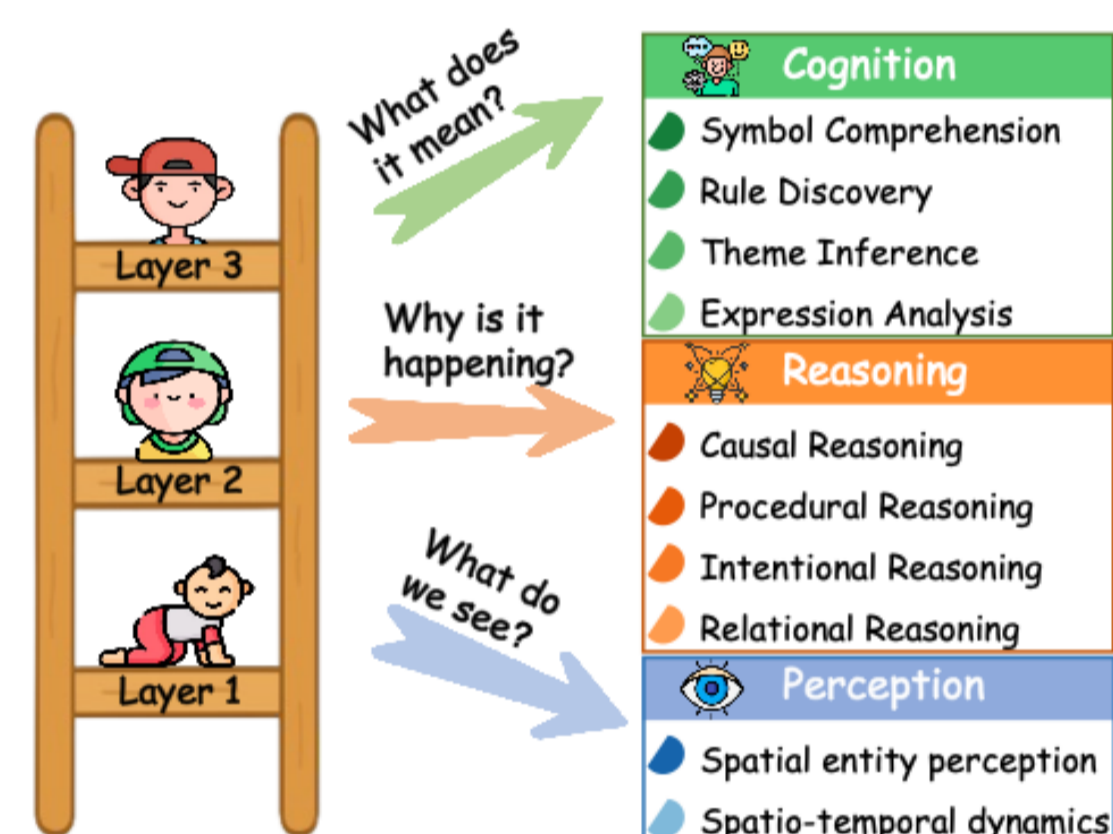
