Reinforcing Video Reasoning Segmentation to Think Before It Segments

Supplementary Material

In this supplementary material, we provide additional experimental details and qualitative analyses for Veason-R1. To begin with, we elaborate on the construction of the chain-of-thought training data (§A). We then provide further implementation details to enhance reproducibility (§B). Next, we analyze the training curves to illustrate the model's optimization dynamics and convergence behavior (§C). This is followed by a discussion of representative failure cases to identify potential areas for improvement (§D). Finally, we include additional qualitative visualizations to further demonstrate the effectiveness of Veason-R1 (§E).

A. More Details of CoT Training Data Construction

Fig. 1 illustrates the full prompt that we use as input to Seed1.5-VL [4] for chain-of-thought generation. The prompt is presented to the model along with a sequence of uniformly sampled video frames and their corresponding timestamps, as well as the predefined object description and a selected keyframe. The prompt guides the model to (i) briefly summarize the overall video content, (ii) justify why the selected keyframe best matches the target expression, and (iii) describe the target's actions and location within the given frame. This structured prompting encourages the model to produce clear and interpretable reasoning based on the visual and contextual information. We observe that Seed1.5-VL tends to closely mimic the style of human-provided exemplars. To promote diversity in the generated CoT responses and avoid overly rigid reasoning patterns, we design three distinct CoT templates—each with slightly different phrasing and logical flow. For each sample, we randomly select one template during prompting, encouraging the model to explore different reasoning trajectories while maintaining consistency with the underlying task structure.

Our CoT corpus comprises 5.8K samples automatically generated from the ReVOS training set, using the first 10 referring expressions per video without manual curation. For each sample, a keyframe is randomly selected from the top-5 frames ranked by target foreground area and provided to Seed1.5-VL together with frame indices, bounding boxes, and the expression to produce the reasoning chain. Although the data are machine-generated, the structured prompt constraints described above encourage grounded and coherent outputs; we will further explore multi-candidate generation with joint model-human validation in future work. We will open-source the CoT data and the full generation pipeline to ensure reproducibility.

B. More Implementation Details

We adopt the same question template for both the SFT and GRPO stages, as illustrated in Fig. 2. During the GRPO stage, we set the maximum response length to 1024 tokens, the maximum gradient norm for clipping to 1.0, and the coefficient for the KL divergence regularization loss to 5e-3. All input frames in both training stages are resized to a resolution of 560 $\times$ 560 pixels. The MLLM's context window includes up to N=6 uniformly sampled frames. During training, we first select an anchor frame t from the top-5 frames with the largest target foreground area to ensure target visibility. The sampling interval is then set to $k = \lfloor T/(N-1) \rfloor$ with starting index $t_{\text{start}} = t \bmod k$, where T is the total number of video frames, providing balanced temporal coverage around the anchor. During inference, we prompt Veason-R1 to generate the keyframe timestamp and the bounding boxes of the referred objects. These predictions are then passed to SAM2 [6], which performs segmentation and mask propagation across the entire video sequence, yielding the final predicted mask sequence.

C. Training Curve Analysis

We visualize the training dynamics of Veason-R1 in Fig. 3. We monitor three key metrics during training: total reward score $\mathcal{R}_{total}$, response length, and the value of KL loss.

As shown in Fig. 3 (a), the model initialized via CoT-based supervision starts with a reward score above 3.0, indicating a reasonable ability to generate structured outputs and locate keyframes. The consistent upward trend in reward reflects effective policy refinement and better alignment with the designed reward function.

It can be observed in Fig. 3 (b) that the average response length initially decreases, suggesting that the model rapidly learns to eliminate redundant tokens and generate concise reasoning. In the later phase, the length slightly increases and then stabilizes, indicating an adjustment towards producing responses that balance brevity and informativeness.

As depicted in Fig. 3 (c), the KL divergence gradually increases as the model departs from the initial supervised policy to explore more reward-driven behaviors. It eventually stabilizes, indicating convergence to a consistent and reward-aligned policy under the regularization constraint.

