CONFIDENTIAL EXTENDED ABSTRACT.
DO NOT DISTRIBUTE ANYWHERE.

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[1,a)]　柳井 啓司[1,b)]

概要

動画における反復動作カウンティングは依然として重要なタスクである．従来の視覚ベース手法は特定の動作パターンへの過学習や解釈可能性の欠如といった性能の頭打ちに直面している．一方，マルチモーダル大規模言語モデルは汎用的な視覚理解と自然言語による推論を通じた解釈可能性を提供するが，反復動作カウンティングにおいては十分なカウンティング精度を達成できていない．本研究では，マルチモーダル大規模言語モデルの推論能力を反復動作カウンティングに活用するフレームワーク ActRep-R1 を提案する．時間認識型タイミング検証を備えた構造化推論データ構築パイプライン，カウント分布の偏りに対処する Random Count Sampling 拡張，およびカウント正規化ハイブリッド報酬を用いた強化学習を統合する．複数のベンチマークにおいて最高精度を達成し，ActRep-R1 の有効性を実証した．

1. はじめに

反復動作カウンティング（Repetitive Action Counting; RAC）は，動画中で人物が周期的動作を何回実行したかを判定するタスクである．フィットネス動作の分析 [7]，周期運動の解析 [10], [19]，動画の時間的理解 [5], [18] など多様な応用を持つ．動作速度の変動，遮蔽，視点変化，動作境界の曖昧さにより，正確なカウンティングは困難である [16], [20]．従来手法は専用アーキテクチャにより着実に進歩してきた．RepNet [4] は時間的自己類似性行列による繰り返しパターン検出，TransRAC [9] は Transformer ベースの密度マップ回帰，視聴覚融合 [28] や動作特徴学習 [12] による精度向上，ESCounts [22] は例示フレームを活用した訓練不要カウンティング，CountLLM [26] は動画をクリップに分割し大規模言語モデル（Large Language Model; LLM）で集約する手法を提案している．しかし，これらは手動設計の回帰ネットワークに依存し，表現力の限界，訓練パターンへの過学習，解釈不能な予測といった問題から性能の頭打ちに達している．マルチモーダル大規模言語モデル（Multimodal Large Language Model; MLLM）[1], [2], [13], [14] は有望な代替手段を提供する．大規模実世界データでの事前学習により広範な視覚理解能力を備え，自然言語による段階的推論を通じた解釈可能性を有する．Chain-of-Thought（CoT） [25] は構造化推論の引き出しに有効であり，GRPO（Group Relative Policy Optimization） [3], [21] は自己回帰損失でなくタスク固有指標を直接最適化する．近年，Video-R1 [6]，VideoChat-R1 [11]，Time-R1 [24]，VAU-R1 [29] がこれらのパラダイムを動画理解に適用し始めている．

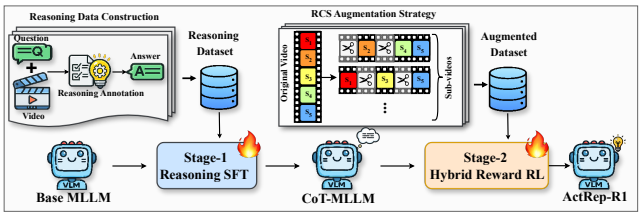


図 1　ActRep-R1 フレームワークの全体構成．**(a)** 推論データ構築，**(b) RCS** 拡張，**(c) 2** 段階訓練．

しかし，RAC 向けの訓練データと評価プロトコルが不足しており，既存 MLLM のカウンティング精度は不十分である．本研究では，MLLM の推論能力を RAC に活用する ActRep-R1 を提案する．主な貢献は以下の通りである．

(1) 時間認識型タイミング検証を備えた構造化推論データ構築パイプラインを提案する．
(2) カウント分布の偏りに対処する RCS を導入する．
(3) カウント正規化ハイブリッド報酬による強化学習訓練フレームワークを設計する．

2. 関連研究

反復動作カウンティング．初期の RAC アプローチは視覚的周期性の検出 [16] や微分幾何的手法 [20] など手動設計の特徴量に依存していた．RepNet [4] は時間的自己類似性行列により繰り返しパターンを同定し，TransRAC [9] は Transformer ベースの密度マップ回帰を導入する．視聴覚

