MIRU 2026
OS1F-07
第29回 画像の認識・理解シンポジウム MIRU2026
2026年8月3日～8月6日　出島メッセ長崎

強化学習を用いたマルチモーダル大規模言語モデルによる動画での反復動作カウンティング推論

仇 禕誠
(QIU YICHENG)

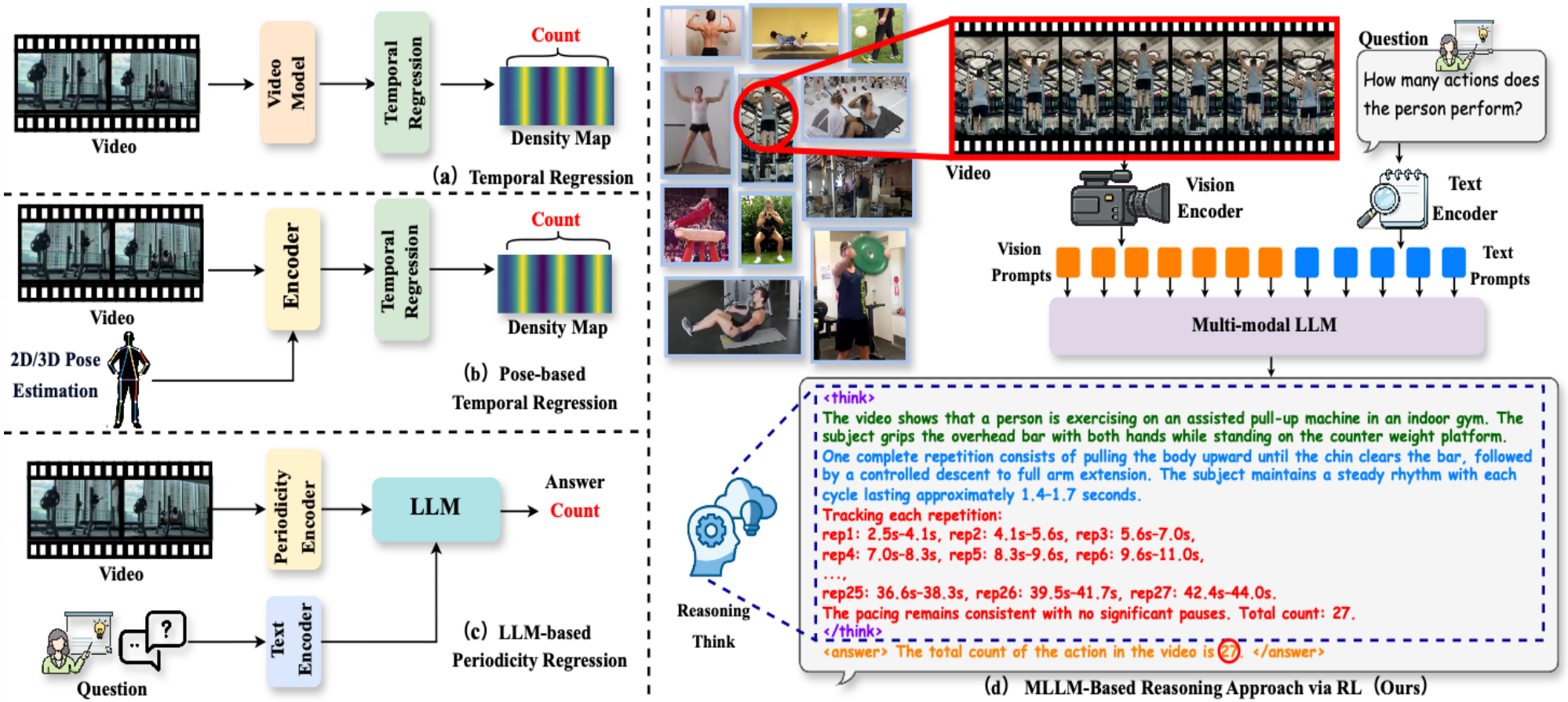
柳井啓司
(YANAI KEIJI)