D. Failure Case Analysis

Fig. 4 presents some failure cases of Veason-R1 on the ReVOS dataset. In the left example, the model generates a correct reasoning trace, accurately identifying the tiger

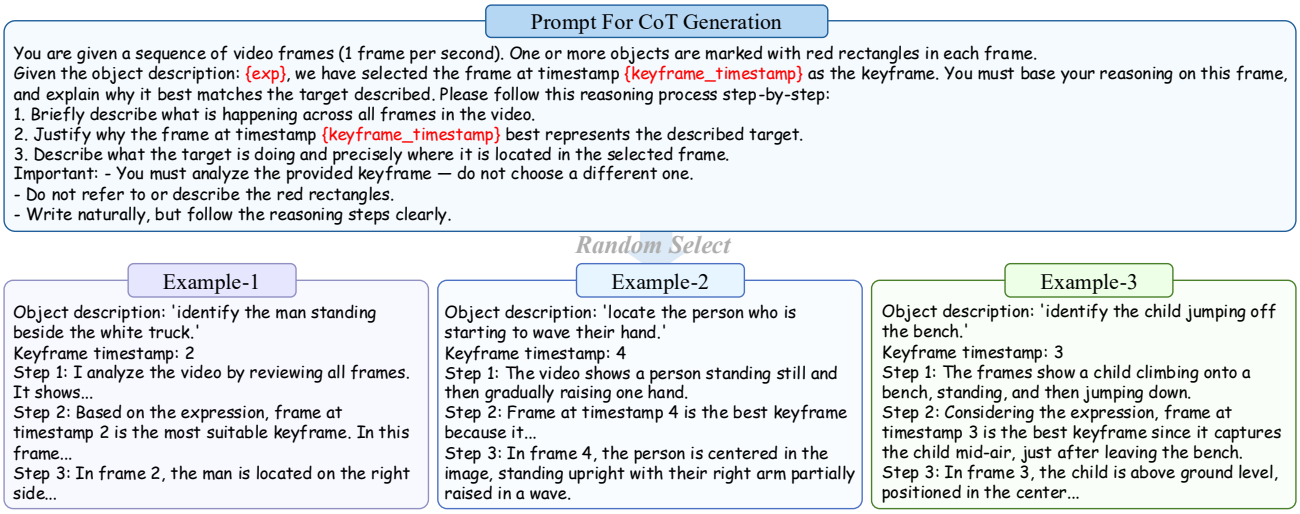


Figure 1. Prompt template for CoT data generation.

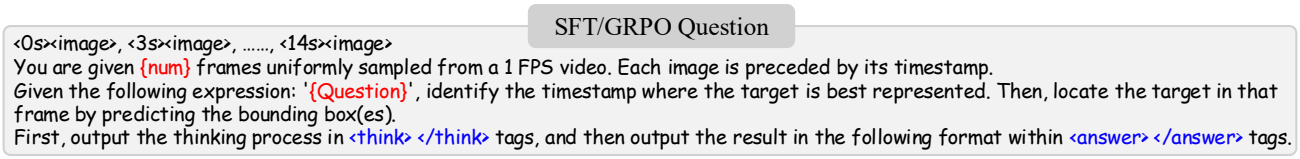


Figure 2. Question template of SFT and GRPO training stage.

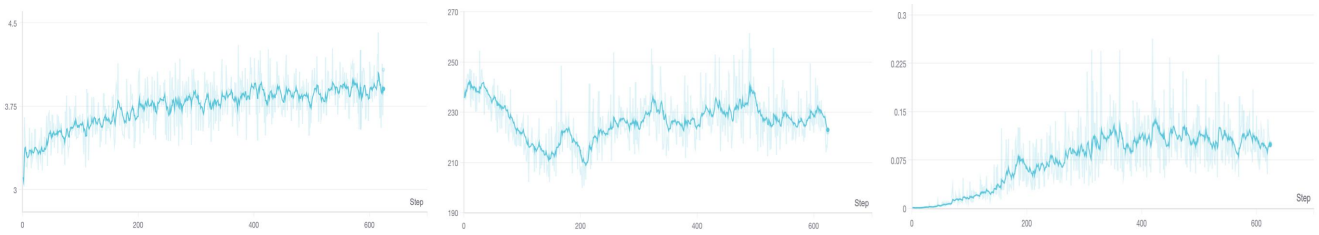
closest to the camera and describing its relation to the background tigers. However, the predicted masks fail to align with the described target, revealing a disconnect between reasoning and visual grounding. This suggests that Veason-R1 struggles to translate reasoning into precise spatial localization, especially in cluttered scenes with visually similar objects. In the right example, Veason-R1 correctly identifies the keyframe showing clear motion cues but ultimately selects the slower ship. The reasoning trace overlooks the constraint "fastest" and focuses solely on movement direction. This indicates the model's limited sensitivity to speed and motion-related concepts, revealing weaknesses in capturing temporal dynamics or comparing relative movements. These limitations suggest the importance of incorporating data that better emphasizes motion perception in future work.