[1] 電気通信大学大学院　情報理工学研究科 情報学専攻
[a)] qiu-y@mm.inf.uec.ac.jp
[b)] yanai@cs.uec.ac.jp

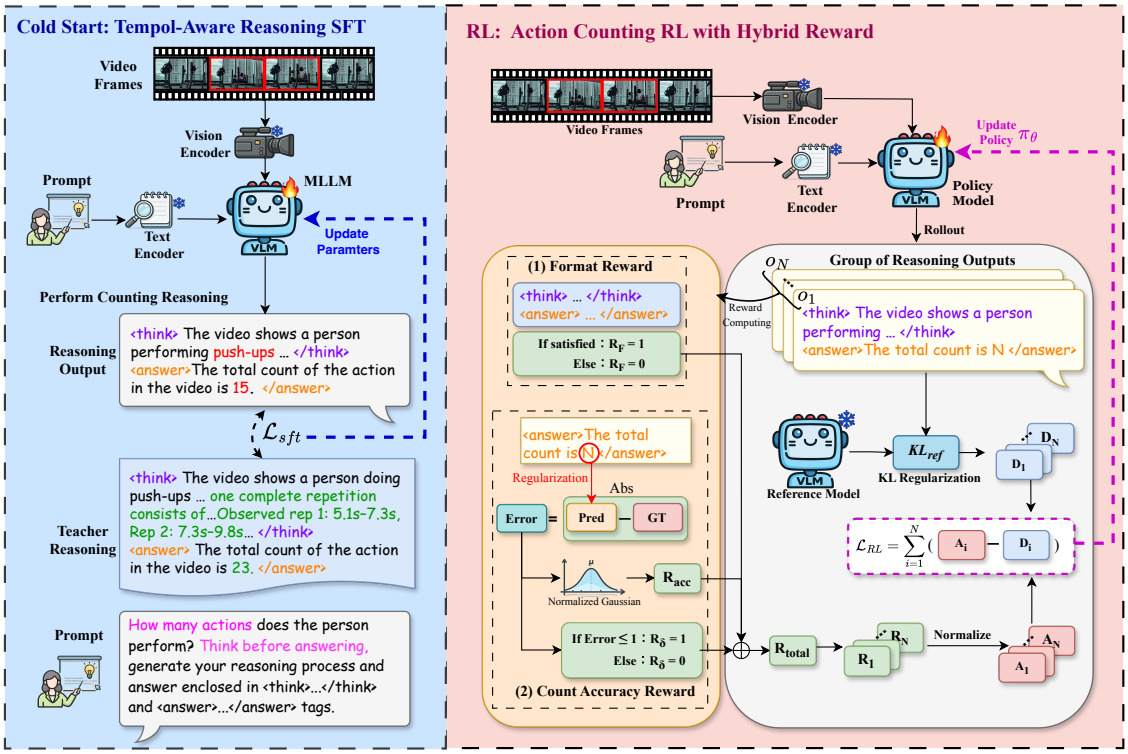


図 2　2 段階訓練パイプライン．**Stage 1（SFT）**：視覚エンコーダとテキストエンコーダを凍結して微調整．**Stage 2（強化学習）**：G=8 の候補応答からカウント正規化ハイブリッド報酬でポリシー勾配を更新．

融合 [28] や動作特徴学習 [12] に基づくカウンティング手法も提案されている．ESCounts [22] は少数例示による訓練不要カウンティング，IVAC-P^{2}L [23] は不規則反復の事前知識を活用する．CountLLM [26] は動画をクリップに分割し LLM でカウントを集約する．

強化学習による推論強化． CoT 推論 [25] は LLM の段階的推論を引き出す手法として広く研究されている．DeepSeek-R1 [3] は GRPO [21] で推論能力を強化し，Video-R1 [6] はこのパラダイムを動画 QA に拡張した．さらに，VideoChat-R1 [11] は時空間知覚の強化，Time-R1 [24] は時間的グラウンディング，VAU-R1 [29] は動画異常理解に本パラダイムを適用している．しかし，RAC に特化した推論フレームワークは未だ提案されておらず，動画内の反復パターンに適した構造化推論の設計と報酬関数が求められる．

3. 提案手法

図 1 に ActRep-R1 の全体構成を示す．構造化推論データ構築とデータ拡張，教師ありファインチューニング（Supervised Fine-Tuning; SFT）による推論形式の学習，GRPO によるカウンティング精度の最適化の 3 段階からなる．

