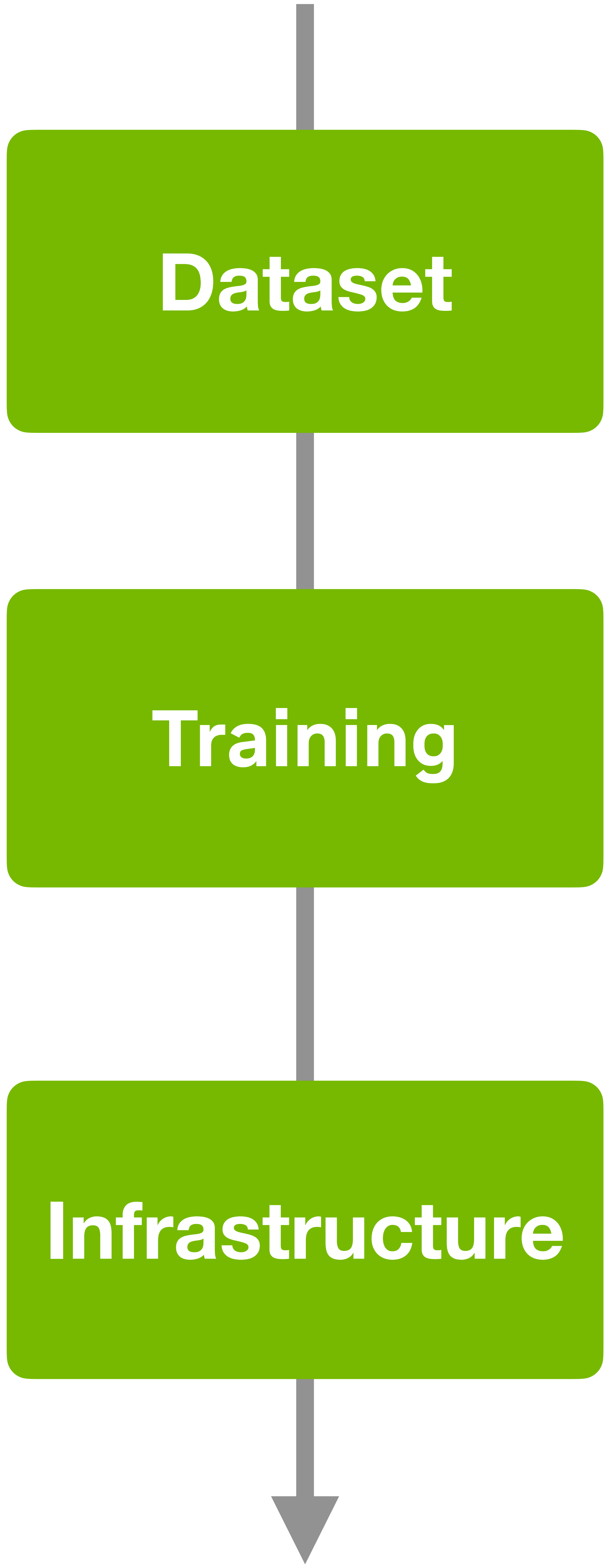
Long-RL - Scaling RL to Long Videos

Long-RL are full-stack solutions for long video VLMs and reasoning.



Dataset

LongVideo Reasoning dataset

A benchmark designed for long video CoT-SFT & RL, with 18k videos (half to 1 hours),
12 video categories, 52k multi-choices questions with reasons and 72k general QAs.

Training

Progressive Training Pipeline

From VILA to LongVILA and to LongVILA-R1,
with long-context extension, high-quantity SFT & CoT SFT and RL (multi-choices).

