_{\text{rd}} + \lambda_{\text{KL}} \,\text{KL}\big(\pi_\phi \,\|\, \pi_{\text{ref}}\big), \tag{9}$$

with $\boldsymbol{\pi}_t^{\text{old}}$ from the previous iterate. Normalizing per frame judges the same trajectory separately on different frames, so a correct switch is not masked by the rest, which is what frame impact means. In Eq. 9 distillation supplies an absolute reference and RL a relative one consistent with the metric, while entropy and KL maintain exploration and prevent drift.

Keyframe selection is optimized at the set level. For compactness, write $P_\theta(\mathcal{A}) = \mathcal{F}_{\text{selector}}(\mathcal{A} \mid \boldsymbol{e}, \mathcal{C})$:

$$\begin{aligned} \mathcal{J}_{\text{key}}(\theta) &= \mathbb{E}_{\mathcal{A}\sim P_\theta}\big[\mathcal{R}^{\text{set}}(\mathcal{A})\big], \\ \nabla_\theta \mathcal{J}_{\text{key}} &= \mathbb{E}_{\mathcal{A}\sim P_\theta}\Big[\big(\mathcal{R}^{\text{set}}(\mathcal{A}) - \bar{\mathcal{R}}\big)\nabla_\theta \log P_\theta(\mathcal{A})\Big]. \end{aligned} \tag{10}$$

where $\bar{\mathcal{R}}$ is the mean reward of the other keyframe sets sampled for the same video. Each sampled set is passed through $\mathcal{F}_{\text{propagator}}$ to construct its trajectories.

C Overall Loss Function

The main text introduces the individual terms of each objective but, for space, not the weighted sums they enter. We restore those sums and give the value of every coefficient, and then describe how each stage converges. Table 9 collects every coefficient in one place. Every coefficient is a constant fixed before training rather than a schedule, and each stage of Table 4 is trained to completion, with its own optimizer state, before the next one begins.

Stage I: Supervised Fine-Tuning. This stage follows the protocol of Sa2VA (Yuan et al., 2026) and introduces no term of our own. Its objective is the token-level cross-entropy of the generated response together with a sigmoid cross-entropy and a Dice term on the predicted mask, combined as $\mathcal{L}_{\text{sft}} = \mathcal{L}_{\text{text}} + 2.0\,\mathcal{L}_{\text{bce}} + 0.5\,\mathcal{L}_{\text{dice}}$. It establishes the mask token $\boldsymbol{e}$ that both of our heads read, and it is the only stage in which the multimodal large language model receives gradient: in Stages II and III it is frozen, and the parameters actually optimized are the propagator and the two lightweight heads of Sect. B.

Stage II: Frame-Impact GRPO. This stage optimizes two objectives with two separate optimizers, and they are never summed into a single scalar. The router objective is Eq. 9, in which $\mathcal{J}_{\text{route}}$ enters with unit weight while the two regularizers take $\lambda_H = 0.01$, and π_{ref} is the router frozen at the value it holds when the stage begins. Within the reward of Eq. 5 we set $\lambda_{\text{g}} = 0.5$, $\lambda_{\text{d}} = 1.0$ and $\lambda_{\text{o}} = 0.2$. The keyframe objective of Eq. 10 is maximized as $\mathcal{L}_{\text{key}} = -\mathcal{J}_{\text{key}}$ and carries no coefficient of its own. Finally, λ_{div} and σ_{div} of Eq. 7 are parameters of the selector's scoring function.

Stage III: Frame-Impact GRPO and Knowledge Distillation. This stage minimizes two sums, which together are the three columns of Table 5. Route distillation collects the four quality terms of Eq. 3, the two trajectory terms of Eq. 4, and one boundary-reweighted term:

$$\mathcal{L}_{\text{rd}} = \lambda_{\text{cal}}\mathcal{L}_{\text{cal}} + \lambda_{\text{hard}}\mathcal{L}_{\text{hard}} + \lambda_{\text{soft}}\mathcal{L}_{\text{soft}} + \lambda_{\text{rank}}\mathcal{L}_{\text{rank}} + \lambda_{\text{swit}}\mathcal{L}_{\text{swit}} + \lambda_{\text{tran}}\mathcal{L}_{\text{tran}}, \tag{11}$$

The first four terms are evaluated on the present set Ω^+, whereas $\mathcal{L}_{\text{swit}}$ and $\mathcal{L}_{\text{tran}}$ are evaluated on all $T-1$ adjacent frame pairs, since a switch into or out of an absent frame is still a switch. In Table 5 the *Quality* column is the first five terms and the *Traj.* column the last two.

Propagation distillation groups the six terms of Eq. 2 into the three levels at which they were defined, attaching one coefficient to each level rather than to each term, so that prediction, state and temporal information can be balanced against one another without re-tuning six quantities that live on different scales:

$$\mathcal{L}_{\text{pd}} = \lambda_{\text{pred}}\,\mathcal{L}_{\text{ml}} + \lambda_{\text{state}}\big(\mathcal{L}_{\text{mf}} + \mathcal{L}_{\text{obp}} + \mathcal{L}_{\text{obs}}\big) + \lambda_{\text{temp}}\big(\mathcal{L}_{\text{tm}} + \mathcal{L}_{\text{gram}}\big). \tag{12}$$

D Dataset Information

We evaluate **FaSTra** on four conventional referring video object segmentation benchmarks, including MeViS, Ref-YouTube-VOS, Ref-DAVIS17, and Ref-SAV, and further evaluate reasoning video object segmentation on ReVOS.

