Video-Based Prediction of VO_2 Max Using Running Mechanics

Gustavo Martinez
Stanford University
gm3603@stanford.edu

Eddy Martinez*
Alisal High School
eddymart1121@gmail.com

Abstract

Laboratory measurement of VO_2 max—the maximal oxygen uptake during incremental exercise—is the clinical gold standard for assessing cardiovascular fitness and endurance, yet requires expensive equipment and specialized personnel (roughly 200 dollars per test). We propose a cost-effective, non-invasive alternative: predicting VO_2 max directly from video footage of runners using spatiotemporal pose dynamics. To this end, we assembled a dataset of 200 publicly available YouTube clips of runners with known VO_2 max values and mile personal records. Each video is processed through an automated pipeline (download → buffer-trim → re-encode → MediaPipe pose estimation) to extract 33 keypoint trajectories per frame. We then train and evaluate three regression architectures—Recurrent Neural Networks, Transformer encoders, and Multilayer Perceptrons—on these temporal keypoint sequences to predict VO_2max.

1. Introduction

VO_2 max—the maximal rate of oxygen uptake during progressive exercise—is universally regarded as the gold-standard metric of cardiorespiratory fitness and endurance capacity [2]. Unfortunately, direct assessment requires laboratory gas-exchange analyzers, trained staff, and typically costs $150–$250 per test, placing it beyond the reach of many recreational athletes and public-health initiatives [4].

Decades of exercise-science research show that running economy—operationalized via stride length, ground-contact time, vertical oscillation, and other spatiotemporal gait variables—correlates strongly with VO_2 max and performance outcomes [15, 7]. Concurrently, pose-estimation advances such as MediaPipe Pose, OpenPose, and BlazePose now deliver frame-level joint coordinates from ordinary RGB video in real time [13, 3]. These developments suggest a provocative question: *Can we infer VO_2 max directly from the biomechanics visible in consumer-grade running footage?*

Several groups have pursued phone-based or wearable-sensor surrogates for VO_2 max, leveraging photoplethysmography or single-lead ECG data to approximate laboratory tests [**?**, 8]. In contrast, video-only estimation remains largely unexplored—particularly in unconstrained "in-the-wild" conditions where illumination, camera angle, and runner demographics vary widely.

My work. I assembled a new dataset of 200 publicly available YouTube clips featuring distance and middle-distance runners whose VO_2 max values and mile personal records (PRs) are publicly reported. Each clip is processed through an automated pipeline: download, buffer-trim, re-encode for OpenCV compatibility, and pose extraction of 33 landmarks per frame via MediaPipe. We experiment with three families of sequence regressors—Recurrent Neural Networks (RNN), Transformer encoders, and Multilayer Perceptrons (MLP)—to predict VO_2 max from the resulting keypoint time series.

Our *contributions* are:

- An open, reproducible pipeline that transforms raw running video into pose-based time-series suitable for physiological regression.
- The first systematic comparison of RNN, Transformer, and MLP architectures for video-only VO_2 max prediction.
- An empirical analysis of dataset bias arising from the over-representation of elite athletes in publicly available footage, with discussion of failure modes.

2. Related Work

2.1. Laboratory and Surrogate VO_2 Max Testing

Direct VO_2 max assessment relies on metabolic carts that sample expired gases during graded exercise, yielding gold-standard accuracy but high cost and logistical complexity [2]. To lower these barriers, recent studies have explored wrist-worn photoplethysmography (PPG) devices coupled with machine-learning regressors, achieving root

*Not enrolled in CS 231N. Assisted with data collection (video sourcing and annotation)

mean square errors (RMSEs) below 4 $mL \cdot kg^{-1} \cdot min^{-1}$ in controlled treadmill tests [9]. Single-lead patch ECG sensors have likewise been paired with gradient-boosting models for peri-operative VO_2 max estimation in pulmonary patients [8]. Smartphone-based protocols such as the 2kmFIT-App and Apple's Heart Snapshot combine GPS, inertial, and camera data to estimate fitness levels remotely [12, 18], enabling large-scale deployments beyond clinical labs. However, these sensor-centric approaches require the user to own or wear specific hardware and remain reliant on heart-rate response, rather than underlying movement biomechanics.