Infrastructure

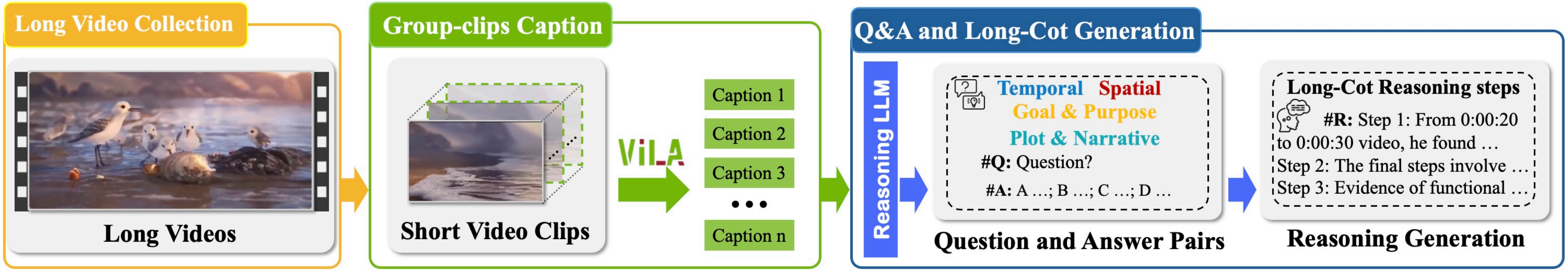
MM-SP & MR-SP (Multi-modal reinforcement sequence parallel)

Systems that enables efficient multi-modal long context SFT & RL with sequence parallelism.

Yukang Chen, Fuzhao Xue, Dacheng Li, et al. Scaling Long-Context Visual Language Models for Long Videos. In ICLR 2025
Yukang Chen, Wei Huang, Baifeng Shi, et al. Scaling RL to Long Videos. In NeurIPS 2025

Long-RL - Scaling RL to Long Videos

Dataset - Data generation process for the LongVideo-Reason dataset



Long Video Collection

Long Videos

Group-clips Caption

Short Video Clips

ViLA

Caption 1

Caption 2

Caption 3

• • •

Caption n

Q&A and Long-Cot Generation

Reasoning LLM

Temporal Spatial

Goal & Purpose

Plot & Narrative

#Q: Question?

#A: A …; B …; C …; D …

Question and Answer Pairs

Long-Cot Reasoning steps

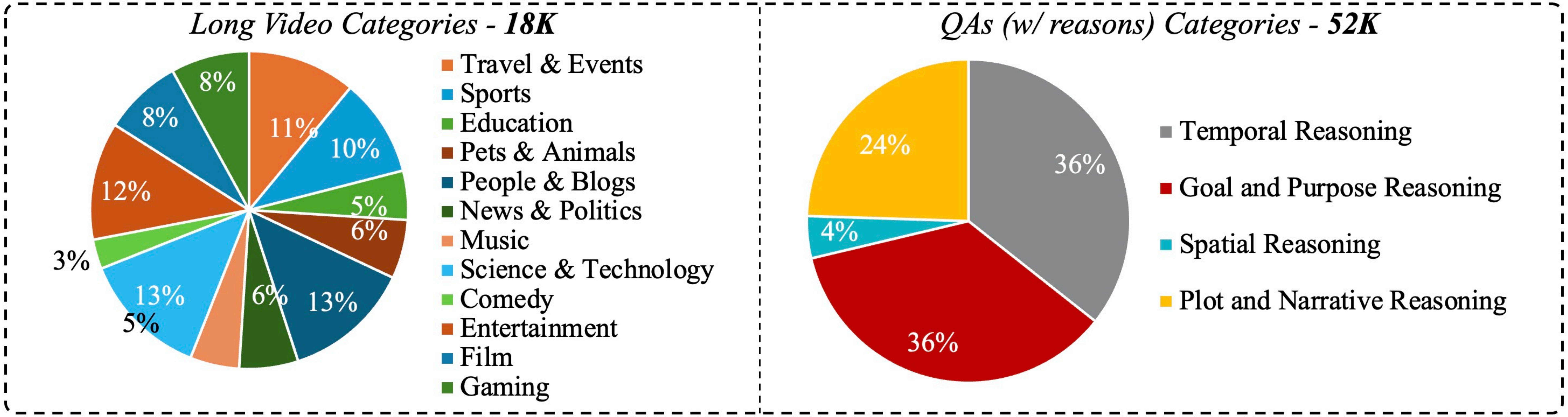
#R: Step 1: From 0:00:20 to 0:00:30 video, he found …

Step 2: The final steps involve …

Step 3: Evidence of functional …

Reasoning Generation

The pipeline turns long videos into short clip captions, then generates Q&A pairs and long CoT reasoning with a reasoning LLM.