MeViS (Ding et al., 2023) is a motion-expression-guided video segmentation benchmark containing 2,006 videos, 8,171 target objects, and 28,570 motion expressions. Unlike conventional referring

Table 9: Loss terms and coefficients of the three training stages of **FaSTra**. Stage I follows the protocol of the base model and introduces no term of our own; Stages II and III optimize only the propagator and the two lightweight heads.

Stage	Objective	Term	Coefficient	Value
I. SFT	$\mathcal{L}_{\text{sft}}$	$\mathcal{L}_{\text{text}}$	–	1.0
		$\mathcal{L}_{\text{bce}}$	–	2.0
		$\mathcal{L}_{\text{dice}}$	–	0.5
II. FI-GRPO	$\mathcal{L}_{\text{router}}$, Eq. 9	$\mathcal{J}_{\text{route}}$	–	1.0
		$\mathcal{H}[\pi_\phi]$	λ_H	0.01
		$\mathcal{L}_{\text{rd}}$	–	1.0
		$\text{KL}(\pi_\phi \Vert \pi_{\text{ref}})$	λ_{KL}	0.05
	$\mathcal{R}^{\text{route}}$, Eq. 5	$\mathcal{R}^{\text{rel}}$	–	1.0
		$\mathcal{R}^{\text{gain}}$	λ_{g}	0.5
		$\mathcal{R}^{\text{down}}$	λ_{d}	1.0
		$\mathcal{R}^{\text{tea}}$	λ_{o}	0.2
	$\mathcal{L}_{\text{key}}$, Eq. 10	$-\mathcal{J}_{\text{key}}$	–	1.0
III. KD	$\mathcal{L}_{\text{rd}}$, Eq. 11	$\mathcal{L}_{\text{cal}}$	λ_{cal}	1.0
		$\mathcal{L}_{\text{hard}}$	λ_{hard}	1.0
		$\mathcal{L}_{\text{soft}}$	λ_{soft}	1.0
		$\mathcal{L}_{\text{rank}}$	λ_{rank}	1.0
		$\mathcal{L}_{\text{swit}}$	λ_{swit}	1.0
		$\mathcal{L}_{\text{tran}}$	λ_{tran}	1.0
	$\mathcal{L}_{\text{pd}}$, Eq. 12	$\mathcal{L}_{\text{ml}}$	λ_{pred}	1.0
		$\mathcal{L}_{\text{mf}} + \mathcal{L}_{\text{obp}} + \mathcal{L}_{\text{obs}}$	λ_{state}	1.0
		$\mathcal{L}_{\text{tm}} + \mathcal{L}_{\text{gram}}$	λ_{temp}	1.0

video datasets that often rely on static appearance cues, MeViS emphasizes object motion and temporal behavior, making it suitable for evaluating motion-aware target understanding.

Ref-YouTube-VOS (Seo et al., 2020) is a large-scale referring video object segmentation benchmark built on YouTube-VOS. It contains 3,978 videos, 7,451 objects, and approximately 15,000 referring expressions. Its diverse object categories, appearance changes, and occlusions make it a standard benchmark for evaluating language-guided target localization and temporally consistent segmentation.

Ref-DAVIS17 (Khoreva et al., 2018) augments DAVIS17 with natural-language referring expressions for individual target objects. Its densely annotated video sequences contain challenging appearance changes, occlusions, object interactions, and fine object boundaries, providing a complementary evaluation to the larger Ref-YouTube-VOS benchmark.

Ref-SAV (Yuan et al., 2026) is introduced with Sa2VA for referring video object segmentation in complex real-world scenes. It contains more than 72K automatically generated object expressions and 2K manually validated video objects, providing diverse scenes and object interactions for evaluating video segmentation models.

ReVOS (Yan et al., 2024) is designed for reasoning video object segmentation, where target objects can be specified by implicit language instructions requiring semantic reasoning, temporal understanding, and world knowledge. It contains 35,074 instruction–mask sequence pairs from 1,042 videos. Following the standard protocol, we report results on its *Referring*, *Reasoning*, and *Overall* subsets.

E Additional Quantitative Experimental Results

E.1 Implementation Details

E.1.1 Training Stage and Optimization.

We first perform supervised fine-tuning (SFT) to establish the base segmentation capability, following the training data composition, data processing, and optimization protocol of Sa2VA (Yuan et al., 2026). We construct the BoT teacher and apply FI-GRPO to optimize the router for more effective dynamic routing. Finally, we perform knowledge distillation to further enhance the model's propagation and routing capabilities. During training Stages II and III, the language model remains frozen. We optimize only the lightweight propagator, with 8.99 M of its 13.94 M parameters receiving gradients, together with the selector and router, which contain 0.285 M parameters in total. Overall, the trainable components account for only about 0.18% of the 5.07 B parameters loaded by the deployed system.

All training stages use AdamW with global gradient-norm clipping at 1.0. Updates with non-finite gradient norms are skipped to prevent numerical instability from affecting the optimizer state. Stage II is trained for 1500 updates with separate optimizers for the router and selector, using learning rates of 1×10^{-4} and 1×10^{-5}, respectively. At each update, we sample $G = 8$ routing assignments and additionally include the teacher assignment, the best single-trajectory baseline, and the current greedy assignment, allowing the advantage in Eq. 8 to be computed against a non-routing baseline. Stage III consists of propagation and route distillation. Propagation distillation is performed for 2000 updates with a learning rate of 1×10^{-4}, using one 12-frame video sample per update and a 50-update linear warmup. During this stage, the propagator image encoder is frozen and its normalization layers remain in inference mode, so optimization focuses on the temporal memory pathway. Route distillation is then conducted for another 2000 updates at 1×10^{-4}, using eight cached routing trajectories per update.