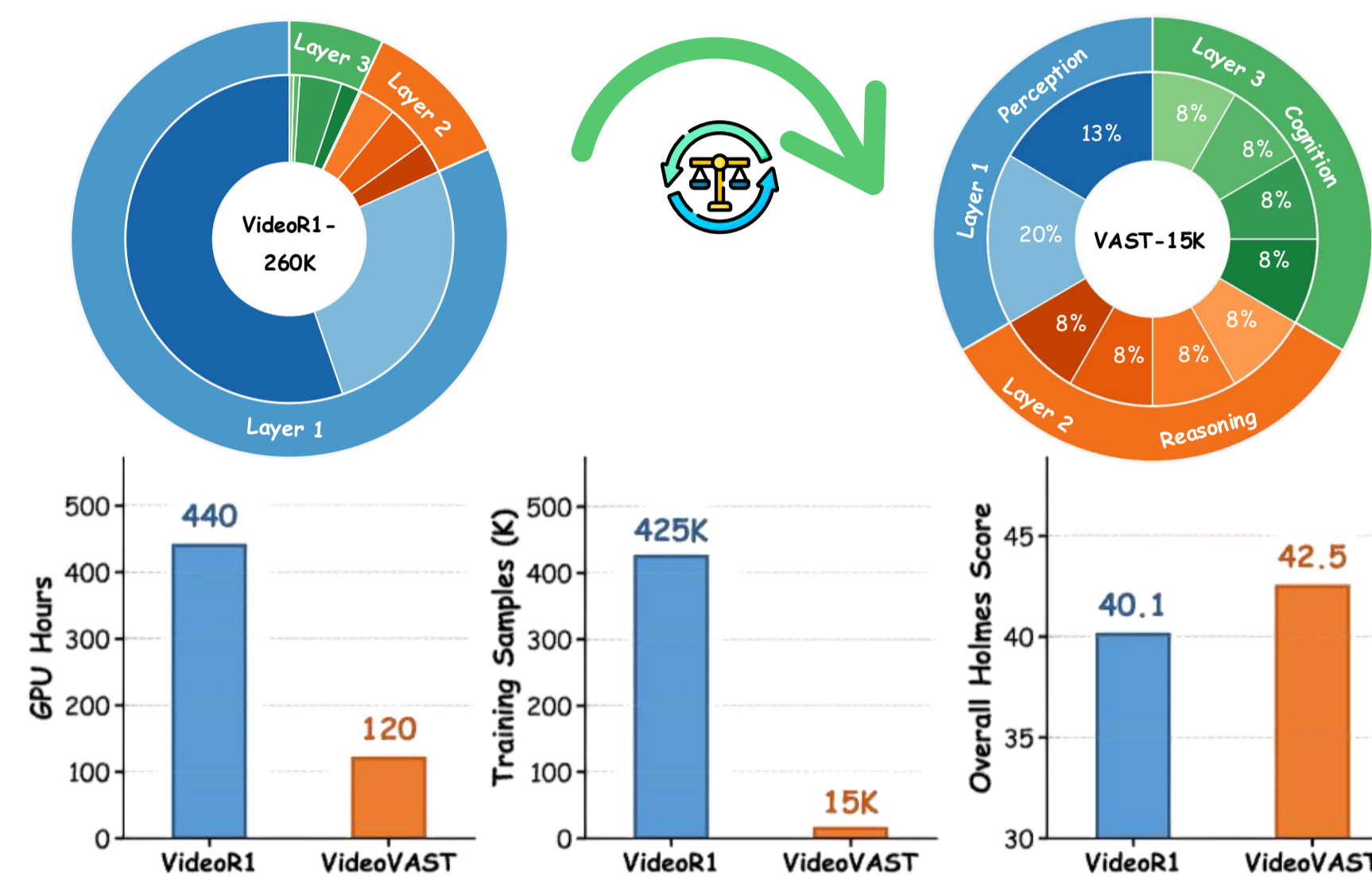