電気通信大学・大学院
柳井研究室

プラチナスポンサー：CyberAgent. DENSO IT LAB KONICA MINOLTA HPC TECH SB Intuitions HUAWEI

目次

- 研究背景
- 関連研究
- 提案手法
- 実験
- まとめ



反復動作カウント（RAC）手法の枠組み比較。(a)–(c) は従来手法、(d) が私たちの提案手法。

研究背景

❶ マルチモーダル入力

動画 …

テキスト 「このシーンを説明して」

音声 …



この動画は
何をしているところ？

❷ トークン化処理

動画トークン …

テキストトークン …

音声トークン …

位置情報の付与など

0 1 2 3 4 5 …

❸ MLLM



統合理解・推論・生成

❹ 出力：自然言語での説明・回答

この動画は、晴れた日に草地で
1人の選手がボールをドリブルし、
その後シュートを決める様子を
示しています。
これはサッカーのプレーです。



理解　推論　説明・回答

動画から特徴を学習し、
動作や時間区間を予測

テキストや音声を組み合わせた
マルチモーダル学習で文脈理解を強化

動画・テキスト・音声を統一的に扱う
MLLMにより、自然言語で説明が可能に

研究背景

反復動作カウント（Repetitive Action Counting、RAC）



・動作カウンティング(RAC) ：
時間軸に沿って反復の進行を追跡し回数を推定する

・動画理解や動作検出における
より高度な時系列理解タスクである。

・幅広い応用される。

CyberAgent. DENSO IT LAB KONICA MINOLTA HPC TECH SB Intuitions HUAWEI

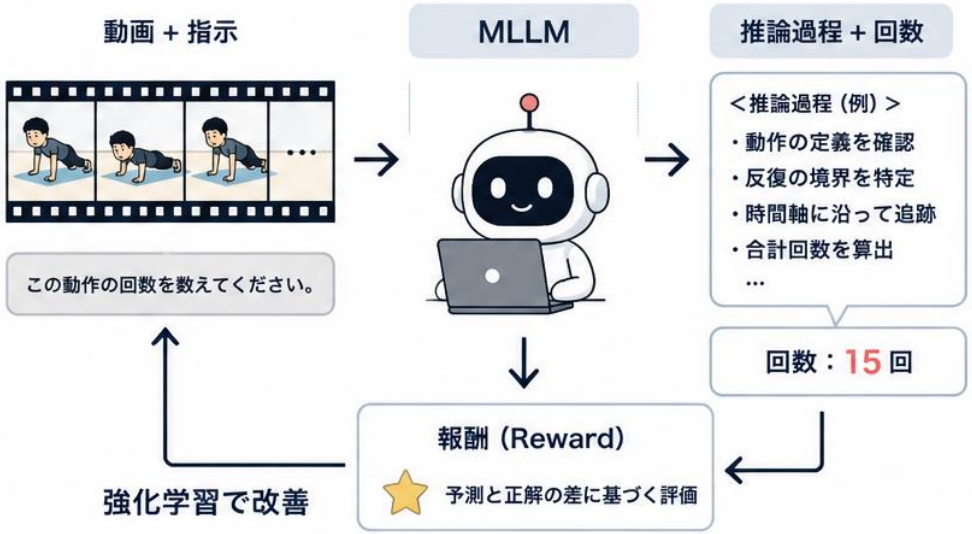
研究目的

なぜ MLLM を用いるのか

汎用視覚理解
未知ドメインへの一般化能力

自然言語による推論過程
「なぜその数か」を説明可能
（解釈性）

マルチタスク対応
計数・動作分類・時間ローカライズを
統一的に扱える



克服すべき 3 つの課題

課題 A
MLLM には
RAC 特化の体系的な
推論パラダイムが存在しない

課題 B
訓練データの
計数分布が
極端に不均衡

課題 C
標準的な SFT は
誤差を一様に扱い、
細粒度な計数最適化が困難

・ MLLMは映像中のRACを高精度かつ解釈可能

・既存の MLLM はRACに十分適応できず

・強化学習による報酬最適化を組み合わせ

関連研究

- **時系列回帰**：

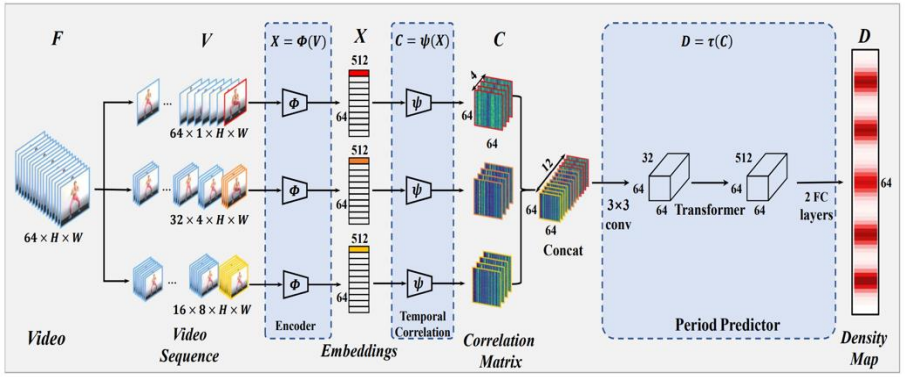
TransRAC [Hu+, CVPR22] ;
ESCounts [Sinha+, ACCV24] ;
ME-RAC [Qiu+, ESWA24]