3.1 構造化推論データ構築

RAC における構造化推論の主な困難は，反復回数が広範囲にわたる点にある．各反復を逐次列挙する形式は高カウント動画でトークン制限を超過するため，本手法では反復回数に依存しない**記述的推論形式**を採用する．具体的には，場面知覚，動作定義，計数推論，回答の 4 フィールドで構成し，反復回数に依存しない一定長の推論チェーンを維持する．各訓練動画について時間セグメント注釈を解析し，セグメント長が大きく変動する動画を変動ペーシング，リズムが一定の動画を均一ペーシングと分類する．この分類に基づき，時間認識型のタイミングヒント，正解動作ラベル，目標カウントを教師 MLLM に与える構造化プロンプトを構成する．教師 MLLM [1] が動画と構造化プロンプトに基づき推論アノテーションを生成する．生成テキストは重複文の退化検出，正解との整合性を確認するカウント照合，文境界での切り詰めと正解カウント保証を行うトークン制限の 3 段階品質制御を経て訓練データを構成する．

3.2 Random Count Sampling（RCS）によるデータ拡張

RAC 訓練集合のカウント分布は低カウントに大きく偏っており，この偏りは予測バイアスを引き起こす．先行研究の Random Count Augmentation（RCA） [23] は逐次クロップに限定され，正確なセグメント注釈を保持できない．RCS はこれらの制約を解決し，任意位置でのランダムセグメント削除により逐次クロップよりも多様な時間構成を生成するとともに，対数スケーリングに基づく分布認識型のサブ動画数決定を導入し，オフライン生成により正確なセグメント注釈を保持する．層別削除比サンプリングにより

表 1　本研究の提案手法と従来手法および商用プロプライエタリ LLM の性能比較.

カテゴリ	手法	Params	RepCount-A		UCFRep		Countix	
			MAE↓	OBO↑	MAE↓	OBO↑	MAE↓	OBO↑
時間回帰手法	RepNet [4]	–	0.995	0.013	0.998	0.009	0.364	0.303
	TransRAC [9]	–	0.443	0.291	0.441	0.430	–	–
	ME-RAC [17]	–	0.353	0.402	0.487	0.460	–	–
	ESCounts [22]	–	0.213	0.563	0.216	0.704	0.276	0.673
	CountLLM [26]	7B	0.162	0.639	0.116	0.839	0.215	0.801
商用プロプライエタリ LLM	Qwen3-VL-Plus [1]	–	0.533	0.331	0.561	0.313	0.594	0.298
	Gemini-3-Pro [8]	–	0.576	0.341	0.612	0.318	0.648	0.302
	GPT-5.4 [15]	–	0.540	0.400	0.573	0.378	0.608	0.361
提案手法	Qwen3-VL-8B + SFT	8B	0.142	0.632	0.164	0.817	0.158	0.798
	Qwen3-VL-8B + SFT + GRPO	8B	**0.104**	**0.697**	**0.111**	**0.844**	**0.119**	**0.829**

低カウント領域を重点的に補完し，元のカウント分布の偏りを是正する．

3.3　強化学習フレームワーク

訓練は図 2 に示す 2 段階で進行する．**SFT 段階**では事前学習済み MLLM [1] の視覚エンコーダとテキストエンコーダを凍結し微調整を行い，`<think>`ブロックに推論チェーン，`<answer>`ブロックに最終カウントを配置する構造化推論形式を学習する．この分離により，GRPO ではカウント抽出と報酬計算を確実に行える．**GRPO 段階**では GRPO [3], [21] に基づき，各プロンプト x に対し G 個の応答 $\{y^{(1)}, \ldots, y^{(G)}\}$ を生成し，以下のポリシー勾配で更新する：

$$\nabla_\theta \mathcal{L} = -\frac{1}{G}\sum_{i=1}^{G} A^{(i)} \nabla_\theta \log \pi_\theta(y^{(i)}|x) \qquad (1)$$

ここでグループ内相対的優位度 $A^{(i)}$ は以下で定義される：

$$A^{(i)} = \frac{R(y^{(i)}) - \mu_R}{\sigma_R + \epsilon} \qquad (2)$$

$R(y^{(i)})$ は i 番目の応答の報酬，μ_R と σ_R はグループ内報酬の平均と標準偏差である．

カウンティングタスクに適した報酬として，精度報酬 R_{acc} と近接報酬 R_δ の 2 成分からなるハイブリッド報酬を設計する：

$$R_{\text{count}}(y) = \underbrace{\exp\left(-\frac{e_{\text{norm}}^2}{2\sigma_{\text{norm}}^2}\right)}_{R_{\text{acc}}} + \underbrace{\mathbb{1}[|\hat{n} - n| \le 1] \cdot \alpha}_{R_\delta} \qquad (3)$$