Motivation

- Decomposes video understanding into Cognition Process
- Provide ability-oriented data for RL training and evaluation
- Improves video reasoning with simple consistency rewards
- Avoids complex temporal modules and reward engineering
- Achieves stronger results with much lower training cost



Dataset Details



Comparison to Existing Approaches

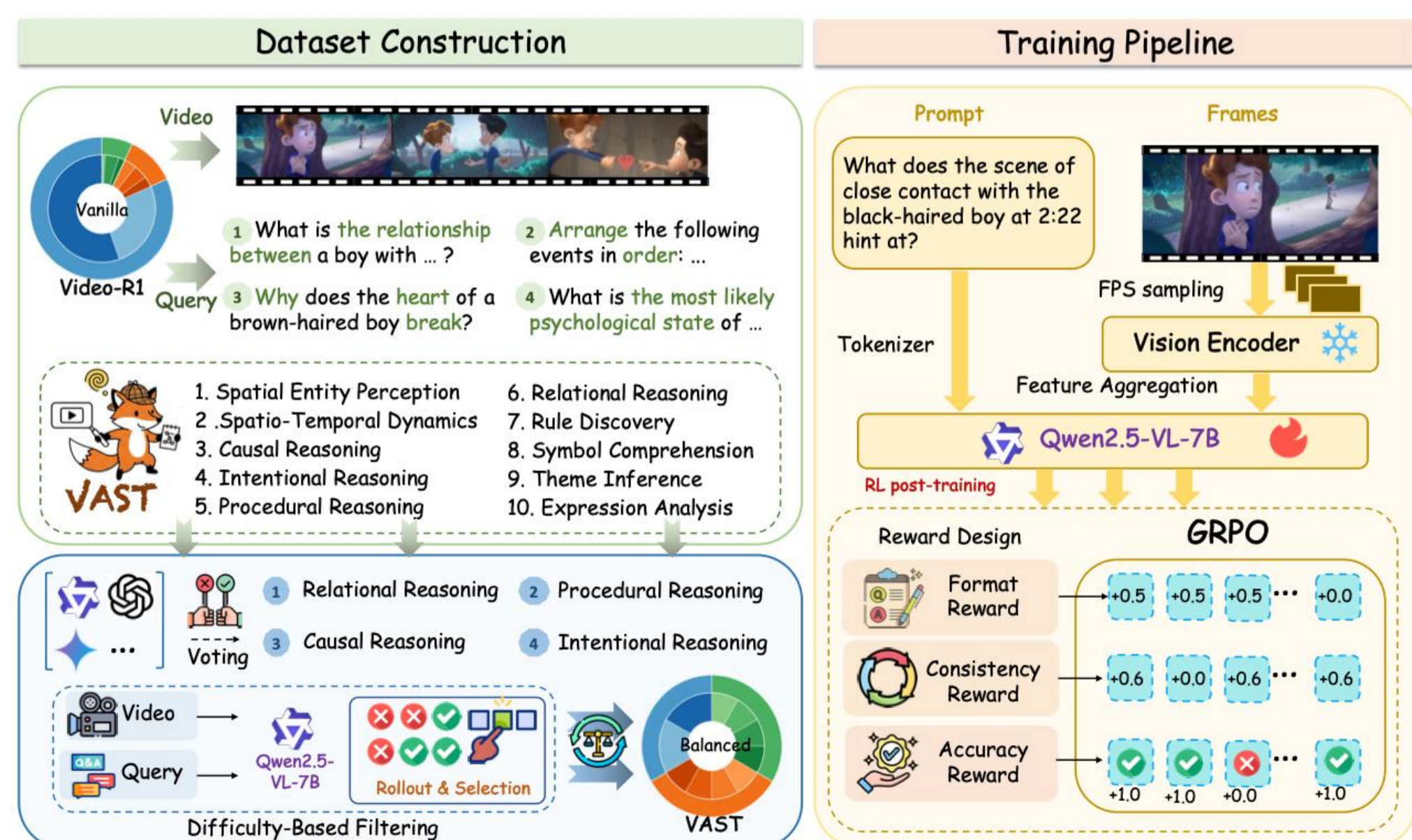
Table 1. Performance comparison on general-purpose video understanding benchmarks. The best results are highlighted in bold, and the second-best results are underlined.

Models	Pub.	RL Only	Frames	Video MMMU	MMVU (mc)	MV Bench	Temp Compass	VSI-Bench	VideoMME	Video Holmes
GPT-4o [20]	-	-	-	61.2	75.4	-	-	34.0	71.9	42.0
LLaMA-VID [31]	ECCV'24	X	-	-	-	41.9	45.6	-	-	-
VideoLLaMA2 [4]	arXiv'24.06	X	-	-	44.8	54.6	-	-	47.9	-
LongVA-7B [61]	TMLR'24	X	-	23.9	-	-	56.9	29.2	52.6	-
VILA-1.5-8B [32]	CVPR'24	X	-	20.8	-	-	58.8	28.9	-	-
VILA-1.5-40B [32]	CVPR'24	X	-	34.0	-	-	-	31.2	60.1	-
Video-25.02-7B [58]	arXiv'25.01	X	-	-	-	58.8	59.7	-	52.6	-
LLaVA-OneVision-7B [28]	TMLR'24	X	-	33.8	49.2	56.7	-	32.4	58.2	-
Kangaroo-8B [33]	arXiv'24.08	X	-	-	-	61.1	62.5	-	56.0	-
Qwen2.5-VL-7B (CoT) [1]	arXiv'25.02	X	16	47.8	59.2	57.4	72.2	27.7	53.1	27.8
Qwen2.5-VL-7B (CoT) [1]	arXiv'25.02	X	32	48.1	60.0	59.0	72.6	30.1	56.6	-
Qwen2.5-VL-7B (CoT) [1]	arXiv'25.02	X	64	<u>50.4</u>	60.0	59.2	72.9	31.4	<u>59.6</u>	-
Video-R1-7B [7]	NeurIPS'25	X	16	49.8	64.2	62.7	72.6	34.6	57.4	<u>40.1</u>
Video-R1-7B-zero [7]	NeurIPS'25	✓	16	49.5	63.8	60.4	70.9	31.8	53.8	-
Video-RFT-7B [49]	NeurIPS'25	X	16	48.2	<u>66.0</u>	<u>64.0</u>	<u>73.5</u>	34.6	55.9	36.4
VideoVAST-7B (Ours)	-	✓	16	50.6	66.1	66.3	74.2	33.5	57.7	42.5