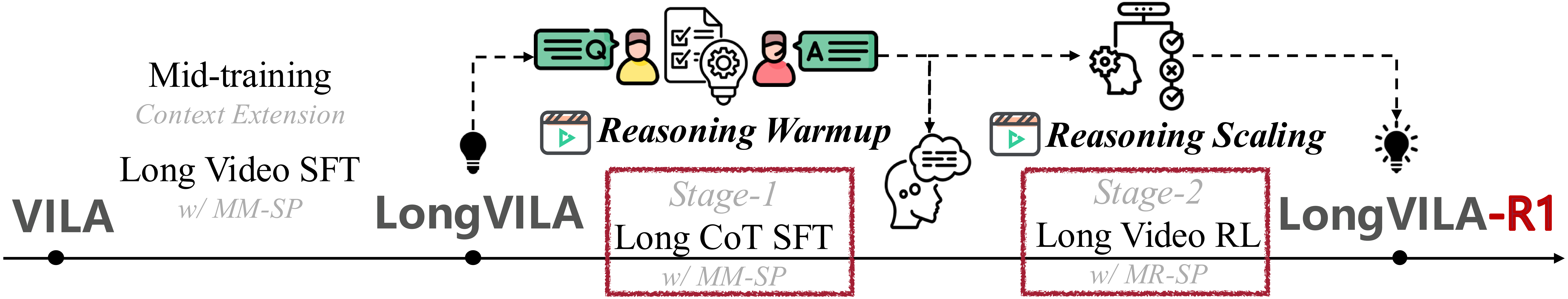


We use the original long videos from shot2story. All datasets used have been approved by the legal team (Ticket DGPTT-3257).

Yukang Chen, Fuzhao Xue, Dacheng Li, et al. Scaling Long-Context Visual Language Models for Long Videos. In ICLR 2025
Yukang Chen, Wei Huang, Baifeng Shi, et al. Scaling RL to Long Videos. In NeurIPS 2025

Long-RL - Scaling RL to Long Videos

Training - from VILA to LongVILA and LongVILA-R1



Mid-training

Context Extension

Long Video SFT

w/ MM-SP

VILA

LongVILA

Reasoning Warmup

Stage-1

Long CoT SFT

w/ MM-SP

Reasoning Scaling

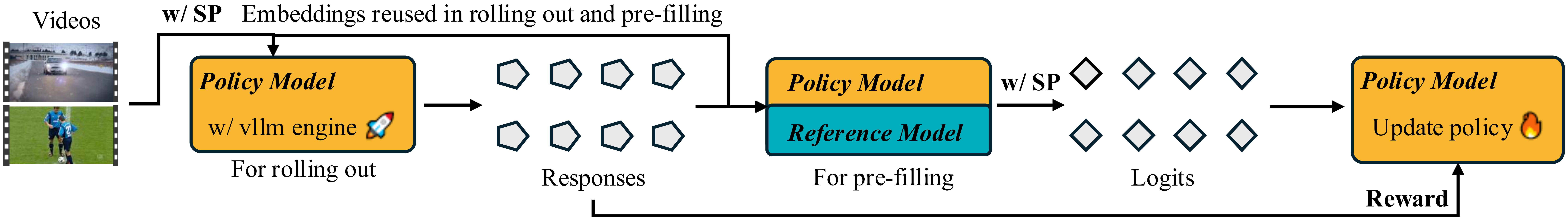
Stage-2

Long Video RL

w/ MR-SP

LongVILA-R1

The pipeline evolves VILA into LongVILA-R1 via long video SFT, CoT SFT, and reasoning RL.



Videos

w/ SP Embeddings reused in rolling out and pre-filling

Policy Model

w/ vllm engine 🚀

For rolling out

Responses

Policy Model

Reference Model

For pre-filling

w/ SP

Logits

Policy Model

Update policy 🔥

Reward

In Long Video RL, we encode video batches with SP, roll out the generations with a policy model, compare them to a reference model to compute rewards, and update the policy using logits.