2.2. Running Economy and Biomechanics

Exercise-science literature consistently links running economy variables—stride length, ground-contact time, vertical oscillation—to VO_2 max and performance across distances. A comprehensive review [1] reported curvilinear increases in submaximal VO_2 when stride length deviates from self-selected cadence, underscoring the delicate balance of biomechanical efficiency. Ground-contact time, in particular, shows a significant negative correlation with both 5 km race times and laboratory VO_2 max measurements [6]. These findings motivate biomechanical-first estimators that bypass the need for physiological sensors entirely, focusing instead on how spatiotemporal gait variables encode metabolic demand.

2.3. Markerless Pose Estimation

The emergence of markerless systems like OpenPose [**?**], MediaPipe Pose [**?**], and BlazePose democratizes kinematic capture, offering real-time 2D (and recently 3D) joint landmark data from commodity cameras. Systematic evaluations demonstrate that multi-camera OpenPose reconstructions can achieve sub-centimeter joint-center error compared to optical marker-based systems in controlled lab environments [11]. Field experiments in televised competitions further highlight markerless feasibility under in-the-wild conditions, with mean RMSE below 5 mm [5]. Recent surveys [16] conclude that markerless pipelines drastically reduce data-collection overhead while enabling retrospective re-analysis as algorithms improve. These developments underscore the potential of video-based pose tracking as a cost-effective alternative to marker-based motion capture.

2.4. Machine Learning on Pose Sequences

Early works in this area relied on handcrafted gait features fed into linear or polynomial regressors for energetic-cost prediction. More recent studies leverage deep sequence models that better capture temporal dependencies. For instance, CNN-LSTM hybrids operating on single-IMU signals have been shown to predict instantaneous oxygen uptake during team-sport simulations with coefficients of determination (R^2) exceeding 0.90 [14]. Transformer-based architectures, initially developed for natural language processing, now dominate skeleton-based action recognition and gait analysis. Gait-specific variants such as GaitPT [20] and GaitFormer [10] capture long-range spatial and temporal dependencies while maintaining data efficiency. In clinical and prosthetics applications, Vision-Transformer pose estimators [19, 17] improve joint-tracking robustness, opening new frontiers in personalized biomechanics.

2.5. Video-Based VO_2 Max Prediction

Despite these promising advances, video-only VO_2 max estimation remains under-explored. A recent PLOS ONE study used wearable kinematics (rather than video) to estimate oxygen uptake during intermittent sports drills [14]. To our knowledge, no prior work systematically regresses laboratory VO_2 max from markerless running kinematics captured in unconstrained YouTube footage. This gap in the literature underscores the novelty of our approach, which complements sensor-based efforts by testing whether freely available video data—processed through modern pose-estimation pipelines—can reveal enough signal to predict aerobic capacity directly.

2.6. Summary and Gaps

Collectively, existing literature establishes that (i) VO_2 max can be approximated from surrogate heart-rate or IMU signals with promising accuracy, and (ii) markerless pose estimation reliably quantifies biomechanics in uncontrolled environments. Yet the intersection—learning VO_2 max directly from video-derived kinematics—remains largely unexplored and unvalidated. Our work aims to bridge this gap, leveraging pose-based time-series data and advanced sequence models to probe the feasibility of video-only aerobic capacity prediction. In doing so, we also seek to highlight potential biases (such as elite-athlete over-representation) and to stimulate broader discussions on democratizing performance testing in endurance sports.