E.1.2 Inference configuration.

At inference, the language model is decoded once using five uniformly sampled frames. For each video, we uniformly sample $N = 9$ candidate frames and select $K = 3$ keyframes using the sequential policy in Eq. 7. These keyframes generate three propagation trajectories, which are routed frame by frame using Eq. 1 with a switching penalty of $\lambda_{\text{swi}} = 0.35$. We impose neither a limit on the number of switches nor a minimum segment length. Keyframe masks are produced by the original SAM2 decoder from the base model, while temporal propagation uses the lightweight propagator. The language model runs in `bfloat16`, whereas the propagator uses full precision. Each frame is represented by 392-1,568 visual tokens, and the final masks are obtained by thresholding the predicted probabilities at 0.5.

E.2 Additional State-of-the-Art Comparisons

Table 10 summarizes comparisons on four conventional benchmarks, including MeViS (Ding et al., 2023), Ref-YouTube-VOS (Seo et al., 2020), Ref-DAVIS17 (Khoreva et al., 2018), and Ref-SAV (Yuan et al., 2026). Table 11 further reports results on the reasoning-oriented ReVOS benchmark (Yan et al., 2024), which places greater emphasis on semantic reasoning and complex target identification. For completeness, we briefly introduce all compared method families below, while grouping different backbone and model-scale variants of the same method together. In addition to the default Qwen3-VL configuration, we further evaluate **FaSTra** with InternVL3 at two model scales, namely **FaSTra**-InternVL3-2B and **FaSTra**-InternVL3-8B, to examine whether the proposed framework generalizes across different MLLM backbones and capacities. These additional results provide a broader evaluation of the robustness and transferability of our multi-trajectory design beyond a single language-model family.

SAMWISE (Cuttano et al., 2025) adapts SAM2 to text-driven video segmentation by incorporating language guidance into video object tracking and segmentation. It aims to preserve the strong temporal segmentation ability of SAM2 while enabling text-conditioned target identification.

Table 10: Additional comparisons on video object segmentation benchmarks. GLUS^A denotes the Additional-SFT variant. Sa2VA-i results follow the 8-frame setting (Nekrasov et al., 2025).

