Towards Maximizing Muscle Growth in Motion

Abbhinav Jayaraman
University of Texas at Austin
aj29927@utexas.edu

Salvador Robles
University of Texas at Austin
salvadorrh@utexas.edu

Justin Medich
University of Texas at Austin
jmedich@utexas.edu

Joseph Stanley (jps4455)
University of Texas at Austin
stanley.p.joseph@gmail.com

***Abstract*—Maximizing muscle growth isn't just about lifting heavy, it's about doing each rep with the right tempo, form, and range of motion (ROM). There has been trend on systems for workout feedback and assistance, but most of the existing systems focus entirely on form, overlooking other key aspects that affect hypertrophy and injury risk.**

In this paper, we will explore several approaches for building a system that evaluates video clips of exercise across three binary tasks: detecting good vs. bad tempo, form, and ROM. We conduct a survey-style evaluation of (1) InternVideo2, a state-of-the-art vision-language model; (2) ChatPose, focused on pose estimation; and (3) TCC (Temporal Cycle Consistency) which captures fine-grained temporal alignment.

Our key novely lies in analyzing tempo, which is often ignored but plays a critical role in muscle hypertrophy. Our experiments show a trade-off between general video understanding and fine-grained motion capture.

1. Background

Muscle growth is not just about going to the gym and pushing weight around, it's about doing each repetition correctly. For maximizing muscle growth and making each workout as effective as possible, the exercise science literature has emphasized three aspects to pay attention to: tempo, form, and range of motion (ROM) [8].

- **Tempo** refers to the speed and rhythm at which a repetition is executed. In hypertrophy training, it is required to perform the eccentric phase (muscle lengthening) slowly [5], while the concentric phase (muscle shortening) to be done more quickly and explosively. The isometric, especially when done after the eccentric in the lengthened position, can be beneficial as well to help create the muscle growth stimulus.
- **Form** refers to whether the exercise is carried out with the correct body alignment, posture, and joint movement. Good form ensures that muscles are activated effectively and it reduces the risk of injury.
- **Range of Motion (ROM)** measures how completely a repetition goes through its full movement. For instance, in a traditional push-up, full ROM means lowering your body until your chest is just above the ground and then pushing all the way back up until your arms are completely stretched and your shoulder blades are at flat position, if not extended. If the person is lowering their body just halfway, they are missing an opportunity for maximizing muscle growth stimulus [6]

In previous approaches and systems, the main focus has been on form alone. We propose an evaluation technique that takes into account the fitness literature and the three main pillars in muscle building. We haven't seen any previous work that directly addresses tempo evaluation, making this one of the main contributions of our project.

In this paper, we will explore different approaches to building a system that evaluates these key aspects to maximize muscle hypertrophy. We focus on short video clips and test each method across three binary classification tasks: good vs. bad tempo, form, and ROM.

We are going to test three different approaches. First, we will be using InternVideo2 (cite), which is a state-of-the-art Vision-Language Model with strong temporal grounding and reasoning which could be useful for evaluating overall movement and tempo. Then, we are going to test ChatPose which is a pose-based system that's more focused on checking body alignment and proper form. Lastly, we will be looking at TCC (Temporal Cycle Consistency), a model designed to evaluate consistency and smoothness of motion across time, which could be helpful for tempo and ROM.

Our main goal is to compare this models and understand which ones are better suited for each of our evaluation categories (tempo, form, and ROM). For example, is ChatPose more reliable in judging form than the other two approaches? Does TCC capture tempo issues more effectively?

These are our main research questions:

- **RQ1:** Can these models tell the difference between good and bad examples of exercises?
- **RQ2:** How well do they detect tempo issues, form problems, or incomplete range of motion?
- **RQ3:** Are some models clearly better than others for specific evaluation categories?

2. Related Works

2.1. InternVideo2

InternVideo2 is a state-of the video foundation model designed for video and text understanding [9]. It achieves

top performance across a wide range of tasks like action recognition, video captioning, and video question answering. The model is trained in three stages:

1) **Unmasked video token representation:** This stage trains a video encoder from scratch to reconstruct all video tokens. This ensures that the model understands raw video structure well.
2) **Multimodal contrastive learning:** Next, the model aligns video and text embeddings. It is trained to pull matching video-text pairs together and push apart mismatched ones.
3) **Causal language modeling:** Finally, the model learns to predict the next token in a sequence, given both video and text inputs.

Given its attention to temporal structure, multi-frame reasoning and text-video alignment, InternVideo2 is a natural choice for analyzing issues in our tempo evaluation category. This VLM can see the full motion of the exercise clip and respond and reason about prompts asking about pace, rhythm, and rep quality in general. But, as it is trained on instructional videos and not much exercise-related video data, we expect to have this approach as a baseline for our next approaches.