E. More Visualization Results

We present additional visual comparisons of Veason-R1-3B with VISA-7B [7] and VideoLISA-3.8B [2] to underscore its robust reasoning capabilities and enhanced fine-grained grounding performance.

The qualitative comparison of Veason-R1 with VISA is shown in Fig. 5. In the left example, a herd of elephants partially occludes the vehicles crossing the road. Veason-R1 successfully identifies the timestamp where both vehicles are most clearly visible and segments them, whereas VISA mistakenly highlights background trees. This reflects Veason-R1's robustness to occlusions and its focus on task-relevant semantics. The right case indicates that Veason-R1 correctly segments the slippers in contact with the puppy's feet, demonstrating strong multi-object localization and spatial reasoning. VISA, by contrast, produces incorrect masks, revealing limitations in handling fine-grained interactions.

Fig. 6 visualizes the segmentation maps of Veason-R1 and VideoLISA on the ReasonVOS dataset. In the left example, Veason-R1 accurately segments the dancing bride in response to a poetic query involving abstract movement descriptions, whereas VideoLISA mistakenly segments the groom. This highlights Veason-R1's superior understanding of implicit and abstract language. In the right example, under a long-duration and high-complexity scenario, Veason-R1 successfully localizes the man riding the bicycle in response to a causal reasoning query, while VideoLISA yields fragmented and inconsistent predictions, demonstrating Veason-R1's stronger capacity for temporal reasoning and grounding in complex visual contexts.



(a) Average Reward Score (b) Average Response Length (c) KL Loss

Figure 3. GRPO training curves.

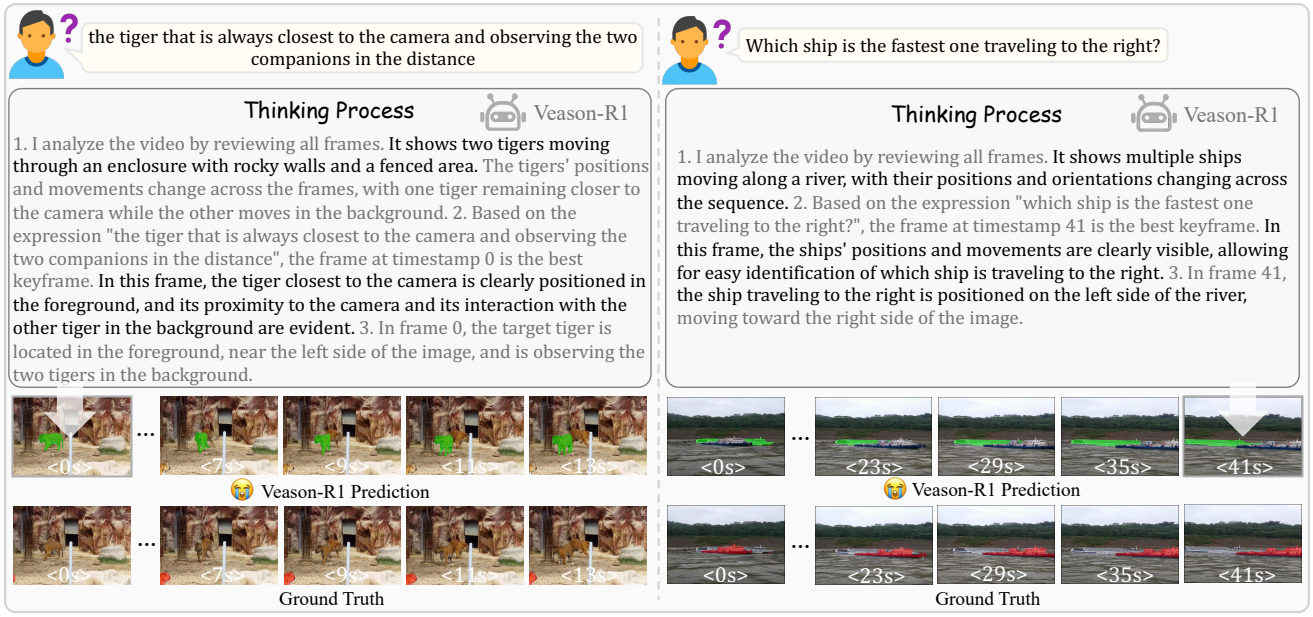


Figure 4. Visualization of failure cases on the ReVOS dataset. Veason-R1 exhibits inconsistencies between its reasoning traces and final segmentation outputs, and shows limited sensitivity to motion-related concepts such as object speed.