Method	MeViS			Ref-YouTube-VOS			Ref-DAVIS17			Ref-SAV		
	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
Specialized Methods												
SAMWISE-RoBERTa (Cuttano et al., 2025)	49.5	46.6	52.4	69.2	67.8	70.6	70.6	67.4	74.5	–	–	–
ReferDINO (Liang et al., 2025)	49.3	44.7	53.9	69.3	67.0	71.5	68.9	65.1	72.9	–	–	–
MPG-SAM 2-BEiT (Rong et al., 2025)	53.7	50.7	56.7	73.9	71.7	76.1	72.4	68.8	76.0	–	–	–
FlowRVS-Wan2.1-T2V-1.3B (Jiang et al., 2026a)	51.1	47.6	54.6	69.6	67.1	72.1	73.3	68.4	78.2	–	–	–
DeRVOS (Cheng et al., 2026)	51.8	48.1	55.4	70.0	68.0	71.9	70.9	68.3	73.5	–	–	–
Training-free / Zero-shot Methods												
SDAM-Qwen2.5-VL (Zhu et al., 2026)	48.6	45.5	51.7	65.3	63.4	67.1	76.0	73.2	78.8	–	–	–
Refer-Agent-Ovis2.5-9B (Jiang et al., 2026b)	54.7	51.6	57.7	71.3	69.0	73.6	–	–	–	–	–	–
CoT-RVS-GPT-4o (Kao et al., 2026)	52.2	48.7	55.7	–	–	–	79.1	75.3	82.9	–	–	–
1–3B MLLM-based Methods												
Sa2VA-InternVL2-1B (Yuan et al., 2026)	41.7	–	–	65.3	–	–	72.3	–	–	–	–	–
Sa2VA-InternVL2.5-1B (Yuan et al., 2026)	47.0	–	–	68.0	–	–	69.5	–	–	–	–	–
Sa2VA-Qwen3-VL-2B (Yuan et al., 2026)	48.8	44.7	53.0	56.3	54.7	57.9	67.1	62.8	71.4	39.1	37.6	40.5
Sa2VA-i-InternVL2.5-1B (Nekrasov et al., 2025)	52.6	49.9	55.3	70.3	68.4	72.2	73.6	70.1	77.1	–	–	–
InstructSeg-Mipha-3B (Wei et al., 2025b)	–	–	–	67.5	65.4	69.5	71.1	67.3	74.9	–	–	–
VideoSEG-O3-Qwen3-VL-2B (Dai et al., 2026)	55.6	52.5	58.6	70.5	68.3	**72.7**	80.0	76.0	84.0	62.9	61.8	64.0
HyperSeg-Qwen2.5-VL-3B (Wei et al., 2025a)	–	–	–	–	–	68.5	–	–	71.2	–	–	–
FaSTra-InternVL3-2B	**61.0**	**57.5**	**64.5**	71.0	69.3	72.8	80.2	76.8	83.5	**67.1**	**66.0**	**68.1**
FaSTra-Qwen3-VL-2B	59.4	55.7	63.1	70.7	68.7	72.7	**82.0**	**78.0**	**86.0**	64.4	63.1	65.7
4–8B MLLM-based Methods												
VideoLISA-LLaVA-Phi3-V-3.8B (Bai et al., 2024)	44.4	41.3	47.6	63.7	61.7	65.7	68.8	64.9	72.7	–	–	–
VideoGLaMM-Phi3-Mini-3.8B (Munasinghe et al., 2025)	45.2	42.1	48.2	–	–	–	69.5	65.6	73.3	–	–	–
RViSeg-3.8B (Shao et al., 2026)	44.1	41.8	46.3	62.8	61.5	64.2	61.2	59.3	64.4	–	–	–
Sa2VA-InternVL2-4B (Yuan et al., 2026)	46.2	–	–	70.0	–	–	73.8	–	–	–	–	–
Sa2VA-InternVL2.5-4B (Yuan et al., 2026)	46.4	43.3	49.5	71.3	69.1	73.5	73.7	69.6	77.8	–	–	–
Sa2VA-Qwen3-VL-4B (Yuan et al., 2026)	55.8	51.4	60.2	67.0	65.3	68.7	75.5	71.1	79.9	42.2	40.9	43.5
Sa2VA-i-InternVL2.5-4B (Nekrasov et al., 2025)	56.6	53.9	59.3	73.2	71.0	75.4	78.6	75.1	82.1	–	–	–
GLUS^A-LLaVA-7B (Lin et al., 2025)	51.3	48.5	54.2	67.3	65.5	69.0	–	–	–	–	–	–
LISA-LLaVA-7B (Lai et al., 2024)	37.2	35.1	39.4	53.9	53.4	54.3	64.8	62.2	67.3	–	–	–
UniPixel-Qwen2.5-VL-7B (Liu et al., 2025)	55.8	53.2	58.3	71.0	69.5	72.4	76.4	72.7	80.1	69.0	68.5	69.6
VISA-Chat-UniVi-7B (Yan et al., 2024)	43.5	40.7	46.3	61.5	59.8	63.2	69.4	66.3	72.5	11.8	13.2	11.3
RViSeg-7B (Shao et al., 2026)	45.8	43.7	48.0	62.8	61.7	64.2	62.3	59.8	64.8	–	–	–
ViLLa-6B (Zheng et al., 2025)	49.4	46.5	52.3	67.5	64.6	70.4	74.3	70.6	78.0	–	–	–
CoT-RVS-online-7B (Kao et al., 2026)	45.9	42.7	49.1	–	–	–	73.9	70.4	77.5	–	–	–
Sa2VA-InternVL2-8B (Yuan et al., 2026)	46.9	–	–	70.7	–	–	75.2	–	–	50.0	48.3	51.7
Sa2VA-InternVL2.5-8B (Yuan et al., 2026)	51.5	–	–	72.3	–	–	75.9	–	–	–	–	–
Sa2VA-i-InternVL2.5-8B (Nekrasov et al., 2025)	59.5	56.6	62.4	73.9	71.7	76.3	79.1	75.6	82.7	–	–	–
VideoSEG-O3-Qwen3-VL-4B (Dai et al., 2026)	60.0	57.0	63.0	74.1	71.5	**76.6**	79.4	75.5	83.2	65.5	64.3	66.7
VIRST-VideoChat-Flash-Qwen2-7B (Hong & Do, 2026)	62.9	60.4	65.4	74.2	72.2	76.1	79.5	75.9	83.1	–	–	–
RCoT-Seg-Qwen2.5-VL-7B (Wen et al., 2026)	54.8	51.9	57.7	73.3	71.5	75.1	78.6	75.0	82.2	–	–	–
FaSTra-InternVL3-8B	64.0	60.7	67.3	72.1	70.2	74.0	81.0	77.3	84.6	71.6	70.7	72.5
FaSTra-Qwen3-VL-4B	**68.7**	**65.4**	**72.0**	**74.6**	**72.5**	**76.6**	**82.0**	**78.5**	**85.6**	**72.2**	**71.1**	**73.3**

ReferDINO (Liang et al., 2025) studies referring video object segmentation with visual grounding foundation models. It exploits vision-language grounding capability to identify referred objects and maintain their segmentation across video frames.

MPG-SAM 2 (Rong et al., 2025) adapts SAM2 to referring video object segmentation by incorporating mask priors and global context. These additional cues provide stronger target guidance for video mask propagation.

Table 11: Additional comparisons on reasoning video object segmentation. We include recent SFT-based, training-free, and reinforcement-learning-based methods on ReVOS (Yan et al., 2024).