- **姿勢ベース**：

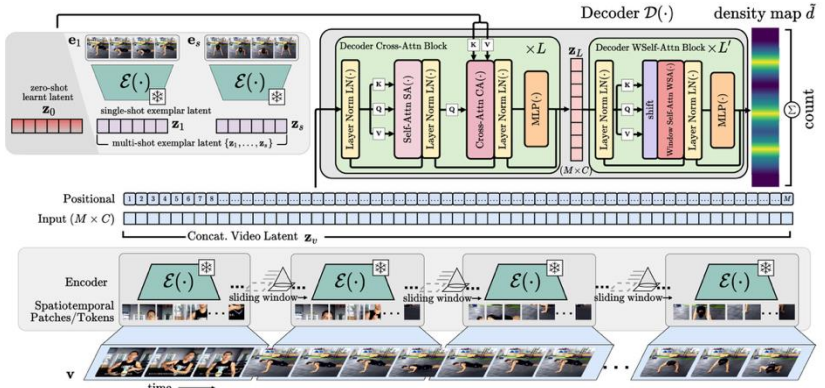
PoseRAC [Yao+, arXiv23] ;
MoCount [Gu+, ACMMM25]

- **言語 LLM ベース**：

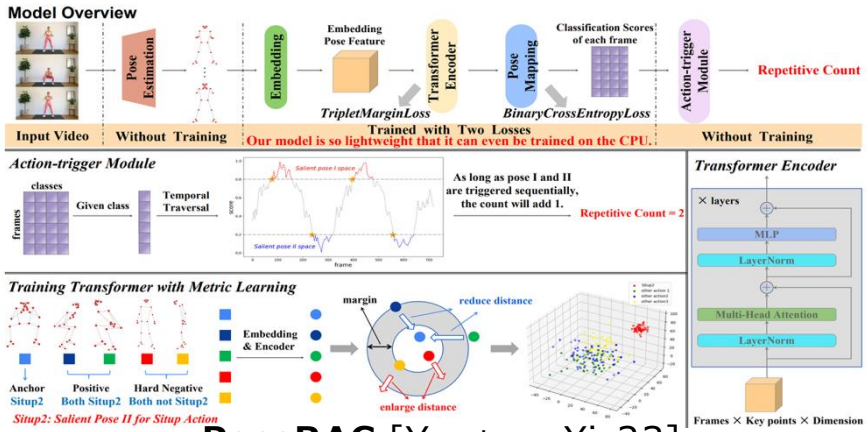
CountLLM [Yao+, CVPR25] ;
OVRCounter [Dwibedi+, arXiv24]



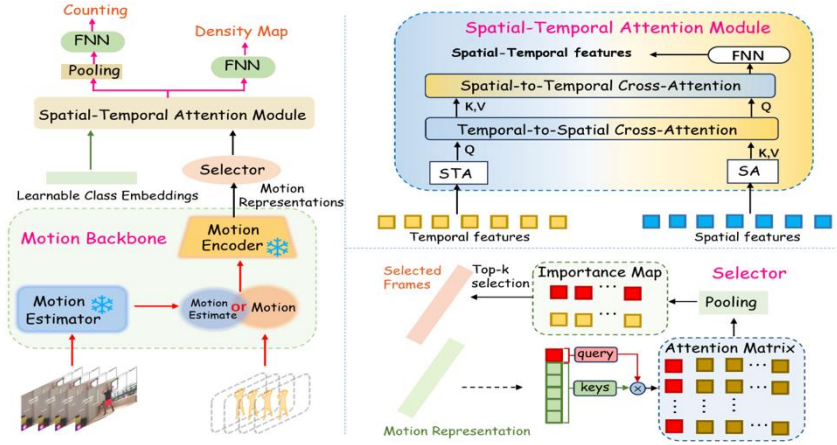
TransRAC [Hu+, CVPR22]