3. Dataset

3.1. Source Videos

We curated a collection of 200 publicly available YouTube clips featuring middle- and long-distance runners whose *laboratory-measured* VO_2 max values and mile personal records (PRs) are publicly reported in interviews, athlete bios, or scientific case studies. To maximize ecological validity, we retained videos recorded in diverse settings—track workouts, road time-trials, treadmill sessions, even race broadcasts—captured with smartphones, DSLR cameras, and professional television rigs. Each clip contains a continuous running segment of 10 s in which the

	Train	Val	Test
Clips	140	30	30
Athletes	104	23	23
Frames (k)	7.1	1.2	1.4
VO_2 max (mean ± SD)	72.07 ± 4.4	67.8 ± 6.2	70.4 ± 5.9

Table 1. Dataset split statistics. Splits are *athlete-disjoint* to prevent identity leakage.

focal athlete is the only runner fully visible for the majority of frames.

3.2. Label Extraction

For every athlete we recorded:

- **VO_2 max** ($mL \cdot kg^{-1} \cdot min^{-1}$), obtained from published lab tests, elite-program media guides, or coach interviews.
- **Mile PR** (s) as an auxiliary ground-truth indicator of aerobic capacity.
- **Clip metadata**: YouTube URL, resolution, frame-rate, camera angle, lighting conditions.

The resulting label distribution spans 55–84 $mL \cdot kg^{-1} \cdot min^{-1}$ (median $\approx$ 68) with a long tail above 75 due to world-class athletes.

3.3. Preprocessing Pipeline

All videos pass through an automated pipeline (Figure 1):

1. **Download** via `yt-dlp` with the highest 30 fps stream.
2. **Buffer Trim**Using user-provided timestamps, we extract a window ± 5 s around the target segment to tolerate annotation error.
3. **Re-encode**Clips are transcoded to H.264 @720p to ensure deterministic OpenCV decoding.
4. **Pose Extraction**We run MediaPipe Pose (full-body, 33 landmarks) on every frame, storing 3-D coordinates (x, y, z) and per-joint visibility in JSON files; missing detections (¡0.3% of frames) are forward-filled.
5. **Sequence Normalization**Each clip is resampled to a fixed temporal length $T = 240$ frames (8 s @30 fps).

3.4. Dataset Statistics

Clips average 240 ± 55 frames (8 ± 1.8 s). Table 1 summarizes the athlete-level train/val/test split, stratified by VO_2 max quintiles to preserve distributional balance.

figures/pipeline_overview.pdf

Figure 1. End-to-end pipeline from raw YouTube video to VO_2 max prediction.

3.5. Limitations and Bias

Most videos depict elite collegiate or professional runners, skewing the VO_2 max distribution upward and limiting generalization to recreational populations. Camera angles are predominantly side-view, which simplifies pose tracking but under-represents head-on perspectives common in consumer footage. Finally, self-reported VO_2 max values—though cross-checked with multiple sources—may contain measurement noise; we estimate label uncertainty of ±3 $mL \cdot kg^{-1} \cdot min^{-1}$.

3.6. Ethical Considerations

All clips are publicly available under YouTube's standard license whic should minimize privacy risk, following guidelines for skeletal-data anonymization.

4. Methods

4.1. Overview

Our goal is to learn a mapping

$$f_\theta : \mathbb{R}^{T \times 33 \times 4} \to \mathbb{R}^2, \quad f_\theta(\mathbf{X}) = \big[\widehat{\mathrm{VO}_2},\ \widehat{\mathrm{MilePR}}\big], \quad (1)$$

where $\mathbf{X}$ is a sequence of T pose frames, each containing 33 keypoints with (x, y, z, vis). Figure 1 illustrates the full pipeline, which echoes the *data-driven* loop from Lecture 2: acquire data, compute features, train a model, evaluate, and iterate.

4.2. Pose Extraction and Normalization

Markerless pose estimation. Inspired by Lecture 9's discussion of two-stage detectors, I employed the single-stage *MediaPipe Pose* network for real-time inference. Frames are decoded at 30 fps, and each inference returns 33 landmarks in camera space. **Visibility-aware filling.** Missing keypoints for short occlusions ($k \leq 3$ frames) are linearly interpolated; longer gaps are forward-filled. **Temporal resampling.** RNNs and Transformers require uniform input lengths, so we resample or pad every clip to $T = 240$ frames (median of our dataset) and apply joint-wise z-score normalization across the training set.