Method	Referring			Reasoning			Overall		
	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
SFT-based Methods									
Sa2VA-i-InternVL2.5-1B (Nekrasov et al., 2025)	58.6	56.4	60.7	48.3	45.9	50.7	53.4	51.2	55.7
HyperSeg-Mipha-3B (Wei et al., 2025a)	58.5	56.0	60.9	53.0	50.2	55.8	55.7	53.1	58.4
InstructSeg-Mipha-3B (Wei et al., 2025b)	57.0	54.8	59.2	51.9	49.2	54.7	54.5	52.0	56.9
Sa2VA-InternVL2.5-4B (Yuan et al., 2026)	63.0	60.5	65.5	56.2	53.2	59.1	59.6	56.8	62.3
Sa2VA-i-InternVL2.5-4B (Nekrasov et al., 2025)	66.5	64.3	68.8	60.9	58.2	63.6	63.7	61.2	66.1
Sa2VA-i-FT-InternVL2.5-4B (Nekrasov et al., 2025)	67.1	64.8	69.4	61.5	58.8	64.2	64.3	61.8	66.8
UniPixel-Qwen2.5-VL-7B (Liu et al., 2025)	65.8	63.9	67.8	61.5	59.4	63.7	63.7	61.7	65.7
VISA-Chat-UniVi-7B (Yan et al., 2024)	52.9	51.1	54.7	39.2	36.7	41.7	46.1	43.9	48.2
VRS-HQ-Chat-UniVi-7B (Gong et al., 2025)	62.1	59.8	64.5	56.1	53.5	58.7	59.1	56.6	61.6
Sa2VA-i-InternVL2.5-8B (Nekrasov et al., 2025)	68.7	66.3	71.1	63.1	60.3	65.8	65.9	63.3	68.5
TrackGPT (IT)-LLaVA-13B (Zhu et al., 2023)	49.5	48.3	50.6	40.5	38.1	42.9	45.0	43.2	46.8
VISA (IT)-Chat-UniVi-13B (Yan et al., 2024)	57.4	55.6	59.1	44.3	42.0	46.7	50.9	48.8	52.9
VISA-Chat-UniVi-13B (Yan et al., 2024)	54.1	52.3	55.8	40.9	38.3	43.5	47.5	45.3	49.7
VRS-HQ-Chat-UniVi-13B (Gong et al., 2025)	63.3	61.1	65.5	56.8	54.1	59.4	60.0	57.6	62.5
VIRST-VideoChat-Flash-Qwen2-7B (Hong & Do, 2026)	70.8	68.8	72.8	66.1	63.9	68.3	68.4	66.3	70.6
CiCVS-3.8B (Zhuge et al., 2026)	64.7	62.4	67.0	58.2	55.7	60.7	61.4	59.1	63.8
RViSeg-3.8B (Shao et al., 2026)	52.6	51.0	54.2	47.4	45.8	49.0	50.3	49.3	51.2
RViSeg-7B (Shao et al., 2026)	52.2	50.8	53.5	48.2	46.5	49.9	50.6	49.5	51.7
RGA3-7B (Wang et al., 2025)	60.5	58.7	62.3	55.4	53.1	57.7	58.0	55.9	60.0
ViLLa-6B (Zheng et al., 2025)	–	–	–	–	–	–	57.0	54.9	59.1
ReVSeg-7B (Li et al., 2025)	65.7	63.3	68.1	58.6	55.4	61.8	62.1	59.3	65.0
Training-free / Zero-shot Methods									
SDAM-Qwen2.5-VL (Zhu et al., 2026)	61.5	59.0	63.9	54.6	51.8	57.5	58.0	55.4	60.7
CoT-RVS-GPT-4o (Kao et al., 2026)	–	–	–	–	–	–	55.9	52.8	59.0
Refer-Agent-Ovis2.5-9B (Jiang et al., 2026b)	65.0	62.3	67.7	57.7	54.3	61.0	61.3	58.3	64.3
RL-based Methods									
Veason-R1-Qwen2.5-VL-7B (Gong et al., 2026)	63.6	60.7	66.5	59.0	55.8	62.2	61.3	58.2	64.4
VideoSeg-R1-Qwen2.5-VL-7B (Xu et al., 2026)	–	–	–	–	–	–	61.1	58.2	64.0
VideoSEG-O3-Qwen3-VL-2B (Dai et al., 2026)	67.5	65.0	70.0	62.0	59.0	65.0	64.8	62.0	67.5
VideoSEG-O3-Qwen3-VL-4B (Dai et al., 2026)	70.3	67.7	72.9	65.1	62.4	67.9	67.7	65.0	70.4
VideoSeg-R1-Qwen2.5-VL-3B (Xu et al., 2026)	–	–	–	–	–	–	58.6	56.4	60.8
RCoT-Seg-Qwen2.5-VL-3B (Wen et al., 2026)	64.3	61.6	66.9	58.2	54.9	61.6	61.2	58.2	64.3
RCoT-Seg-Qwen2.5-VL-7B (Wen et al., 2026)	65.9	63.2	68.7	59.7	56.6	62.8	62.8	59.9	65.7
FaSTra-InternVL3-2B	63.1	60.7	65.4	55.6	52.7	58.6	59.3	56.7	62.0
FaSTra-Qwen3-VL-2B	67.6	65.5	69.6	62.1	58.6	65.6	64.8	62.1	67.6
FaSTra-InternVL3-8B	66.6	64.3	68.8	61.2	58.5	63.9	63.9	61.4	66.4
FaSTra-Qwen3-VL-4B	**72.7**	**70.3**	**75.1**	**67.0**	**64.3**	**69.7**	**69.8**	**67.3**	**72.4**

FlowRVS (Jiang et al., 2026a) formulates referring video segmentation from a flow-matching perspective. Instead of relying solely on the conventional target-identification-and-propagation pipeline, it models the transformation from video representations to target masks under language guidance.

DeRVOS (Cheng et al., 2026) decouples consistent trajectory generation from multimodal understanding for referring video object segmentation. This formulation separates temporal trajectory modeling from language-based target identification.

SDAM (Zhu et al., 2026) presents a training-free spatio-temporally decoupled framework for reasoning video segmentation. It further employs adaptive object memory to maintain target information across video frames without task-specific training.

Refer-Agent (Jiang et al., 2026b) formulates referring video object segmentation as a collaborative multi-agent process. Reasoning and reflection are used to progressively improve target identification and segmentation without additional task-specific training.

CoT-RVS (Kao et al., 2026) studies zero-shot chain-of-thought reasoning for video segmentation. In addition to the standard setting, the main comparison also includes CoT-RVS-online, which performs reasoning in an online video-processing setting.

Sa2VA (Yuan et al., 2026) couples a multimodal large language model with SAM2 for dense grounded understanding of images and videos. The MLLM performs language-conditioned target reasoning, while SAM2 predicts and propagates the corresponding masks across video frames.