Yukang Chen, Fuzhao Xue, Dacheng Li, et al. Scaling Long-Context Visual Language Models for Long Videos. In ICLR 2025
Yukang Chen, Wei Huang, Baifeng Shi, et al. Scaling RL to Long Videos. In NeurIPS 2025

Long-RL - Scaling RL to Long Videos

Infrastructure - MR-SP (multi-modal reinforcement sequential parallel)

One Video
Chunk 1
Chunk 2
Chunk 3
GPU 1
GPU 2
GPU 3
Vision Tower
Vision Tower
Vision Tower
Text Prompt
All-Gather
Embeddings
copy
vLLM engine (LLM)
Rollout 1
Rollout 2
Rollout n
(1) Gathering Global Input Embedding
bs
sequence
(2) Padding Sequence
(3) Sharding to Local GPU
GPU 1 GPU 2 GPU 3 GPU 3 GPU 2 GPU 1
1 1 1 2 2 2 3 3 3 3 3 3 2 2 2 1 1 1
1 1 2 2 3 3 3 3 2 2 1 1
1 1 2 2 3 3 3 3 2 2 1 1
1 2 3 3 2 1
GPU 1
GPU 2
GPU 3
LLM
LLM
LLM
Pad Token to Enable P2P Communication

Stage 1 - Rollout with Paralleled Encoding

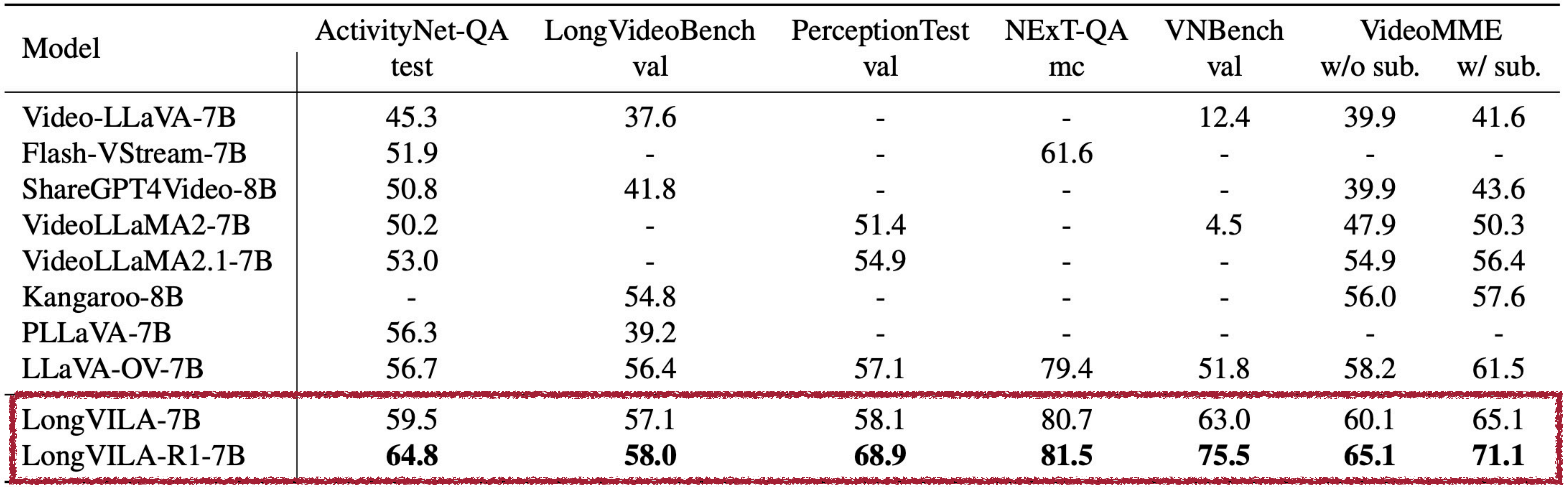
Stage 2 - Prefilling with Sequence Parallelism

Stage 1: Video Chunks are embedded in parallel on GPUs; then gathered, with text prompt, fed to the vLLM engine for response generation.
Stage 2: Gathered input embeddings are padded and sharded across GPUs to prefill the LLM, with P2P communication.