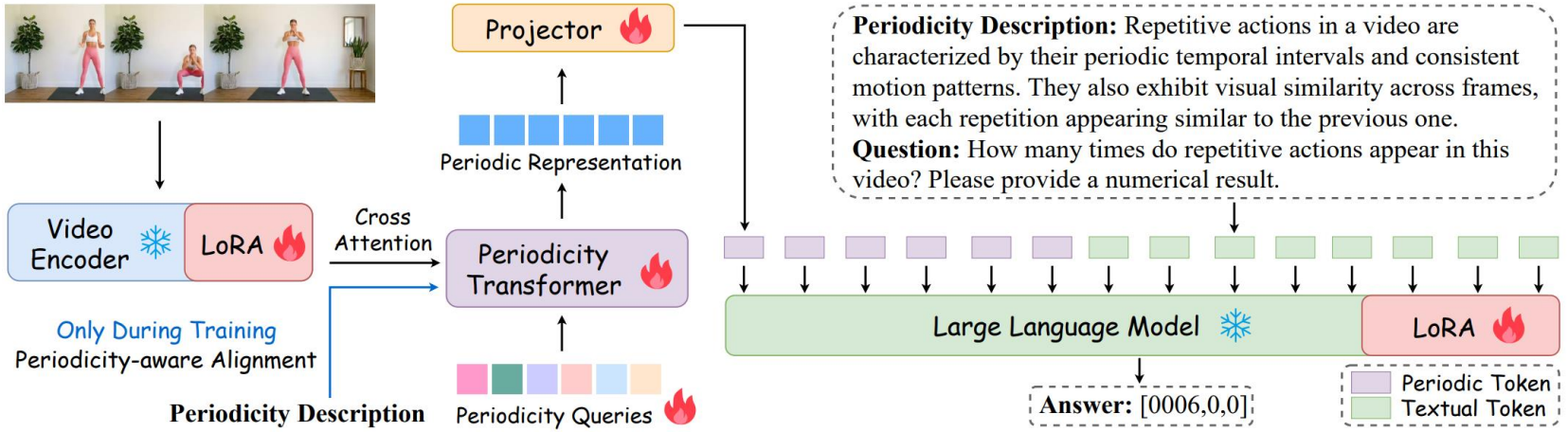
ESCounts [Sinha+, ACCV24]



PoseRAC [Yao+, arXiv23]

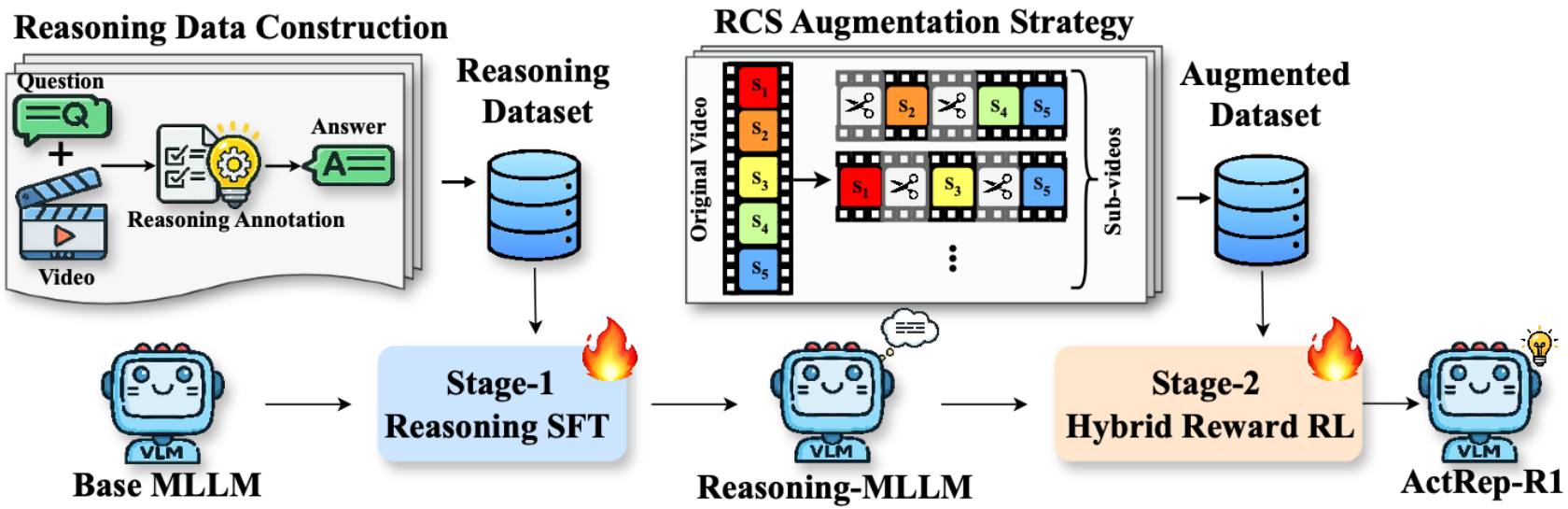


MoCount[Gu+, ACMMM25]