Sa2VA-i (Nekrasov et al., 2025) revisits Sa2VA by improving the consistency between its training and inference procedures. In addition to standard Sa2VA-i variants, our ReVOS comparison includes the fine-tuned Sa2VA-i-FT variant.

InstructSeg (Wei et al., 2025b) formulates multiple visual segmentation tasks within a unified instructed segmentation framework based on multimodal large language models. It supports both image and video segmentation from language instructions.

VideoSEG-O3 (Dai et al., 2026) introduces a multi-turn reinforcement learning framework for reasoning video object segmentation. It performs iterative temporal-spatial exploration with segmentation feedback rather than relying on a single fixed reasoning step.

VideoSeg-R1 (Xu et al., 2026) applies reinforcement learning to reasoning video object segmentation. It combines video reasoning with target segmentation and optimizes the reasoning process with reinforcement-learning objectives.

RCoT-Seg (Wen et al., 2026) introduces reinforced chain-of-thought reasoning for video reasoning and segmentation. It explicitly models keyframe reliability, allowing the model to adapt its frame selection when the current visual evidence is unreliable.

HyperSeg (Wei et al., 2025a) develops a hybrid segmentation assistant with a fine-grained visual perceiver. It aims to unify multimodal reasoning with detailed pixel-level perception for visual segmentation tasks.

VideoLISA (Bai et al., 2024) extends language-instructed reasoning segmentation from images to videos. It connects language reasoning with video-level segmentation so that a language-conditioned target can be segmented consistently across temporal frames.

VideoGLaMM (Munasinghe et al., 2025) extends multimodal large language models to pixel-level visual grounding in videos. It enables language-based interaction with video content while associating textual responses with spatially grounded target regions.

CiCVS (Zhuge et al., 2026) improves reasoning video object segmentation through context-infused trajectories. It focuses on strengthening video context modeling and frame consistency during target segmentation.

RViSeg (Shao et al., 2026) develops a reasoning-centric video object segmentation framework based on multimodal large language models. It emphasizes joint semantic reasoning and temporally consistent target segmentation in videos.

GLUS (Lin et al., 2025) unifies global and local reasoning within a single large language model for video segmentation. Following the setting in the main comparison table, we report its Additional-SFT variant, denoted as GLUS^{A}.

LISA (Lai et al., 2024) introduces reasoning segmentation with a large language model. It connects the hidden representation of a dedicated segmentation token to a mask decoder, providing an interface between language-level reasoning and pixel-level mask prediction.

UniPixel (Liu et al., 2025) unifies object referring and segmentation for pixel-level visual reasoning. It integrates language-based object understanding with fine-grained pixel-level prediction in a unified multimodal framework.

VISA (Yan et al., 2024) introduces reasoning video object segmentation based on large language models. It requires the model to identify targets from implicit language instructions and generate

temporally consistent video masks. The ReVOS comparison includes both standard VISA models and the instruction-tuned VISA (IT) variant.

VRS-HQ (Gong et al., 2025) focuses on high-quality video reasoning segmentation by emphasizing temporal information during target reasoning and mask prediction. It is designed to improve segmentation quality when target identification depends on temporal video evidence.

ViLLa (Zheng et al., 2025) addresses video reasoning segmentation with a large language model. It combines language-based reasoning over video content with pixel-level target segmentation.

ReVSeg (Li et al., 2025) incentivizes the reasoning chain for video segmentation with reinforcement learning. It decomposes the reasoning process into semantic interpretation, temporal evidence selection, and spatial grounding before producing the final segmentation.

VIRST (Hong & Do, 2026) introduces a video-instructed reasoning assistant for spatio-temporal segmentation. It jointly models video-level reasoning and spatial-temporal target segmentation within a unified framework.

TrackGPT (Zhu et al., 2023) studies tracking from human-intent instructions. Instead of receiving an explicit target specification, the model first reasons about the intended object from language and subsequently tracks it through the video.

RGA3 (Wang et al., 2025) studies object-centric video question answering with visual grounding and referring. Its object-level grounding capability provides an additional comparison for reasoning-oriented target identification on ReVOS.

Veason-R1 (Gong et al., 2026) applies reinforcement learning to video object segmentation and explicitly encourages the model to reason before segmentation. Its optimization targets both reasoning-related decisions and the resulting segmentation quality.

E.3 Additional Comparisons of Model Architectures and Inference Efficiency

Table 12 summarizes the deployed architecture at the two backbone scales. The vision encoders of Qwen3-VL-2B and Qwen3-VL-4B are nearly identical in structure and parameter count, whereas the language model grows from 2.03B to 4.41B parameters. Thus, scaling the backbone primarily increases the reasoning capacity behind the mask token rather than the visual encoding capacity. Both variants further inject visual features into multiple language-model layers and explicitly encode temporal and spatial positions, allowing the final mask representation to preserve multi-level visual and temporal information.

A stronger asymmetry appears between frame segmentation and temporal propagation. As shown in Table 6 in the main text, SAM2 requires 2442.04 GFLOPs per frame, compared with only 142.46 GFLOPs for EdgeTAM, corresponding to a 17.14× reduction in arithmetic. Most of this gap comes from the image encoder and memory-attention modules, while the mask decoder and memory encoder contribute little to the difference. This supports our design choice of using the expensive segmentor only for candidate-frame initialization and relying on the lightweight propagator for dense temporal processing.

