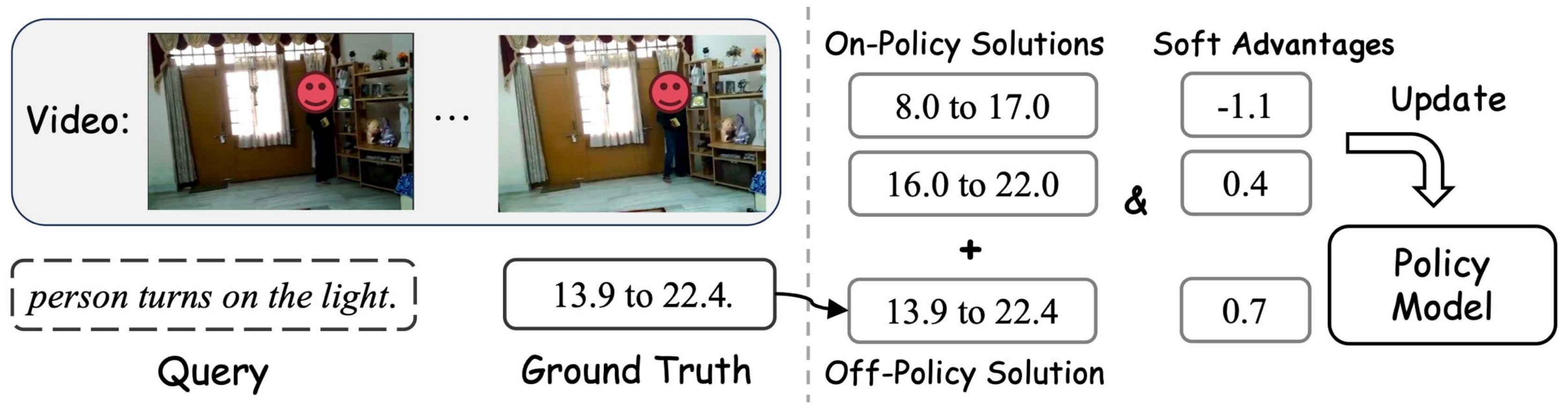
TempSamp-R1: Effective Temporal Sampling with Reinforcement Fine-Tuning for Video LLMs

Nankai University
ByteDance
Yunheng Li[1], Jing Cheng[2], Shaoyong Jia[2], Hangyi Kuang[1], Shaohui Jiao[2], Qibin Hou[1,3], Ming-Ming Cheng[1,3]
[1]VCIP, School of Computer Science, Nankai University, [2]ByteDance Inc., [3]NKIARI, Futian, Shenzhen

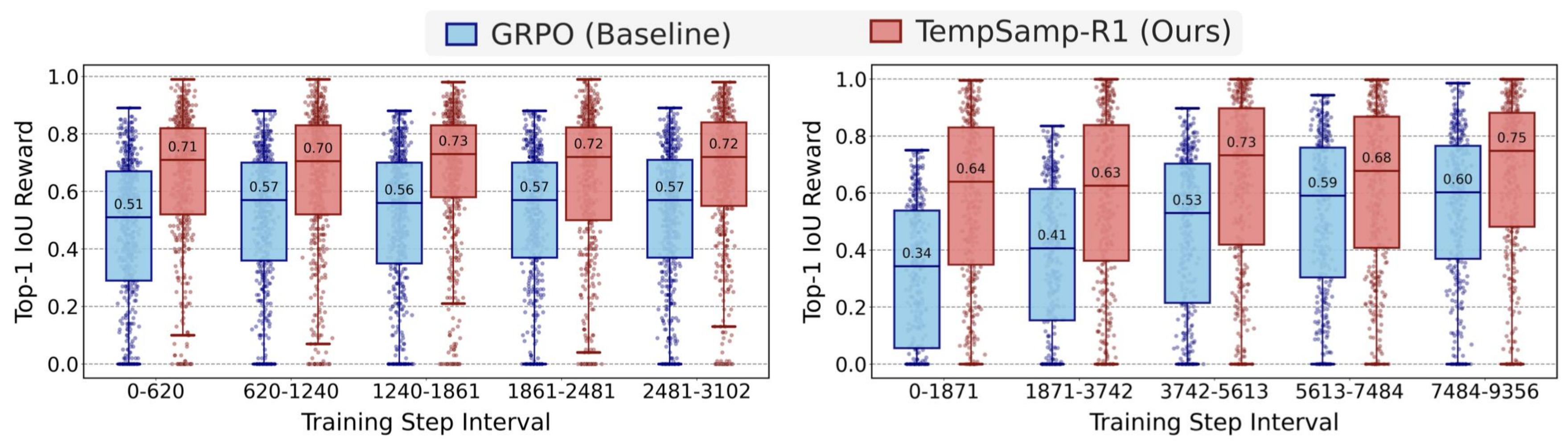
NEURAL INFORMATION PROCESSING SYSTEMS

Motivation

➢ GRPO methods use video high-quality annotations (e.g., timestamps) only for evaluation (e.g., IoU reward), not dynamic learning.



➢ The vast temporal search space severely hinders effective exploration.