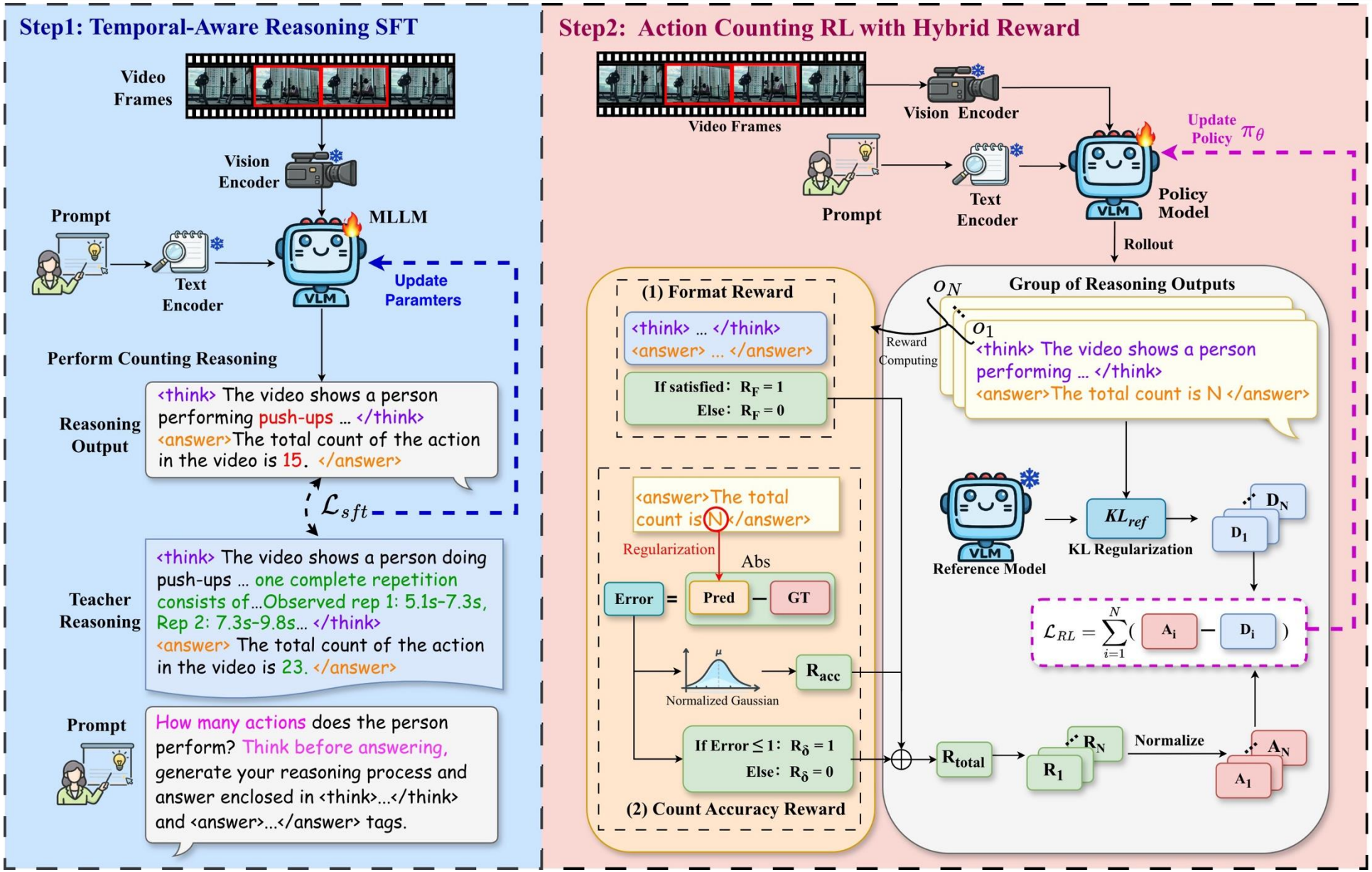
CountLLM [Yao+, CVPR25]

提案手法



- ActRep-R1: RACに特化した推論フレームワーク を提案する。
- 既存のRACデータに基づく、MLLMの学習に特化したデータセットを構築した。
- 場面理解・動作定義・時間的追跡・最終カウントという段階的な推論過程を学習する。

提案手法



学習順序：

・構造化推論データSFT により学習させる。

・データ拡張戦略を用いた強化学習

・「動画を見て、考えて、正確に数える」能力を獲得

実装

設定

学習方法：事後学習、強化学習GRPO

設備: RTX PRO 6000

データセット：
RepCount-A [Hu+, CVPR22] ; UCFRep [Zhang+, CVPR20] ;
Countix [Dwibedi+, CVPR20] ; OVR [Dwibedi+, arXiv24] ;
QUVA [Runia+, CVPR18]