Yukang Chen, Fuzhao Xue, Dacheng Li, et al. Scaling Long-Context Visual Language Models for Long Videos. In ICLR 2025
Yukang Chen, Wei Huang, Baifeng Shi, et al. Scaling RL to Long Videos. In NeurIPS 2025

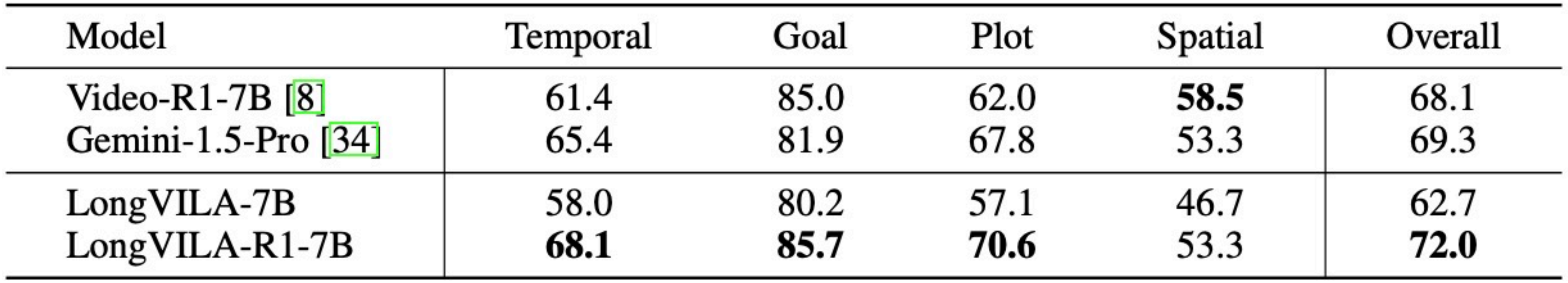
Long-RL - Scaling RL to Long Videos

Performance -10 benchmarks & scaling ablations on accuracy and efficiency

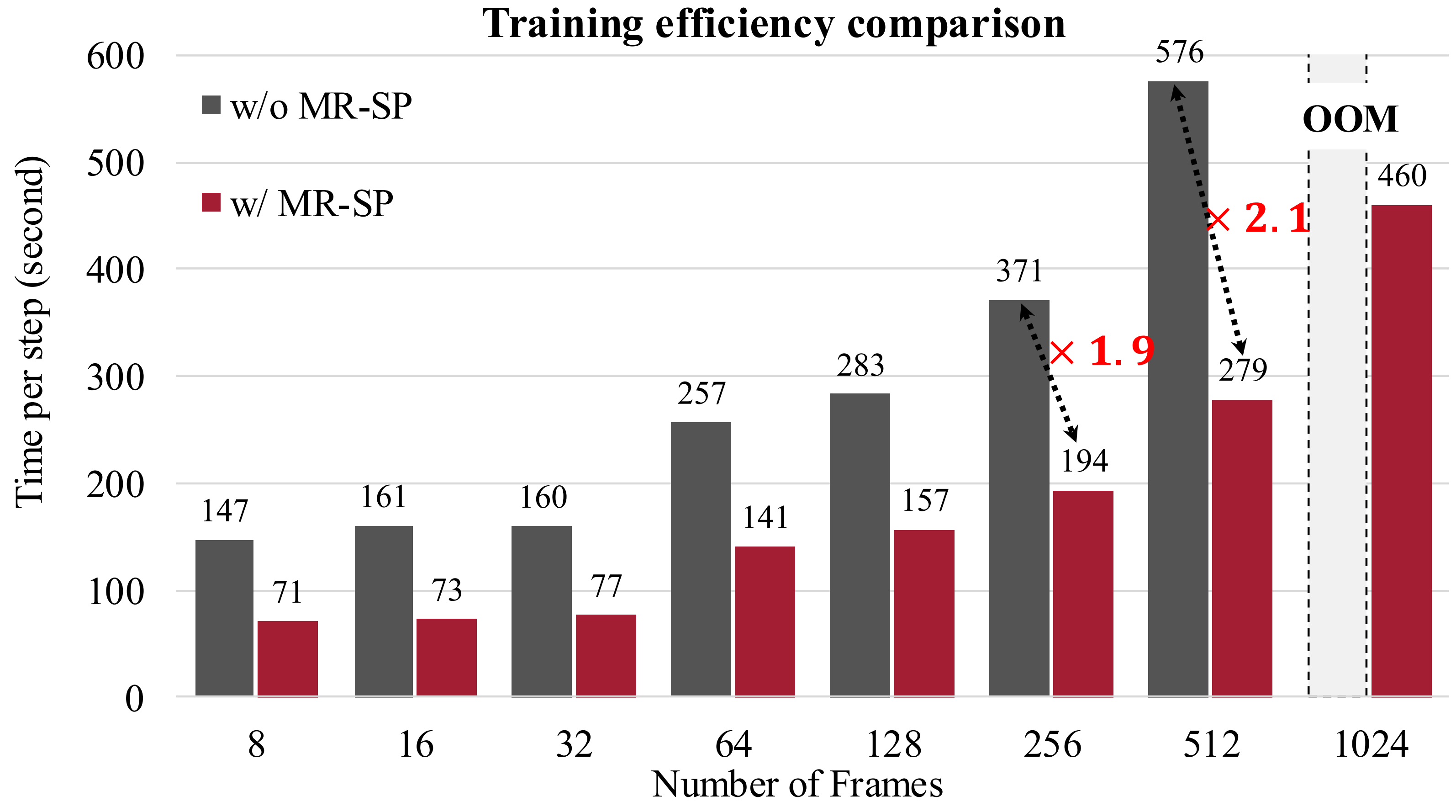


Model	ActivityNet-QA test	LongVideoBench val	PerceptionTest val	NExT-QA mc	VNBench val	VideoMME w/o sub.	VideoMME w/ sub.
Video-LLaVA-7B	45.3	37.6	-	-	12.4	39.9	41.6
Flash-VStream-7B	51.9	-	-	61.6	-	-	-
ShareGPT4Video-8B	50.8	41.8	-	-	-	39.9	43.6
VideoLLaMA2-7B	50.2	-	51.4	-	4.5	47.9	50.3
VideoLLaMA2.1-7B	53.0	-	54.9	-	-	54.9	56.4
Kangaroo-8B	-	54.8	-	-	-	56.0	57.6
PLLaVA-7B	56.3	39.2	-	-	-	-	-
LLaVA-OV-7B	56.7	56.4	57.1	79.4	51.8	58.2	61.5
LongVILA-7B	59.5	57.1	58.1	80.7	63.0	60.1	65.1
LongVILA-R1-7B	**64.8**	**58.0**	**68.9**	**81.5**	**75.5**	**65.1**	**71.1**

LongVILA-R1 has consistent improvement upon LongVILA across video benchmarks.



Model	Temporal	Goal	Plot	Spatial	Overall
Video-R1-7B [8]	61.4	85.0	62.0	**58.5**	68.1
Gemini-1.5-Pro [34]	65.4	81.9	67.8	53.3	69.3
LongVILA-7B	58.0	80.2	57.1	46.7	62.7
LongVILA-R1-7B	**68.1**	**85.7**	**70.6**	53.3	**72.0**



Better efficiency on long video frames.

Yukang Chen, Fuzhao Xue, Dacheng Li, et al. Scaling Long-Context Visual Language Models for Long Videos. In ICLR 2025