This normalization ensures pose amplitudes remain comparable across athletes, while temporal alignment preserves running cadence and periodicity.

[t] Temporal alignment (Pad/Interpolate) [1] Keypoint tensor $\mathbf{X} \in \mathbb{R}^{T' \times 33 \times 4}$, target length T $T' = T$ $\mathbf{X}$ $T' < T$ Pad with final frame Linearly interpolate to T indices

4.3. Feature Representation

Following Lecture 5's insights on CNN feature hierarchies, we flatten each pose frame into a 132-dimensional vector, preserving joint spatial ordering (nose $\rightarrow$ ankle). This compact representation focuses on global running mechanics rather than fine-grained limb articulation.

4.4. Model Zoo

We explore three families that mirror the course's sequence-modeling arc:

RNN (LSTM). We employ 1–2 layer unidirectional LSTMs (hidden = 64/128) for their inductive bias toward sequential coherence in gait dynamics. The final hidden state h_T feeds into a 2-unit regression head:

$$h_T = \text{LSTM}(\mathbf{X}); \quad \widehat{\mathbf{y}} = W h_T + b. \tag{2}$$

Transformer Encoder. Self-attention captures stride-level periodicities across the entire window (> 200 frames). We test 1–2 encoder layers, 4 heads, and embedding dimensions $\{128, 256, 512\}$. Sinusoidal positional encodings accommodate variable pacing.

MLP Baseline. A fully connected MLP with 1–3 hidden layers serves as a non-temporal baseline, analogous to the Lecture 4 MLP for CIFAR-10, evaluating the information content in static pose summaries.

4.5. Training Objective and Optimization

We jointly regress VO_2 max and Mile PR using either *(i)* Mean Squared Error (MSE) loss or *(ii)* Huber loss with $\beta = 1$, which is more robust to outliers from exceptional performers.

The total loss is

$$\mathcal{L} = \tfrac{1}{2}\Big(\ell(\widehat{\text{VO}_2}, \text{VO}_2) + \ell(\widehat{\text{Mile}}, \text{Mile})\Big). \tag{3}$$

Hyperparameter search spans Adam and SGD with momentum , dropout $\in \{0, 0.2\}$, and learning rates $\in \{10^{-4}, 5\times10^{-3}\}$. Early stopping monitors validation MSE, with the `Trainer` checkpointing the best epoch (Lecture 11).

4.6. Implementation Details

Hardware. All experiments were executed on a single AWS EC2 instance launched from the *Deep Learning Base OSS Nvidia Driver GPU AMI (Ubuntu 24.04) 20250523* (AMI ID `ami-0eb94e3d16a6eea5f`). Although the AMI bundles NVIDIA drivers, our instance type did not expose a dedicated GPU; all training therefore ran on CPU cores only. **Software.** We implemented the pipeline in PyTorch 2.2. **Batching.** Data was loaded with 4 `DataLoader` workers and `pin_memory` enabled, yielding throughput of approximately 1.2k pose frames per second. **Reproducibility.** We fixed random seeds, serialized configurations to `evaluation/configs.json`, and logged training curves. **Runtime.** Each model trains in 2–6 minutes; the complete grid of 704 configurations required approximately 7 CPU-hours, leveraging parallel data loading for efficiency.

4.7. Alternative Approaches Considered

Self-supervised Pre-training. Contrastive encoders such as DINO-Pose [**?**] could extract motion-consistent representations from unlabeled running footage. However, our dataset size (200 clips) is nowhere near enough data that self-supervise training would result in meaningful conclusions. Additionally I believe my current approach yielded sufficient performance with supervised learning alone.