基盤モデル：　Qwen3-VL、Qwen2.5-VL、などの事前学習MLLM







実験

主の性能結果

商用 LLMと比較して大幅な性能向上を達成した。

より小規模なモデルで高いカウント精度を実現

Method	Params	RepCount-A (Hu et al. 2022)				UCFRep (Zhang et al. 2020)				Countix (Dwibedi et al. 2020)			
		MAE↓	OBO↑	OBZ↑	CR↑	MAE↓	OBO↑	OBZ↑	CR↑	MAE↓	OBO↑	OBZ↑	CR↑
TransRAC (Hu et al. 2022)	–	0.443	0.291	0.160	–	0.441	0.430	0.237	–	–	–	–	–
PoseRAC (Yao, Cheng, and Zou 2023)	–	0.236	0.560	–	–	0.312	0.452	–	–	–	–	–	–
ME-RAC (Qiu, Niu, and Sha 2024)	–	0.353	0.402	0.241	–	0.487	0.460	0.276	–	–	–	–	–
MoCount (Gu et al. 2025)	–	0.174	0.649	0.297	–	0.122	0.805	0.424	–	–	–	–	–
ESCounts (Sinha, Stergiou, and Damen 2024)	–	0.213	0.563	0.245	–	0.216	0.704	0.381	–	0.276	0.673	0.319	–
CountLLM (Yao et al. 2025)	7B	0.162	0.639	0.435	–	0.116	0.839	0.571	–	0.215	0.801	0.545	–
Qwen3-VL-Plus*† (Bai et al. 2025a)	–	0.545	0.388	0.180	0.942	0.578	0.365	0.170	0.938	0.612	0.345	0.160	0.935
Gemini-3-Pro*† (Google DeepMind 2025)	–	0.576	0.341	0.158	0.925	0.612	0.318	0.146	0.921	0.648	0.302	0.139	0.918
GPT-5.4*† (Singh et al. 2025)	–	0.540	0.400	0.193	0.938	0.573	0.378	0.182	0.934	0.608	0.361	0.173	0.931
Video-R1 (Feng et al. 2025)	7B	0.388	0.405	0.205	0.924	0.412	0.372	0.185	0.916	0.438	0.348	0.168	0.908
VideoChat-R1 (Li et al. 2025)	7B	0.402	0.388	0.192	0.908	0.428	0.355	0.172	0.900	0.455	0.332	0.155	0.892
Time-R1 (Wang et al. 2025b)	7B	0.395	0.398	0.198	0.912	0.420	0.365	0.178	0.904	0.448	0.340	0.162	0.895
ActRep-R1 (Qwen2.5-VL-3B + SFT + RL)	3B	0.195	0.498	0.352	0.942	0.218	0.685	0.462	0.945	0.212	0.668	0.441	0.938
ActRep-R1 (Qwen3-VL-4B + SFT + RL)	4B	0.128	0.641	0.468	0.955	0.138	0.793	0.558	0.958	0.146	0.776	0.540	0.952
ActRep-R1 (Qwen2.5-VL-7B + SFT + RL)	7B	0.135	0.652	0.482	0.958	0.148	0.806	0.572	0.962	0.152	0.789	0.554	0.955
ActRep-R1 (Qwen3-VL-8B + SFT)	8B	0.142	0.632	0.474	0.952	0.164	0.817	0.583	0.956	0.158	0.798	0.562	0.948
ActRep-R1 (Qwen3-VL-8B + SFT + RL)	**8B**	**0.104**	**0.697**	**0.518**	**0.964**	**0.111**	**0.844**	**0.612**	**0.968**	**0.119**	**0.829**	**0.596**	**0.962**