Yukang Chen, Wei Huang, Baifeng Shi, et al. Scaling RL to Long Videos. In NeurIPS 2025

Long-RL - Scaling RL to Long Videos

Taboo Game

Question: This a video of two people playing the game Taboo on the topic of 'machine learning'. One person describes a word, and the other person guesses it. Your task is to analyze it and choose the most likely word being described from the following options:

A: Logistic Regression - Pruning - Attention Mechanism
B: Support Vector Machine - Quantization - Robotics
C: Neural Network - Pruning - Gradient Boosting
D: Neural Network - Quantization - Transformer

The word “Logistic Regression” is not actually mentioned; instead, the description “**similar to the human brain, with a lot of layers of nodes and weighted connections” clearly refers to a Neural Network, a fundamental machine learning model originating in the 1950s**. The second clue, about **“cutting the digits of an algorithm and localizing it to an edge device,” matches Quantization, which reduces bit-width to make models more efficient. The mention of “Optimus Prime and Megatron” directly points to Transformer, named after the Transformers franchise and used widely in modern NLP**. Together, Neural Network, Quantization, and Transformer form a coherent chain of concepts matching the video’s clues. Support Vector Machine, Logistic Regression, and Gradient Boosting are not described. Options A and C omit Transformer or include unrelated terms. Therefore, the best answer is D.
<answer>D</answer>✅