ここで $e_{\text{norm}} = |\hat{n} - n|/(n + 0.1)$ はカウント正規化誤差である．R_{acc} はカウント正規化により，カウント数の大小に関わらず同一の相対的精度に対し一貫した報酬を付与する．二値報酬のスパース性や線形ペナルティの一様な勾配を回避する設計である．R_δ は誤差 ± 1 以内の予測に追加報酬を与える．応答からカウントを抽出できない場合は $R_{\text{count}} = 0$ とする．加えて`<think>`...`<answer>`タグの整合性を検証するフォーマット報酬 $R_{\text{format}} \in \{0, 1\}$ を定義し，総報酬 $R(y) = R_{\text{count}}(y) + R_{\text{format}}(y)$ とする．

4.　実験

4.1　実験設定

データセット．3 つのベンチマークで評価する．**RepCount-A** [9] は 9 種のフィットネス動作を含む最も広く使用される RAC ベンチマークである．**UCFRep** [27] は 23 カテゴリに拡張される．**Countix** [4] はフィットネス以外の多様な動作を含み汎化性能のテストとして最も挑戦的である．評価指標として正規化平均絶対誤差（Mean Absolute Error; MAE）と OBO（Off-By-One）を使用する．ベースモデルは事前学習済み MLLM [1] を用い，SFT で視覚エンコーダとテキストエンコーダを凍結した微調整，GRPO でカウント報酬に基づく強化学習を行う．

4.2　主要結果

表 1 に 3 ベンチマークでの比較結果を示す．ゼロショット MLLM は汎用動画理解を有しても RAC では精度が不十分であるのに対し，SFT のみで MAE を 0.142 に低減し，構造化推論がアーキテクチャ変更なしに体系的なカウンティングを教示できることを示した．さらに GRPO により，SFT が確立した推論形式の上でカウンティング精度が校正され，従来の時間回帰手法と商用プロプライエタリ LLM の双方を上回って全 3 ベンチマークで最高精度を達成した．これは，専用アーキテクチャを必要としない汎用 MLLM ベースのアプローチの有効性を示している．

4.3　定性的分析

図 3 に ActRep-R1 の推論出力例を示す．ActRep-R1 は正確な場面知覚，精密な動作定義，正しい計数推論を多様な動画にわたり示す．商用プロプライエタリ LLM は強力な汎用動画理解を有しながらも信頼性の高いカウンティングを達成できず誤カウントを生じる傾向があるのに対し，ActRep-R1 は構造化推論によりタスクを解釈可能なサブタスクに分解し，信頼性の高いカウンティングを実現する．