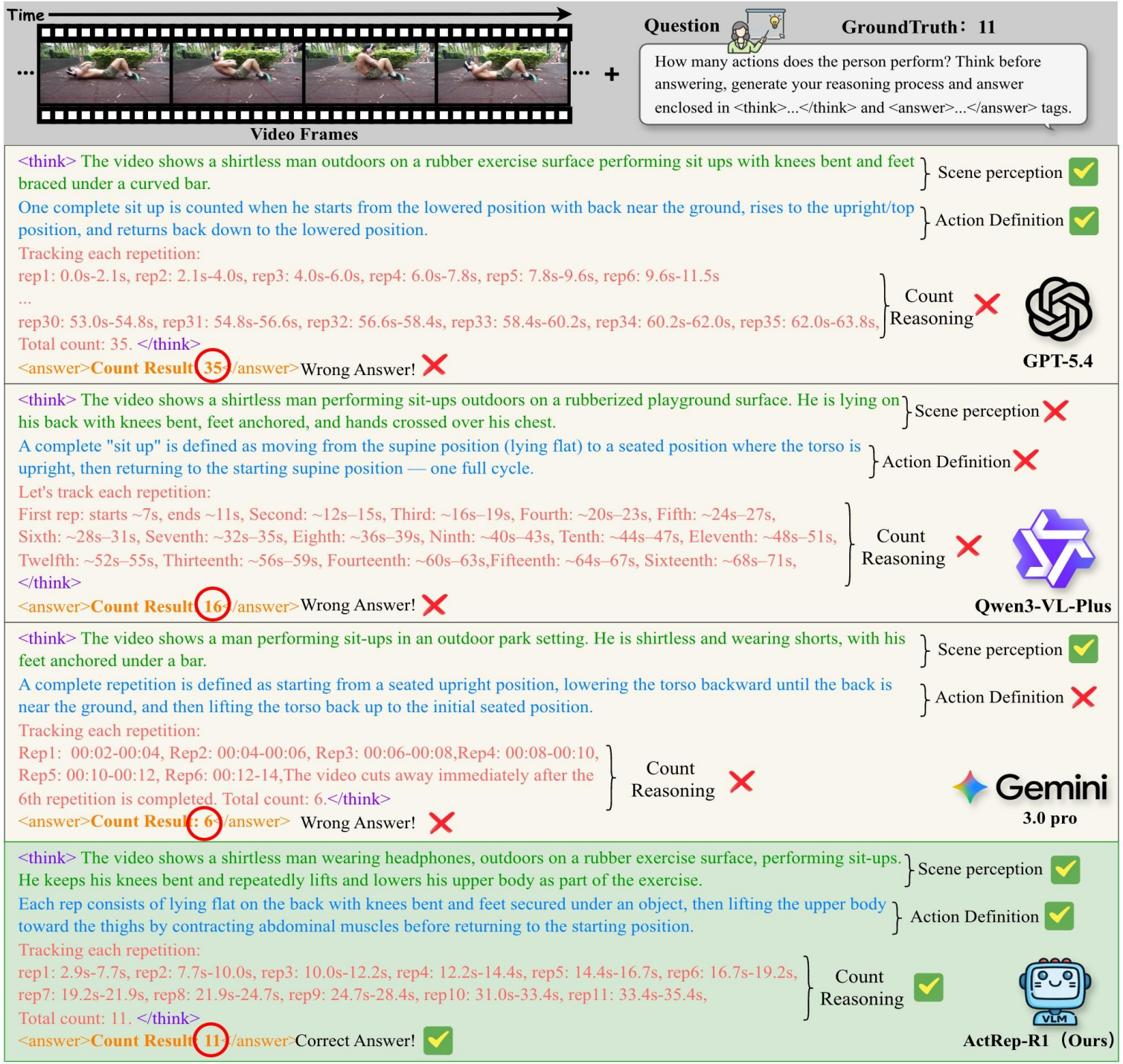
Table 1: Performance comparison of our method with various previous methods and proprietary commercial models on mainstream RAC datasets. *: evaluated zero-shot without RAC-specific training; †: closed-source proprietary models accessed via API.

実験

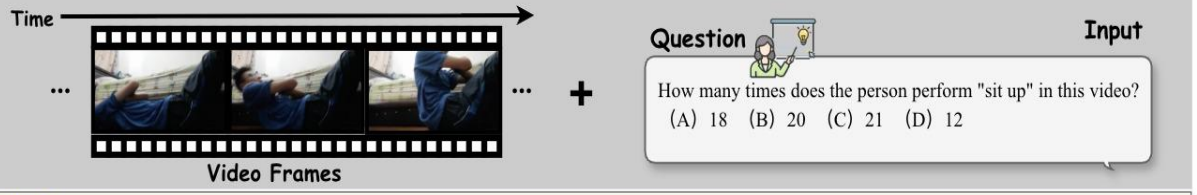
定性分析

・商用モデルが誤カウントする場合

・提案手法は時間軸に沿って反復動作を追跡し、正確な回数を推定



実験



<think>
The person starts lying down, then repeatedly raises the torso to a seated crunch position and returns down. Counting each full sit-up cycle visible in the sequence gives 12 repetitions before stopping and reaching toward the camera.
</think>
<answer> A </answer> Answer Misaligned ❌

Thinking Misjudged ❌

GPT-5.4

<think>
Counting each sit-up performed by the person in the video, it appears that they complete a total of 21 repetitions.
</think>
<answer> D </answer> Answer Misaligned ❌

Thinking Correct ✅

Qwen3-VL-Plus

<think>
The video shows a person performing sit-ups on the floor next to a bed.
A complete sit-up is defined as starting from a lying position, lifting the torso up towards the knees, and lowering back down. Tracking each repetition: ,..., The closest option is 12.
</think>
<answer> D </answer> Answer Misaligned ❌