More Details

Frames	SFT	CR	Video MMMU	MMVU (mc)	MV Bench	Temp Compass	VSI-Bench	VideoMME (wo sub)	Video Holmes	Overall
16	X	X	49.1	66.1	65.2	74.7	34.8	57.2	41.5	55.5
	✓	X	47.3	63.0	66.0	73.0	33.5	57.2	39.8	54.3
	X	✓	50.6	66.1	66.3	74.2	33.5	57.7	42.5	55.8
	✓	✓	49.9	63.2	65.5	73.0	36.5	58.0	40.4	55.2
32	X	X	48.7	65.4	66.6	73.2	37.4	60.6	38.6	55.8
	✓	X	49.9	62.4	66.9	73.5	35.0	59.9	37.7	55.0
	X	✓	51.2	67.0	67.4	74.9	34.3	59.4	39.7	56.3
	✓	✓	50.1	63.0	66.9	73.6	37.6	60.8	41.6	56.2
64	X	X	50.3	66.9	66.8	73.1	38.8	62.2	38.6	56.7
	✓	X	50.1	62.6	68.2	73.4	37.1	61.6	39.7	56.1
	X	✓	51.4	67.0	68.2	74.7	36.5	62.4	37.0	56.7
	✓	✓	50.6	62.7	68.0	73.8	39.5	62.1	39.2	56.6

Framework



- VAST-Eval: comprehensive evaluation across 7 video reasoning benchmarks
- VAST-Performance: consistent improvements over Video-R1 on key benchmarks
- VAST-Bench: stronger gains on reasoning and cognition-oriented abilities
- VAST-Efficiency: achieves strong performance with much lower training cost
- VAST-Ablation: validates ability stratification and consistency reward design

Visualization

(a) Question: What is the direct cause of the wound on the woman's neck?
Options: A. Steak knife accidentally cut. B. Car accident impact caused. C. self-harm behavior. D. The man killed with a knife. E. Supernatural power cutting. F. Red wine bottle fragments pierce in.

VideoVAST: <think>The video sequence shows a man holding a knife, which is a direct indication of the cause of the wound on the woman's neck. The presence of the knife and the man's actions suggest that the wound is a result of him using the knife. </think><answer>D</answer>

Video-R1: <think> Hmm, it seems like the most plausible option is A. Steak knife accidentally cut. The man is holding a knife, and the woman is lying on a table, which suggests that the wound might have been caused by an accidental cut from a steak knife. </think><answer>A</answer>

(b) Question: What does the focus on the white, orange and blue coat on the ground at 2:04 imply?
Options: A. The owner has been killed. B. Clothes are discarded. C. Shooting props bloopers. D. The weather suddenly turns cold. E. Clothes were stolen. F. The owner changes the style.

VideoVAST: <think>The focus on the white, orange and blue coat on the ground at 2:04 implies a significant event, given the context of the video which seems to be a dramatic or intense scene. The presence of the coat on the ground, along with the other elements of the scene, suggests a possible murder or death, as this is a common prop in horror or thriller genres to indicate a violent event. </think><answer>A</answer>

Video-R1: Option C suggests that it's a shooting props blooper. This is an interesting option. In film production, sometimes props are left on the ground by mistake, and the focus on the coat could be highlighting this blooper. This seems like a plausible explanation. <answer>C</answer>

(c) Question: What was the intention of the man in black aiming his gun at the man in red at 0:20?
Options: A. Self-defense counterattack. B. Threatening to extort money. C. Misidentification. D. Prevent alarm. E. Vent anger. F. carry out a cover-up.

VideoVAST: <think>The man in black is aiming his gun at the man in red, which suggests a threatening action. Given the context of the scene, it is likely that the man in black is threatening to extort money from the man in red. </think><answer>F</answer>

Video-R1: <think> The most plausible intention of the man in black is to threaten to extort money from the man in red. This fits the context of the scene where the man in black is holding a gun and the man in red is on the ground. </think><answer>B</answer>

(d) Question: Which item is closest to the core theme of this film?
Options: A. The wisdom of the weak rising up. B. The awareness of natural ecological protection. C. The preciousness of innocent childhood. D. The importance of exercise. E. Those who harm others will ultimately bring misfortune upon themselves. F. The warning of out-of-control technology.

VideoVAST: <think>The film appears to be a fable or a story that involves a character who is initially in a position of power or control but ends up being punished for their actions. This is a common theme in fables where the moral is often about the consequences of one's actions. The other options are either too broad or specific to the context of the film, but the theme of punishment for harming others is a central element in the narrative structure of the film. </think><answer>E</answer>

Video-R1: <think> The film shows a young boy running through a park, interacting with nature, and eventually falling asleep on the grass. The boy's actions and the setting suggest a focus on childhood and innocence. </think><answer>C</answer>