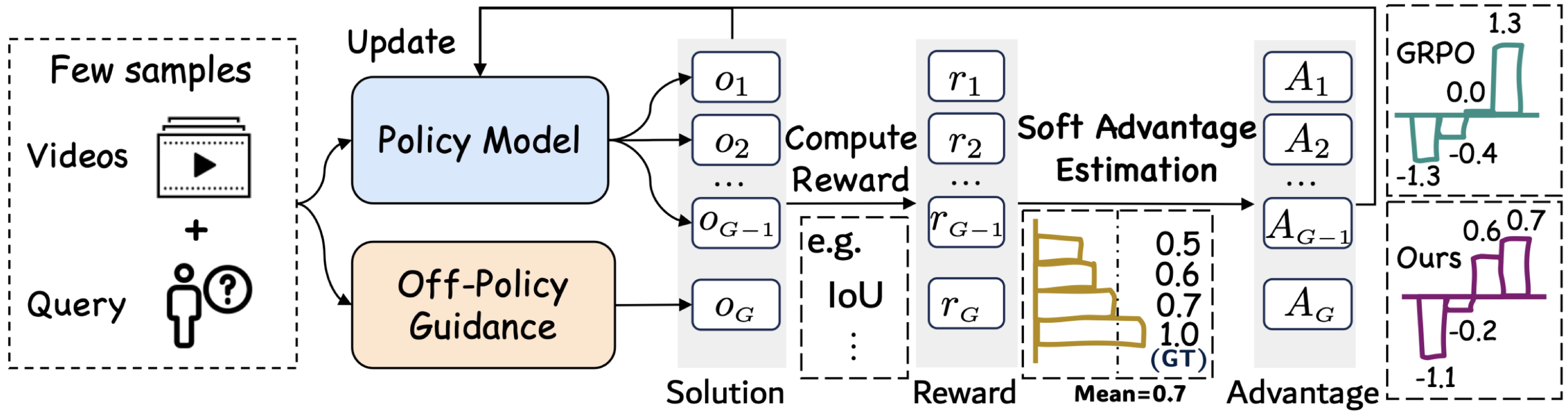


(a) Training on Charades-STA (b) Training on ActivityNet Caption

TempSamp-R1: Mix-policy sampling & Non-linear reward shaping

➢ To address the above limitation, we introduce a mixed-policy training strategy that incorporates external off-policy solutions to provide accurate and query-specific temporal grounding.

➢ To improve stability under skewed reward distributions, we apply a non-linear transformation to the rewards prior to advantage computation.



Experiments: Performance Comparison and Ablation Analysis

TempSamp-R1 (unified CoT/no-CoT) and its Mixed CoT (per-query better predictions) achieve strong performance.

		Charades-STA				ActivityNet Captions				QVHighlights	
Method	Type	mIoU	R1@0.3	R1@0.5	R1@0.7	mIoU	R1@0.3	R1@0.5	R1@0.7	mAP	HIT@1
Supervised Fine-Tuning (SFT) Methods											
UnLoc-L [62]	SFT	-	-	60.8	38.4	-	-	48.3	30.2	-	-
Timechat [44]	SFT	-	-	46.7	23.7	-	-	-	-	21.7	37.9
HawkEye [58]	SFT	49.3	72.5	58.3	28.8	39.1	55.9	34.7	17.9	-	-
TRACE [14]	SFT	-	-	61.7	41.4	-	-	37.7	24.0	-	-
VideoChat-T [67]	SFT	-	79.4	67.1	43.0	-	-	-	-	27.0	55.3
iMOVE [28]	SFT	57.9	79.8	68.5	45.3	49.3	67.2	50.7	32.4	-	-
Reinforcement Learning (RL) Methods based on Qwen2.5-VL-7B											
Qwen2.5-VL-7B [32]	-	29.0	-	24.2	11.1	21.1	-	15.8	7.5	-	-
VideoChat-R1 [32]	RL	60.8	-	71.7	50.2	-	-	-	-	-	-
VideoChat-R1-thinking [32]	RL	59.9	-	70.6	47.2	-	-	-	-	-	-
TimeZero [59]	RL	-	83.3	72.5	47.9	-	68.6	47.3	26.9	-	-
TempSamp-R1(no-CoT)	RL	61.7	83.3	73.6	52.2	52.1	72.8	55.4	34.2	**30.0**	**57.6**
TempSamp-R1 (CoT)	RL	**62.1**	**83.6**	**74.1**	**52.9**	**52.4**	**73.4**	**56.0**	**34.7**	28.3	54.9
TempSamp-R1 Mixed CoT	RL	64.2	85.0	76.0	56.3	54.9	75.7	58.7	37.6	29.3	63.7

Few-shot performance comparison of SFT, GRPO, and TempSamp-R1.

	50 videos		100 videos		200 videos		500 videos		
Method	R1@0.5	mIoU	R1@0.5	mIoU	R1@0.5	mIoU	R1@0.5	mIoU	Training Time
SFT	44.8	41.9	46.5	42.6	45.2	42.7	51.4	46.2	93 min
GRPO	36.2	38.4	39.3	40.8	43.5	43.8	55.3	49.8	338 min
TempSamp-R1 (Ours)	**46.7**	**44.7**	**54.0**	**49.1**	**58.2**	**51.8**	**64.0**	**55.1**	218 min

Ablation results on variants with mixed-policy rewards and alternative advantage shaping.

Method	R1@0.3	R1@0.5	R1@0.7
GRPO (baseline)	81.2	68.9	46.0
Mixed-policy	77.8	63.0	41.3
Reward downscaling	81.2	70.3	48.1
Advantage anchoring	81.8	70.7	49.1
Non-linear reward shaping	**82.9**	**72.1**	**49.6**

Skewness of the advantage distributions during training for different variants.