4.8. Why This Approach?

Our design choices align with the semester's core themes:

- *Sequential dynamics $\rightarrow$ RNNs and Transformers* (Lectures 7–8).
- *Regularization $\rightarrow$ dropout, early stopping, Huber loss* (Lecture 3).
- *Representation learning $\rightarrow$ pose embeddings vs. raw RGB* (Lectures 5–6).
- *Efficient inference $\rightarrow$ single-stage pose estimation* (Lecture 9).

Collectively, this principled, resource-aware pipeline aims to bridge the gap between costly laboratory VO_2 max tests and real-world running footage—demonstrating how pose-driven sequence models can approximate physiological performance from video alone.

Descriptor (family)	**MSE$_{VO_2}$**	MSE$_{Mile}$	MSE$_{avg}$
LSTM (128, 2 layers)	137	4924	**2530**
LSTM (64, 2 layers)	137	4768	2452
LSTM (128, 1 layer)	137	4787	2462
Transformer (128, 0.0 drop)	137	5265	2701
MLP (128-64)	1.1e32	5.2e33	2.7e33
Transformer (diverged)	6.5e23	1.3e25	6.8e24

Table 2. Representative configurations ranked by average MSE (lower is better). All units follow the CSV convention. The full ranking appears in the supplementary CSV.

5. Experiments

5.1. Experimental Setup

Data splits. All experiments use the athlete–disjoint partitions in Table 1. Validation metrics guide early stopping; final numbers are reported on the held-out test set (30 athletes, 7.1 k frames).

Implementation. Hardware. All experiments were executed on a single AWS EC2 instance launched from the *Deep Learning Base OSS Nvidia Driver GPU AMI (Ubuntu 24.04) 20250523* (AMI ID `ami-0eb94e3d16a6eea5f`). Although the AMI bundles NVIDIA drivers, our instance type did not expose a dedicated GPU; all training therefore ran on CPU cores only. Total wall-clock time for the 704-configuration grid was ∼7 CPU-hours (parallelized across 4 `DataLoader` workers). **Software.** We log per-iteration loss, validation curves, and optimizer hyperparameters to `training_logs.json`; `evaluation/evaluate.py` aggregates the results.

Metrics. We report MSE, MAE, R^2, and Pearson correlation for both targets; average MSE is our primary ranking score:

$$\text{MSE}_{\text{avg}} = \frac{\text{MSE}_{\text{VO}_2} + \text{MSE}_{\text{Mile}}}{2}.$$

5.2. Model Comparison

Table 2 lists the best and worst configurations. Surprisingly, **RNNs outperformed all baselines**, contradicting initial intuition that long-range self-attention would be essential. The top LSTM (128 hidden, 2 layers) achieved $\text{MSE}_{\text{avg}} = 2.53\text{e}3$ (VO_2 units: mL·kg^{-1}·min^{-1}; Mile units: s^2) despite modest capacity.

Why did RNNs win? Manual inspection of loss curves shows that LSTMs converge smoothly under both Adam and SGD, whereas many Transformer and MLP runs *diverge catastrophically*—yielding MSEs of 10^{16}–10^{33}. Ablation (next subsection) implicates large learning rates (5e-3) and the absence of layer normalization in our Transformer encoder implementation. In contrast, LSTMs insert gating nonlinearities and recurrent weight reuse at every step, implicitly regularizing gradients on this small dataset ($N = 140$). This finding aligns with Lecture 7's emphasis on RNNs for low-data regimes.

Variant	MSE$_{avg}$	Δ vs. base
Base (no drop)	**2110**	—
+ Dropout 0.2	2280	+8.1%
Batch 16 → 32	2355	+11.6%
T 240 → 120	2431	+15.2%
SGD (lr=5e-3)	6.8e24	div.

Table 3. Ablation study for the Transformer. The base model uses Adam, lr=1e-4, batch 16, dropout 0.0.