Table 13 further shows that this asymmetry translates into lower end-to-end inference cost at both backbone scales. The per-frame latency decreases from 43.36 to 33.95 ms with Qwen3-VL-2B and from 43.35 to 32.76 ms with Qwen3-VL-4B. Although `FaSTra` introduces a modest fixed overhead and higher peak memory because multiple trajectories are propagated concurrently, the lower per-frame cost dominates for typical video lengths. For example, at 4B, inference time decreases from 5.13 to 4.14 s for 100 frames and from 13.80 to 10.69 s for 300 frames.

The break-even analysis leads to the same conclusion. With the deployed candidate-frame setting, the estimated break-even occurs at roughly 7 frames for the 4B model and 12 frames for the 2B model, both shorter than the videos in the benchmarks considered here. Therefore, the additional initialization cost is amortized over practical RVOS sequences, while the lightweight propagation path provides increasing savings as video length grows.

Finally, the 17.14× reduction in propagation GFLOPs translates to only about a 1.32× reduction in measured per-frame latency, because frame decoding, resizing, and mask post-processing already account for a large fraction of the remaining runtime. This indicates that further shrinking the

Table 12: Configuration of every module the deployed **FaSTra** system loads, at the two backbone sizes we evaluate.

Module	Property	Sa2VA-Qwen3-VL-2B	Sa2VA-Qwen3-VL-4B
Vision encoder	Layers / width / feed-forward	24 / 1024 / 4096	24 / 1024 / 4096
	Attention heads	16	16
	Patch / spatial merge / temporal patch	16 / 2 / 2	16 / 2 / 2
	DeepStack injection layers	{5, 11, 17}	{5, 11, 17}
	Visual tokens per frame	392–1,568	392–1,568
	Output width	2048	2560
	Parameters	406.96 M	415.35 M
Language model	Layers / width / feed-forward	28 / 2048 / 6144	36 / 2560 / 9728
	Query heads / key–value heads	16 / 8	32 / 8
	Head dimension	128	128
	Positional encoding	interleaved M-RoPE, sections $(24, 20, 20)$, base 5×10^6	
	Normalization / activation	RMSNorm ($\epsilon = 10^{-6}$) / SwiGLU	
	Vocabulary / context length	151,674 / 262,144	
	Output embedding	untied, 310.63 M	untied, 388.29 M
	Parameters (with output embedding)	2,030.67 M	4,410.08 M
Mask-token projection	Parameters	4.72 M	7.21 M
Frame segmentor (Ravi et al., 2025)	Backbone	Hiera-L, 1024^2 input, 7 memory slots	
	Parameters	224.43 M	224.43 M
Propagator (Zhou et al., 2025)	Backbone	RepViT-M1 with spatial perceiver, 1024^2 input, 7 memory slots	
	Parameters	13.94 M	13.94 M
TDF selector	Parameters	0.151 M	0.151 M
CT router	Parameters	0.134 M	0.134 M
Total	Parameters loaded at inference	2.68 B	5.07 B

Table 13: Inference cost of **FaSTra** compared with fixed-keyframe propagation at two backbone scales. *Fixed* denotes the per-expression cost, while *Marginal* denotes the per-frame cost. Memory reports peak allocation at the shortest and longest lengths, and the last two columns show the fitted total time for 100 and 300 frames. **FaSTra** uses $N = 9$ candidate frames in this measurement.

Backbone	Pipeline	Fixed (ms/expr.)	Marginal (ms/frame)	Mem., 30 fr. (GB)	Mem., peak (GB)	Total time (s) 100 fr.	Total time (s) 300 fr.
Sa2VA-Qwen3-VL-2B	Fixed keyframe	569	43.36	8.8	12.8	4.90	13.58
	FaSTra	682	**33.95**	8.9	16.3	**4.08**	**10.87**
Sa2VA-Qwen3-VL-4B	Fixed keyframe	791	43.35	14.2	18.1	5.13	13.80
	FaSTra	867	**32.76**	14.4	21.5	**4.14**	**10.69**

propagator would provide limited practical benefit. Instead, the main efficiency gain comes from allocating the heavyweight segmentation model only to informative frames and using lightweight modules for the majority of temporal processing.

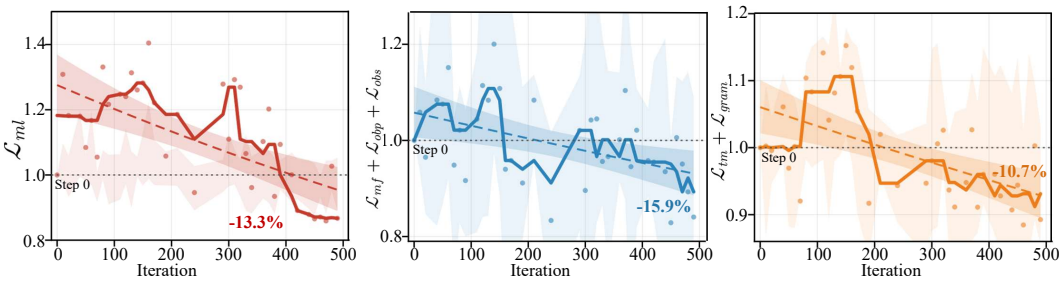

Figure 5: Additional analysis of distillation loss convergence.

E.4 Additional Analysis of Distillation Loss Convergence

Fig. 5 analyzes the optimization behavior of the three level propagation distillation. From left to right, we report the prediction-level logit loss $\mathcal{L}\mathrm{ml}$, the state-level feature loss $\mathcal{L}\mathrm{mf} + \mathcal{L}\mathrm{obp} + \mathcal{L}\mathrm{obs}$, and the temporal-level relation loss $\mathcal{L}\mathrm{tm} + \mathcal{L}\mathrm{gram}$. Despite the stochastic fluctuations during training, all three objectives exhibit clear downward trends, decreasing by 13.3%, 15.9%, and 10.7%, respectively. This consistent convergence indicates that the student progressively aligns with the BoT teacher at the prediction, internal-state, and temporal-relation levels, supporting the effectiveness of our tri-level propagation distillation.