F. Multi-Instance Evaluation on ViCaS

To assess Veason-R1's multi-instance capability, we conduct a zero-shot evaluation on ViCaS [1], a benchmark designed for multi-object video referring segmentation. Veason-R1 is **not** trained on any ViCaS data. As shown in Tab. 1, Veason-R1 achieves 24.5 and 18.2 $\mathcal{J}\&\mathcal{F}$ on the validation and test sets, respectively, outperforming both VideoLISA and Video-LLaVA-Seg by a clear margin—despite the latter being explicitly fine-tuned on ViCaS. This result demonstrates that our keyframe-first reasoning paradigm, combined with structured CoT supervision and GRPO-based spatial alignment rewards, equips the model with a strong multi-object grounding ability that transfers effectively to unseen benchmarks without task-specific adaptation.s

Table 1. **Zero-shot multi-instance evaluation on ViCaS.** Veason-R1 is evaluated without any ViCaS training data. † denotes methods fine-tuned on ViCaS.

Methods	val	test
LMPM [3]	8.4	6.3
DsHmp [5]	13.3	10.4
VideoLISA [2]	10.7	7.8
Video-LLaVA-Seg† [1]	20.5	16.5
Veason-R1	**24.5**	**18.2**

References

[1] Ali Athar, Xueqing Deng, and Liang-Chieh Chen. Vicas: A dataset for combining holistic and pixel-level video understanding using captions with grounded segmentation. In *Proceedings of the Computer Vision and Pattern Recognition*

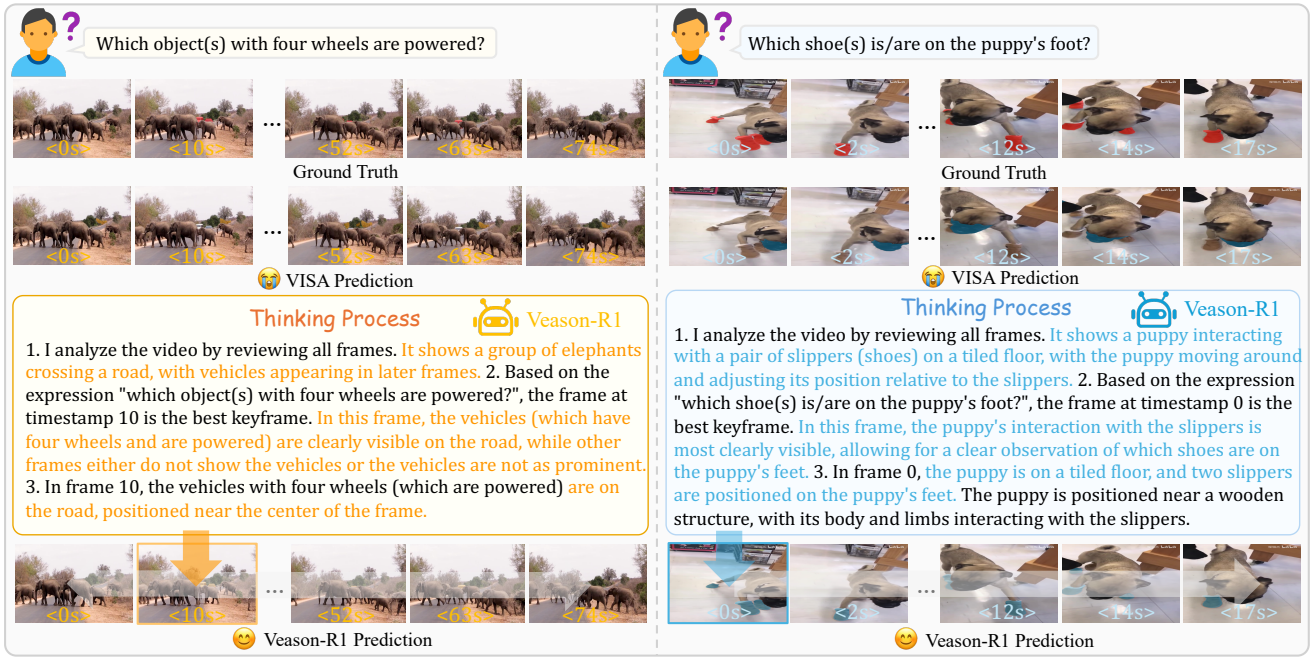


Figure 5. More visualization on the ReVOS dataset.