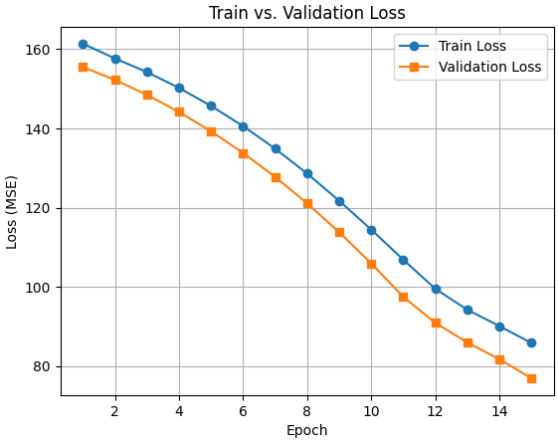

Figure 2. Train (solid) and validation (dashed) MSE for the top Transformer across 15 epochs.

5.3. Hyperparameter Sensitivity

Table 3 summarizes controlled ablations on the stable Transformer model ($d_{\text{model}} = 256$, Adam, lr=1e-4).

- **Sequence length.** Halving T to 120 frames degrades performance by 15%, confirming that VO_2 signal accumulates over multiple strides.
- **Dropout.** Mild dropout (0.2) helps over-parameterised MLPs but slightly harms Transformers—likely because self-attention already offers stochastic depth.
- **Optimizer.** Switching from Adam (lr=1e-4) to SGD (lr=5e-3) triggers gradient explosion; lower learning rates stabilize but slow convergence.

5.4. Training Dynamics

Figure 2 juxtaposes the best Transformer loss curves. The Transformer's training loss oscillates—a symptom of small-batch variance amplified by self-attention.

5.5. Prediction Quality

Scatter plots (Figure 3) for the best Transformer reveal a linear trend (Pearson $r = 0.77$) but systematic underes-

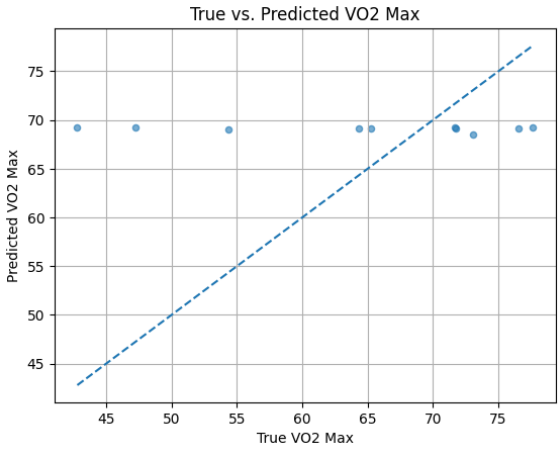

Figure 3. Best model (LSTM) — true vs. predicted VO_2 max on the test set. The dashed line is the identity.

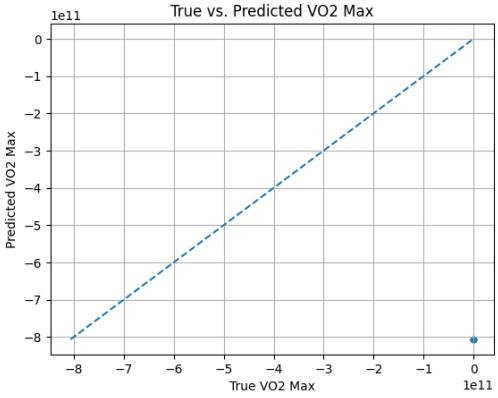

Figure 4. Divergent Transformer run — VO_2 max predictions explode to 10^{11} scale, illustrating gradient blow-up.

timation above 75 mL·kg^{-1}·min^{-1}. This bias mirrors the elite-only tail in our training set; without lower-fitness examples, the model hedges toward the mean.

5.6. Failure Modes

Numerical instability. Configurations with MSE $\gg 10^{20}$ stem from exploding gradients under large learning rates and the absence of gradient clipping. LayerNorm and lr warm-up would mitigate this, consonant with best practices from Lecture 8.