Yukang Chen, Fuzhao Xue, Dacheng Li, et al. Scaling Long-Context Visual Language Models for Long Videos. In ICLR 2025
Yukang Chen, Wei Huang, Baifeng Shi, et al. Scaling RL to Long Videos. In NeurIPS 2025

Long-RL - Scaling RL to Long Videos

Football Game - 2022 FIFA Argentina v.s. Netherlands

Question: You are shown a 30-minute video segment from a football game. The score remains 2–2 throughout this period, and no goals are scored. Based on players' physical condition, tactical behavior, emotional state, and overall match performance observed during extra time, which team is more likely to win the upcoming penalty shootout?

A: The Netherlands win the penalty shootout 4–2, thanks to their momentum advantage from coming back in regulation time.
B: Argentina win the penalty shootout 4–3, with their goalkeeper playing a decisive role and key players maintaining composure.
C: The Netherlands win the shootout 5–4 after a flawless penalty performance and two missed shots by Argentina.
D: Argentina score a goal in the final minutes of extra time and win 3–2 without going to penalties.

<think>The score remains 2–2 throughout extra time, indicating a high-intensity match with both teams committed to winning… **Argentina's goalkeeper, who initially struggled, regains confidence and makes a crucial save, demonstrating his improved focus and determination. Key players from both teams, such as Messi and de Ligt, are seen holding their heads in disappointment or celebration, reflecting the emotional highs and lows that often decide penalty shootouts**… Given the evidence, Argentina is more likely to win the penalty shootout 4–3, with their goalkeeper playing a decisive role and key players maintaining composure. </think>
<Answer> B </Answer> ✅

Yukang Chen, Fuzhao Xue, Dacheng Li, et al. Scaling Long-Context Visual Language Models for Long Videos. In ICLR 2025
Yukang Chen, Wei Huang, Baifeng Shi, et al. Scaling RL to Long Videos. In NeurIPS 2025

Long-RL - Scaling RL to Long Videos

Texas Hold’em - 2025 Triton Super High Roller Series

Question: Thomas Mühlöcker (wearing sunglasses and a navy-blue shirt) and Fedor Holz (wearing a black jacket). Thomas Mühlöcker holds the Ace of Clubs and the King of Hearts, while Fedor Holz's cards remain unknown. The community cards are as follows: Flop: 7 of Clubs, Jack of Hearts, 2 of Hearts; Turn: 8 of Diamonds; River: King of Diamonds (visible in the video). Based on the betting patterns observed, assume you are Thomas Mühlöcker and know that your best possible hand is a pair of Kings. After analyzing the video, infer Fedor Holz's hand.

A. Club King, Club Jack B. Heart 3, Heart 10 C. Spade King, Club King D. Diamond Ace, Spade Ace

<think>Fedor Holz’s actions suggest a medium-strength but playable hand. Pre-flop, he raised to 40,000 and called a 3-bet to 115,000—consistent with hands like suited connectors, high cards, or mid-to-high pocket pairs, but likely not premiums he'd 4-bet. **On the flop (7♣ J♥ 2♥), his check-call of a 90,000 c-bet implies decent equity—possibly top pair, a draw, or a strong second-best hand. A monster (e.g., set) would likely raise to deny equity, while weak or missed hands would fold. The turn (8♦) brings more draws; Fedor checks again, likely controlling the pot with top pair or similar. Thomas’s check-back reinforces this line. Hands like 3♥ 10♥ or weak Jacks are unlikely due to pre-flop and post-flop passivity. Premiums like AA or KK would probably have been 4-bet.** The line best fits K♣ J♣—top pair, strong kicker, pot control, and flexibility to respond on the river.</think>
<answer>A. Club King, Club Jack</answer> ✅

Yukang Chen, Fuzhao Xue, Dacheng Li, et al. Scaling Long-Context Visual Language Models for Long Videos. In ICLR 2025
Yukang Chen, Wei Huang, Baifeng Shi, et al. Scaling RL to Long Videos. In NeurIPS 2025