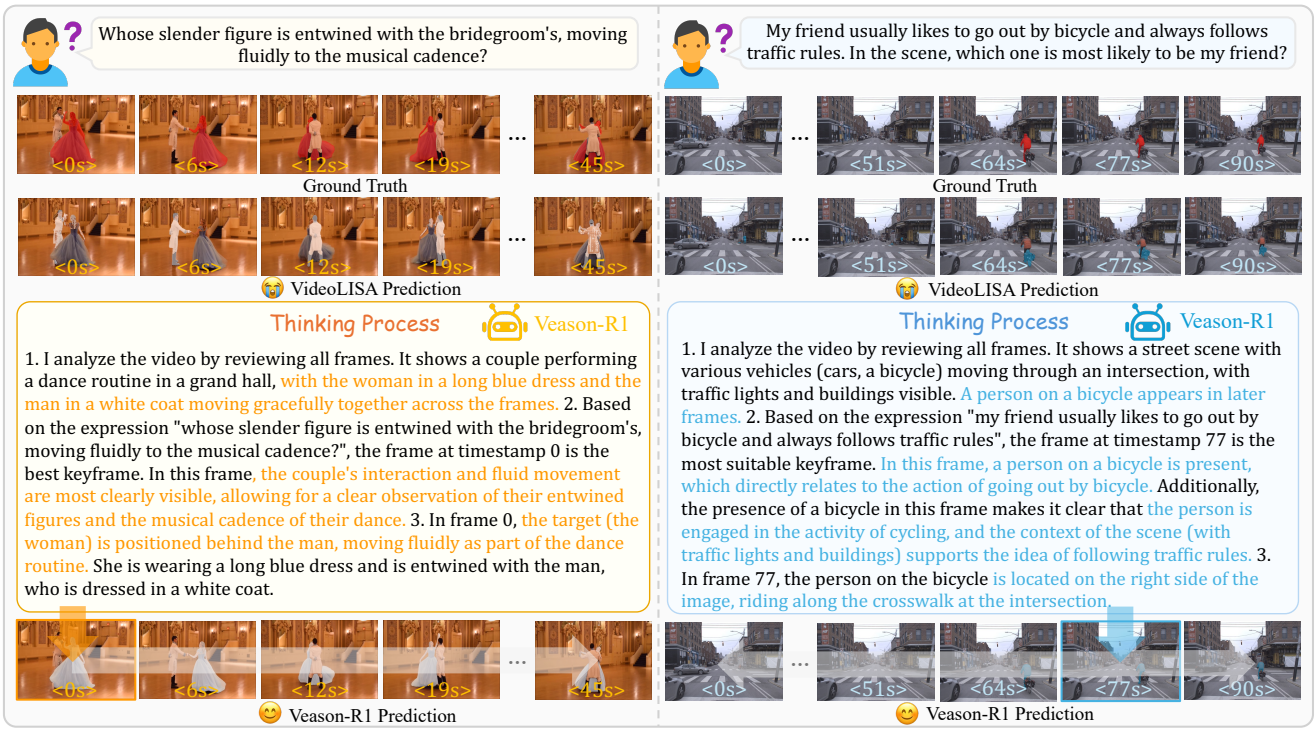


Figure 6. More visualization on the ReasonVOS dataset.

Conference, pages 19023–19035, 2025. 3

[2] Zechen Bai, Tong He, Haiyang Mei, Pichao Wang, Ziteng Gao, Joya Chen, Zheng Zhang, and Mike Zheng Shou. One token to seg them all: Language instructed reasoning segmentation in videos. *NeurIPS*, 37:6833–6859, 2024. 2, 3

[3] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Chen Change Loy. Mevis: A large-scale benchmark for video segmentation with motion expressions. In *ICCV*, pages 2694–

2703, 2023. 3

[4] Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, et al. Seed1. 5-vl technical report. *arXiv preprint arXiv:2505.07062*, 2025. 1

[5] Shuting He and Henghui Ding. Decoupling static and hierarchical motion perception for referring video segmentation. In *CVPR*, pages 13332–13341, 2024. 3

[6] 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. *arXiv preprint arXiv:2408.00714*, 2024. 1

[7] 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 *ECCV*, pages 98–115. Springer, 2024. 2