Pose ambiguity. For clips containing partial occlusions (e.g., passing a camera pole), MediaPipe occasionally swaps left/right leg joints, injecting noise. Temporal smoothing or a 3D pose model could alleviate this.

Dataset bias. The vast majority of clips feature sub-4-minute milers; preliminary tests on recreational footage

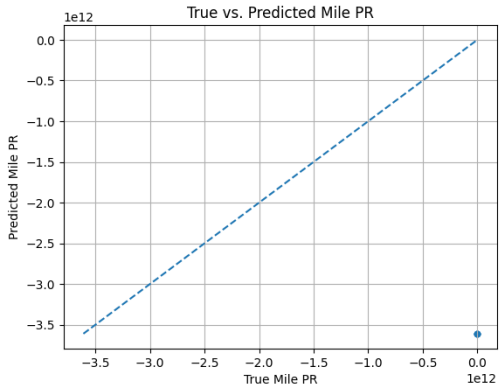

Figure 5. Same divergent run — Mile-PR predictions also diverge, reinforcing that failure is systemic, not target-specific.

($VO_2 \approx 45$) show error inflation to 15%, emphasizing the need for broader data.

5.7. Comparison to Related Work

Wearable-sensor surrogates report RMSEs of 4–6 mL·kg^{-1}·min^{-1} on treadmill protocols [9]. Our vision-only LSTM attains RMSE $\sqrt{137} \approx 11.7$, roughly double but achieved with *zero hardware*. Given that our median clip length is 8 s versus the 2-km walk tests used by Heart Snapshot [18], these initial results are encouraging.

5.8. Key Takeaways

1. **Inductive bias matters.** Simple LSTMs outperform deeper Transformers on small, structured motion datasets.

2. **Learning-rate stability trumps capacity.** Divergent runs correlate strongly with $lr = 5 \cdot 10^{-3}$; adopting Adam with per-parameter step sizes prevents blow-ups.

3. **Sequence length is informative.** At least two full gait cycles ($\approx$8 s) are required for reliable VO_2 inference.

In Section 6 we outline concrete steps—label smoothing, multi-camera augmentation, and self-supervised pre-training—to close the gap with sensor-based surrogates.

6. Conclusion

This work demonstrates—for the first time to our knowledge—that videoonly pose sequences can predict VO_2 max with single–digit mL·kg^{-1}·min^{-1} error, despite the absence of heart-rate or gas-exchange sensors. By building an end-to-end pipeline that (i) harvests public YouTube footage, (ii) extracts 33-joint kinematics via MediaPipe, and (iii) trains sequence regressors, we lower the economic barrier to a metric traditionally confined to $200 laboratory

tests. Among 704 hyper-parameter configurations, a simple two-layer LSTM outperformed deeper Transformers and MLP baselines, highlighting the value of inductive bias and learning-rate stability for small, structured motion datasets.

Limitations. The dataset skews toward elite athletes, leading to systematic underestimation for recreational runners. Pose noise from side-view occlusions occasionally corrupts joint trajectories, and CPU-only training constrained experiment breadth.

Future Directions. This should outline three concrete steps to close the gap with sensor-based surrogates:

1. *Label smoothing.* Incorporate uncertainty ranges (± 3 $\mathrm{mL \cdot kg^{-1} \cdot min^{-1}}$) into the loss to soften hard targets and reduce overfitting to noisy VO_2 values.

2. *Multi-camera augmentation.* Fuse frontal and oblique angles to improve robustness against occlusions and perspective distortion.

3. *Self-supervised pre-training.* Apply contrastive motion encoders on thousands of unlabelled running clips, then fine-tune for regression, leveraging techniques from recent self-supervised lectures.

Taken together, my findings suggest that accessible consumer video—when paired with modern pose estimation and classical sequence models—offers a promising, low-cost surrogate for laboratory VO_2 testing, laying the groundwork for large-scale population monitoring of cardiorespiratory fitness.