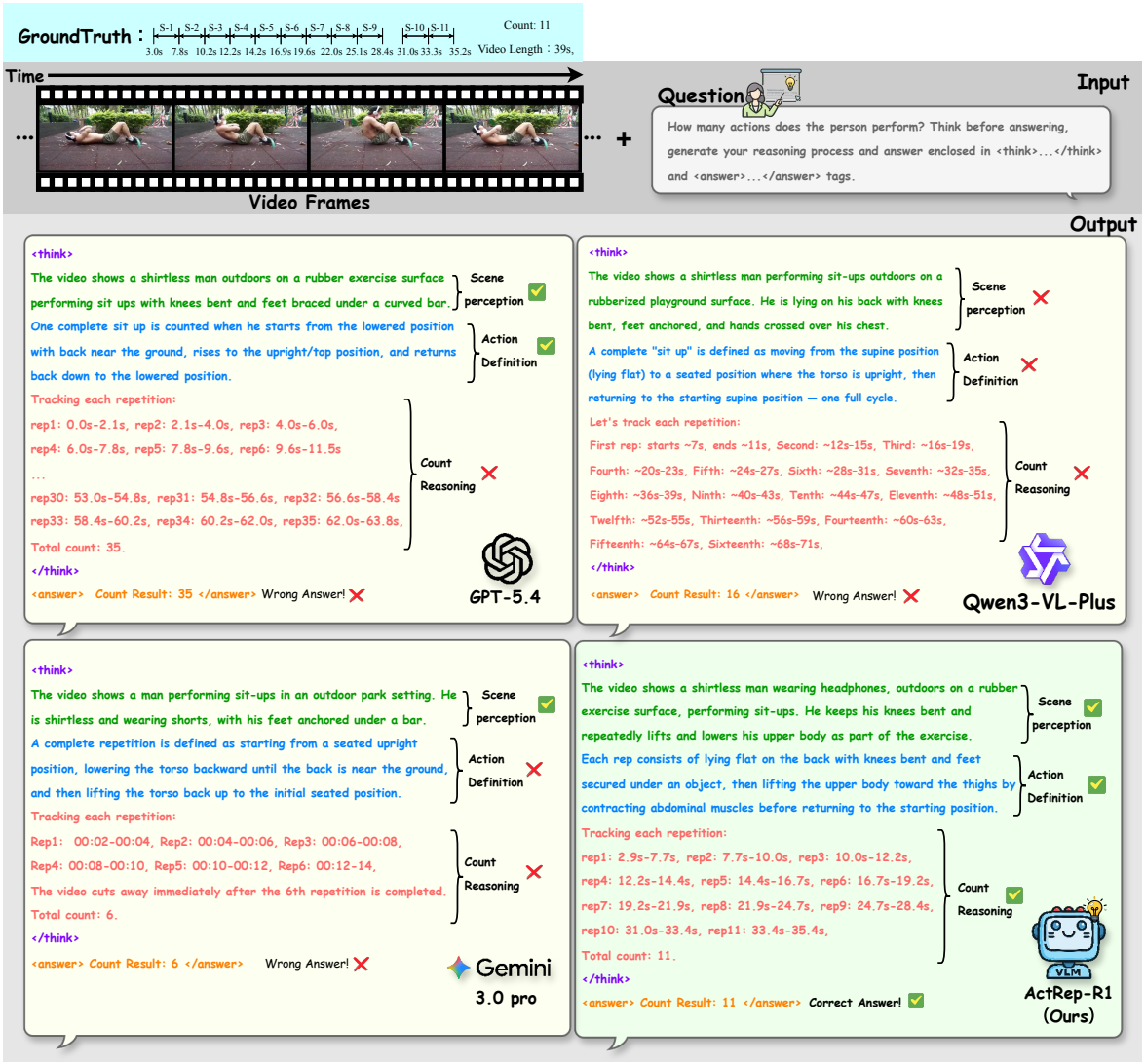


図 3 ActRep-R1 と商用プロプライエタリ LLM の定性比較．ActRep-R1 は段階的推論によりカウンティングの各ステップを明示し正確な予測を出力するのに対し，商用プロプライエタリ LLM は体系的な推論戦略を欠き誤カウントする傾向がある．

5. おわりに

本研究では，MLLM に構造化推論を教示し強化学習で精度を最適化する RAC フレームワーク ActRep-R1 を提案した．記述的推論形式によるカウンティングのサブタスク分解，RCS によるカウント分布の偏り是正，カウント正規化ハイブリッド報酬による GRPO の 3 つの手法を統合し，複数の RAC ベンチマークで既存手法を上回る最高精度を達成した．専用の時間回帰ネットワークを必要とせず，汎用 MLLM のみで RAC が高精度に解決可能であることを実証した．

本研究は初期段階の成果であり，今後は各構成要素の寄与を検証するアブレーション実験，より多様なデータセットでの汎化性能評価，および実装詳細の整備を進める．加えて，時間的接地やテスト時スケーリングなどの機能拡張も検討する．

参考文献

[1] Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y. and Zhu, K.: Qwen3-VL Technical Report, *arXiv preprint arXiv:2511.21631* (2025).

[2] Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H. and Lin, J.: Qwen2.5-VL Technical Report, *arXiv preprint arXiv:2502.13923* (2025).

[3] DeepSeek-AI: DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning, *arXiv preprint arXiv:2501.12948* (2025).

[4] Dwibedi, D., Aytar, Y., Tompson, J., Sermanet, P. and Zisserman, A.: Counting out time: Class agnostic video repetition counting in the wild, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10387–10396 (2020).

[5] Dwibedi, D., Aytar, Y., Tompson, J. and Zisserman, A.: Ovr: A dataset for open vocabulary temporal repetition counting in videos, *arXiv preprint arXiv:2407.17085* (2024).

[6] Feng, K., Gong, K., Li, B., Guo, Z., Wang, Y., Peng, T., Wu, J., Zhang, X., Wang, B. and Yue, X.: Video-r1: Reinforcing video reasoning in mllms, *arXiv preprint arXiv:2503.21776* (2025).