E.5 Additional Analysis of FI-GRPO Reward Dynamics

Fig. 6 shows the evolution of the FI-GRPO reward during training. The overall complementarity reward $\mathcal{R}^{\mathrm{overall}}$ increases steadily from near zero to approximately 0.5, indicating that the learned policy progressively favors more beneficial trajectory assignments. Among its components, the relative-quality reward $\mathcal{R}^{\mathrm{rel}}$ shows the clearest upward trend, while the degradation penalty $-\mathcal{R}^{\mathrm{down}}$ becomes less negative over training, suggesting fewer harmful routing decisions relative to the best single-trajectory baseline. The gain reward $\mathcal{R}^{\mathrm{gain}}$ and teacher-alignment reward $\mathcal{R}^{\mathrm{tea}}$ remain positive and gradually stabilize. Despite local fluctuations caused by trajectory sampling, the consistent increase in total reward demonstrates that FI-GRPO effectively improves frame-wise routing decisions and better exploits complementary trajectories.

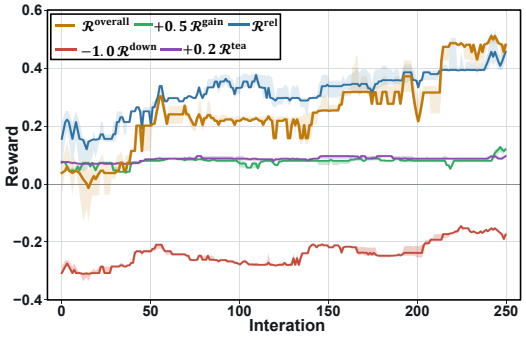

Figure 6: Evolution of the routing reward and its weighted components during FI-GRPO training.

F Additional Qualitative Experimental Results

F.1 Prediction Results

We further conduct qualitative experiments on representative video sequences to examine how **FaSTra** behaves under challenging temporal conditions, including occlusion, pose variation, close interactions, and long-term identity changes. Figs. 7-9 visualize the selected keyframes, candidate trajectories, predicted quality matrices, and final segmentation results, providing a more detailed view of how temporal complementarity and dynamic routing contribute to robust video segmentation.

Temporal complementarity and dynamic routing. Across Figs. 7-9, the most reliable trajectory clearly changes over time rather than remaining fixed. For example, in the crutch sequence, the yellow trajectory is more reliable in the earlier frames, while the blue and green trajectories become preferable after the target moves farther across the scene. A similar pattern appears in the chasing



Figure 7: Additional qualitative results of **FaSTra** on representative video sequences. The visualization shows the selected keyframes, candidate trajectories, the predicted quality matrix, and the final segmentation results compared with other methods (Yuan et al., 2026; Dai et al., 2026).

and dancing sequences, where different trajectories dominate at different temporal regions as the target undergoes large motion, scale variation, or pose changes. The predicted quality matrices reflect these shifts, and the final routing path correspondingly switches between trajectories instead



Figure 8: Additional qualitative results of `FaSTra` on representative video sequences.

of committing to a single source. These examples visually support our central observation that trajectories initialized from different keyframes provide complementary temporal coverage.

Robust semantic disambiguation in complex scenes. The examples also show that `FaSTra` is more reliable when the language query requires distinguishing among visually similar or interacting



Figure 9: Additional qualitative results of **FaSTra** on representative video sequences.

instances. In Fig. 8, the queries "the fighter who is instructing the other one" and "the dancer who keeps their hands up near their face" require relational and pose-based reasoning rather than simple appearance matching. Likewise, the query "the animal that is being chased" in Fig. 7 and "the only person dancing while the rest of the room watches" in Fig. 9 require maintaining the intended role across time. In these cases, Sa2VA and VideoSEG-O3 frequently produce incomplete masks, lose the target, or confuse it with nearby instances, whereas **FaSTra** preserves the queried identity more consistently throughout the sequence.

Long-term identity preservation under appearance changes. The qualitative results further demonstrate the advantage of dynamic routing under prolonged temporal variation. In the indoor crutch sequence, the target changes position and scale substantially; in the forest interaction sequence, the two fighters overlap and exchange relative positions; and in the crowded dancing scenes, the target is surrounded by multiple similar people while undergoing strong pose changes. Under these conditions, predictions from a single trajectory can become unreliable after occlusion or appearance variation, whereas **FaSTra** switches to trajectories initialized from more favorable temporal locations. Consequently, the final masks remain more complete and temporally stable across difficult frames, indicating that the proposed routing mechanism improves not only per-frame localization but also long-term identity preservation.

F.2 Failure Cases and Limitation

Despite the improved temporal consistency, Figures 7–9 reveal several remaining limitations. Boundary quality can degrade around thin structures and articulated body parts, such as limbs, hands, and the crutch, especially under large motion or partial occlusion. In crowded or highly interactive scenes, such as the fighter and dancing sequences, overlapping instances can also introduce local mask ambiguity even when the target identity is preserved correctly. Moreover, the router can only select among the available candidate trajectories; when all trajectories are simultaneously degraded by severe occlusion, motion blur, or inaccurate keyframe masks, routing alone cannot fully recover the target. Future work will therefore focus on finer boundary-aware propagation and more adaptive trajectory generation to improve robustness in these challenging cases.