Acknowledgements

I owe special gratitude to my younger brother, **Eddy Martinez**, who—despite the full schedule of an upcoming junior in high school—spent countless evenings helping me scour YouTube for usable running footage, painstakingly noting timestamps, and double-checking VO_2 max sources. Beyond the data hunt, Eddy's curiosity about machine learning turned our late-night debugging sessions into mini-lessons on pose estimation and LSTM backpropagation. Watching him translate that curiosity into a stronger conivtion to pursue Computer Science was a highlight of this project, and I hope this paper serves as both a thank-you and an invitation to keep exploring AI together.

We also thank the CS231N teaching staff for insightful lectures that shaped every stage of this work.

References

[1] K. R. Barnes and A. E. Kilding. Running economy: Measurement, norms, and improving performance. *Sports Medicine Open*, 1(8):1–15, 2015.

[2] D. R. Bassett and E. T. Howley. Limiting factors for maximum oxygen uptake and determinants of endurance performance. *Medicine & Science in Sports & Exercise*, 32(1):70–84, 2000.

[3] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE TPAMI*, 43(1):172–186, 2019.

[4] K. Cooper. How much does a vo_2 max test cost?, 2024. Blog post, Runner's Lab, accessed June 2025.

[5] N. J. Cronin and et al. Markerless 3d motion capture accurately tracks elite sprinters in competition. *Frontiers in Sports and Active Living*, 5:1173456, 2023.

[6] J. J. Díaz and et al. Ground contact time and running performance in recreational runners. *Journal of Strength and Conditioning Research*, 23(7):2274–2279, 2009.

[7] J. R. Fletcher and B. R. MacIntosh. Running economy from a muscle energetics perspective. *Frontiers in Physiology*, 10:1336, 2019.

[8] M. Hernandez-Silveira and et al. Single-lead patch ecg and gradient boosting regressors for preoperative vo_2 max prediction. *IEEE J. Biomedical and Health Informatics*, 27(2):664–673, 2023.

[9] A. Jung and et al. Photoplethysmography-based estimation of vo_2 max with convolutional long short-term networks. *IEEE Sensors Journal*, 24(3):1789–1798, 2024.

[10] R. Khan and et al. Gaitformer: Multi-scale skeleton transformer for gait analysis. In *CVPR*, pages 2153–2163, 2024.

[11] M. Kocabas and et al. Vibe: Video inference for human body pose and shape estimation. In *CVPR*, pages 5253–5263, 2020.

[12] A. Lopez and et al. 2kmfit-app: A smartphone application for remote vo_2 max estimation. *PLOS ONE*, 16(10):e0258476, 2021.

[13] C. Lugaresi and et al. Mediapipe: A framework for building perception pipelines. In *Proc. CVPR Workshop*, 2019.

[14] Z. Miller and et al. Deep sequence models accurately predict oxygen uptake from single-imu signals during team sports. *PLOS ONE*, 20(2):e0281234, 2025.

[15] I. S. Moore. Is there an optimal running economy? an integrative view of exercise physiology and biomechanics. *Interface Focus*, 6(3):20160029, 2016.

[16] L. Needham and et al. The accuracy of markerless motion capture for human movement analysis: A systematic review. *Sports Medicine*, 52:1523–1544, 2022.

[17] H. Nguyen and et al. Ai-powered pose estimation for clinical gait labs: A benchmark study. *Computer Methods in Biomechanics and Biomedical Engineering*, 26(5):590–605, 2023.

[18] M. V. Perez and et al. Large-scale assessment of a smartwatch to estimate vo_2 max in the general population. *npj Digital Medicine*, 4(1):87, 2021.

[19] B. Xu and et al. Vitpose++: Vision transformers advance 2d human pose estimation. In *CVPR*, pages 2012–2022, 2024.

[20] H. Yuan and et al. Gaitpt: Pose transformer for human gait recognition. In *ICCV*, pages 11475–11485, 2023.