[7] Fieraru, M., Zanfir, M., Pirlea, S. C., Olaru, V. and Sminchisescu, C.: Aifit: Automatic 3d human-interpretable feedback models for fitness training, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9919–9928 (2021).

[8] Google DeepMind: Gemini 3 Pro Model Card (2025).

[9] Hu, H., Dong, S., Zhao, Y., Lian, D., Li, Z. and Gao, S.: Transrac: Encoding multi-scale temporal correlation with transformers for repetitive action counting, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19013–19022 (2022).

[10] Huang, S., Ying, X., Jiang, J., Rong, Z., Shang, H. and Zha, H.: Camera calibration from periodic motion of a pedestrian, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3025–3033 (2016).

[11] Li, X., Yan, Z., Meng, D., Dong, L., Zeng, X., He, Y., Wang, Y., Qiao, Y., Wang, Y. and Wang, L.: VideoChat-R1: Enhancing Spatio-Temporal Perception via Reinforcement Fine-Tuning, *arXiv preprint arXiv:2504.06958* (2025).

[12] Li, X. and Xu, H.: Repetitive action counting with motion feature learning, *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 6499–6508 (2024).

[13] Liu, H., Li, C., Wu, Q. and Lee, Y. J.: Visual instruction tuning, *Advances in Neural Information Processing Systems*, Vol. 36, pp. 34892–34916 (2023).

[14] Maaz, M., Rasheed, H., Khan, S. and Khan, F. S.: Video-ChatGPT: Towards detailed video understanding via large vision and language models, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 12585–12602 (2024).

[15] OpenAI: GPT-5 System Card, *arXiv preprint arXiv:2601.03267* (2025).

[16] Pogalin, E., Smeulders, A. W. and Thean, A. H.: Visual quasi-periodicity, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8 (2008).

[17] Qiu, Y., Niu, L. and Sha, F.: Multipath 3D-Conv encoder and temporal-sequence decision for repetitive-action counting, *Expert Systems with Applications*, Vol. 249, p. 123760 (2024).

[18] Qiu, Y., Sha, F. and Niu, L.: Efficient Temporal Attention with State Space Model for Temporal Action Localization, *International Conference on Neural Information Processing*, Springer, pp. 183–197 (2024).

[19] Ran, Y., Weiss, I., Zheng, Q. and Davis, L. S.: Pedestrian detection via periodic motion analysis, *International Journal of Computer Vision*, Vol. 71, pp. 143–160 (2007).

[20] Runia, T. F., Snoek, C. G. and Smeulders, A. W.: Real-world repetition estimation by div, grad and curl, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9009–9017 (2018).

[21] Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y. and Guo, D.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models, *arXiv preprint arXiv:2402.03300* (2024).

[22] Sinha, S., Stergiou, A. and Damen, D.: Every shot counts: Using exemplars for repetition counting in videos, *Proceedings of the Asian Conference on Computer Vision*, pp. 3056–3073 (2024).

[23] Wang, H., Cheng, Z.-Q., Du, Y. and Zhang, L.: IVAC-$\mathbf{P^2L}$: Leveraging Irregular Repetition Priors for Improving Video Action Counting, *IEEE Transactions on Multimedia*, Vol. 27, pp. 8325–8339 (2025).

[24] Wang, Y., Wang, Z., Xu, B., Du, Y., Lin, K., Xiao, Z., Yue, Z., Ju, J., Zhang, L., Yang, D., Fang, X., He, Z., Luo, Z., Wang, W., Lin, J., Luan, J. and Jin, Q.: Time-R1: Post-Training Large Vision Language Model for Temporal Video Grounding, *arXiv preprint arXiv:2503.13377* (2025).

[25] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q. V. and Zhou, D.: Chain-of-thought prompting elicits reasoning in large language models, *Advances in Neural Information Processing Systems*, Vol. 35, pp. 24824–24837 (2022).

[26] Yao, Z., Cheng, X., Huang, Z. and Li, L.: Countllm: Towards generalizable repetitive action counting via large language model, *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 19143–19153 (2025).

[27] Zhang, H., Xu, X., Han, G. and He, S.: Context-aware and scale-insensitive temporal repetition counting, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 670–678 (2020).

[28] Zhang, Y., Shao, L. and Snoek, C. G.: Repetitive activity counting by sight and sound, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14070–14079 (2021).

[29] Zhu, L., Chen, Q., Shen, X. and Cun, X.: VAU-R1: Advancing Video Anomaly Understanding via Reinforcement Fine-Tuning, *arXiv preprint arXiv:2505.23504* (2025).