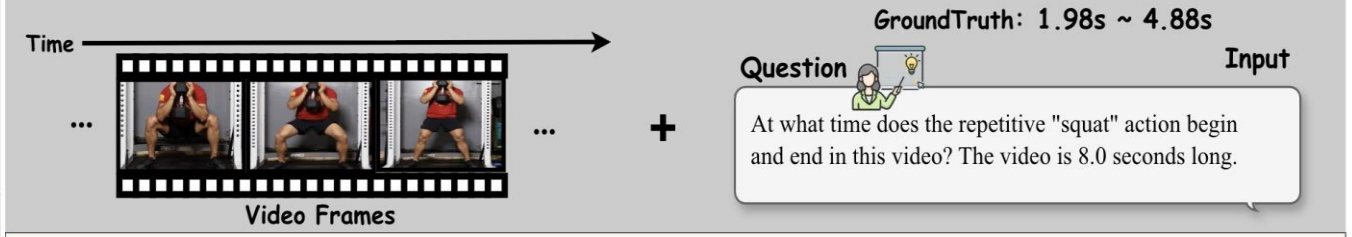
Thinking Correct ❌

Gemini 3.0 pro

<think>
The video shows a person performing sit-ups repeatedly while lying on their back on a wooden floor. Each rep consists of lifting the upper body upward from the floor by contracting the abdominal muscles until the elbows reach close to the knees, then lowering back down. Tracking the repetitions ,..., I count 21 complete sit-up cycles.
</think>
<ansewr> C </answer> Answer Correct ✅

Thinking Correct ✅

ActRep-R1（Ours）



<think>
The video shows a person performing pull-ups on a bar in an indoor gym setting. The repetitive action appears to span the majority of the video, with the subject gripping the bar and performing controlled pull-up movements throughout the clip.
</think>
<answer> 0.00 - 8.00 </answer> tIoU = 0.363 ❌

GPT-5.4

<think>
The repetitive "squat" action begins at the start of the video and continues until the end.
</think>
<answer> [0.00s, 8.00s] </answer> tIoU = 0.363 ❌

Qwen3-VL-Plus

<think>
The video shows a man performing dumbbell goblet squats in a home gym. He begins the squatting motion at approximately 00:01 and completes the visible repetitions by approximately 00:06. The remaining portion of the video shows him standing still.
</think>
<answer> [1.0s, 6.0s] </answer> tIoU = 0.483 ❌

Gemini 3.0 pro

<think>
The video shows a person performing squats inside a home gym while holding a dumbbell at chest level. The squatting motion begins at approximately 2.0 seconds when the person starts lowering into the squat position, and the last repetition completes at approximately 4.8 seconds when they return to standing.
</think>
<answer> [2.02s, 4.83s] </answer> tIoU = 0.969 👍

ActRep-R1（Ours）

MIRU 2026
第29回 画像の認識・理解シンポジウム MIRU2026
2026年8月3日～8月6日 出島メッセ長崎
プラチナスポンサー：CyberAgent. DENSO IT LAB KONICA MINOLTA HPC TECH SB Intuitions HUAWEI

まとめ

・MLLM と強化学習を用いた RAC 手法 ActRep-R1 を提案

・構造化推論 SFT と強化学習により、高精度かつ説明可能なカウンティングを実現

・既存手法および商用 MLLM と比較し、RAC における有効性を確認

・**詳細はポスター発表 OS1F-07 をご覧ください**

参考文献

[1] H. Hu et al., "TransRAC: Encoding Multi-scale Temporal Correlation with Transformers for Repetitive Action Counting", CVPR 2022.
[2] Y. Qiu et al., "Multipath 3D-Conv Encoder and Temporal-Sequence Decision for Repetitive-Action Counting", Expert Systems with Applications, 2024.
[3] S. Sinha et al., "Every Shot Counts: Using Exemplars for Repetition Counting in Videos", ACCV 2024.
[4] Z. Yao et al., "PoseRAC: Pose Saliency Transformer for Repetitive Action Counting", arXiv:2303.08450, 2023.
[5] R. Gu et al., "MoCount: Motion-Based Repetitive Action Counting", ACM Multimedia 2025.
[6] Z. Yao et al., "CountLLM: Towards Generalizable Repetitive Action Counting via Large Language Model", CVPR 2025.
[7] D. Dwibedi et al., "OVR: A Dataset for Open Vocabulary Temporal Repetition Counting in Videos", arXiv:2407.17085, 2024.
[8] H. Zhang et al., "Context-Aware and Scale-Insensitive Temporal Repetition Counting", CVPR 2020.
[9] D. Dwibedi et al., "Counting Out Time: Class Agnostic Video Repetition Counting in the Wild", CVPR 2020.
[10] T. Runia et al., "Real-World Repetition Estimation by Div, Grad and Curl", CVPR 2018.
[11] S. Bai et al., "Qwen2.5-VL Technical Report", arXiv:2502.13923, 2025 / Qwen Team, "Qwen3-VL Technical Report", arXiv:2511.21631, 2025.
[12] Z. Shao et al., "DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models", arXiv:2402.03300, 2024.