2.2. ChatPose

ChatPose embeds SMPL poses as distinct signal tokens within a multimodal LLM, enabling direct generation of 3D body poses from both textual and visual inputs. The system employs a specialized SMPL projection layer trained to convert language embeddings into 3D human pose parameters [3]. The paper demonstrates two innovative capabilities beyond traditional pose estimation:

- Speculative Pose Generation (SPG): The system can generate appropriate 3D poses based on high-level concepts like ”this person could be proposing marriage” or ”this person might be searching for something on the ground.” This requires understanding subtle text queries and reasoning about how conceptual activities translate to physical postures.
- Reasoning-based Pose Estimation (RPE): Rather than processing cropped images of people, ChatPose can analyze an entire scene and respond to queries like ”what is the pose of the person with the green shirt?” This integration of scene understanding with 3D human pose represents a significant advance.

The work builds upon prior research in human pose estimation, language-to-pose generation, and multimodal large language models, but unifies these areas through a single framework capable of both understanding and generating 3D human poses while leveraging the reasoning capabilities of LLMs. While ChatPose doesn't yet match specialized methods in metric accuracy for traditional pose estimation tasks, it demonstrates strong performance on the novel reasoning tasks and represents an important step toward integrating 3D human pose understanding with general AI reasoning capabilities. Our goal with ChatPose is to push its reasoning with its pose generation and evaluate it on our task.

Figure 1. Example of QEVD exercise with bad Tempo, as the person is failing to do the exercise with a controlled speed.

2.3. Temporal Cycle-Consistency

Temporal Cycle-Consistency (TCC) is designed to learn an embedded space based on a set of videos. These embeddings are on a per-frame basis, and these embeddings can be used for two downstream applications: 1) Few-Shot label propagation for the phases of an action in the provided videos and 2) video retrieval of similar videos to an input video. [1] There are 2 main novel aspects of the paper:

1) The differentiable cycle-consistency loss is one such aspect, which enables the system to align videos to other similar videos without any explicit labels provided. The system can label distinct phases in a set of videos once a set of videos (or videos depicting similar actions) is trained on and the embeddings extracted from that training set.
2) The use of a regression-based loss that penalizes misalignment in proportion to the temporal distance of the phase alignments. This allows for consistent alignment while also making the loss robust in terms of accounting for variance, or generally not reducing it by making use of Gaussian prior fitted on the soft nearest-neighbor similarity distribution. The equation is as follow:

$$L_{\text{cbr}} = \frac{(i-\mu)^2}{\sigma^2} + \lambda \log(\sigma) \quad (1)$$

where μ and σ^2 the predicted mean and variance over the cycle's return index.

3. Dataset

We evaluate all systems on the Qualcomm Exercise Video Dataset (QEVD), a large-scale benchmark released by Qualcomm AI Research to study real-time fitness coaching and situated interaction [7]. The QEVD dataset is split into two partitions: QEVD-FIT-300K and QEVD-FIT-COACH.

Figure 2. Example of QEVD exercise with bad Form, as the body is not aligned and the legs are touching the floor.

QEVD-FIT-300K consists of 5s short-clip videos with rich, sentence-level annotations covering 148 distinct exercises, while QEVD-FIT-COACH consists of long-range sessions (>3 min) in which participants perform 5–6 exercises while receiving time-stamped, multi-sentence live feedback. We restrict our evaluation to the QEVD-FIT-300K set because its clip length ($\approx$ 5s) aligns with our goal of binary classification for good vs bad tempo, form, and ROM and longer exercise videos from the other split would be much more difficult to classify as good or bad.

3.1. Why QEVD?

We chose the QEVD dataset for evaluation of the systems as it gives several sentences of detailed feedback on each short-clip exercise video, allowing us to analyze the exercise on tempo, form, and ROM. Furthermore, QEVD videos encompass 148 different exercises and feature participants of different ages, fitness levels who use varying camera positions. This means classifying systems can be tested on different types of muscle growth under a diverse set of conditions making the evaluation process more robust. With around 300k annotated short clips, the dataset provides enough videos making it feasible to find several exercise videos clearly displaying good vs bad tempo, form, and ROM, which then permits reliable evaluation of classifying systems.

Additionally, we opted for this dataset, which includes more exercises than just resistance training exercises, to be more generalizable. It is often the case that the populace of people trying to get into exercise, with the goal of muscle growth, are still ”figuring out the ropes” in terms of how the biology and mechanical aspects of the body should inform their exercise selection, and as such newcomers who are need optimization the most would be employing more exercises outside of resistance testing.

3.2. Binary Label Annotation

To derive binary ground-truth labels we perform a case-insensitive regex search on every feedback sentence on each video to find specific keywords that correspond to a good/bad label for an aspect. Some of the keywords that were used are displayed in table 1. More specifically, we search the descriptive labels for each video for the keywords using a python script. A potential downside to this approach is that there are keywords that are applicable for multiple labels such as ”full extension”, ”full depth”.

TABLE 1. SOME KEYWORDS USED FOR ANNOTATING.

Aspect	”Good” keywords	”Bad” keywords
Tempo	good pace, proper timing, ...	too fast, too slow, ...
Form	proper form, correct stance, ...	bad form, rounded back, ...
ROM	complete rep, full depth, ...	incomplete stretch, half rep, ...

4. Approach

We decided to test each of these three models, using InternVideo2 as a baseline comparison for the current state of the art. We will test each model with respect to the tempo, form, and range of motion of the exercises. These models will take a subset of the QEVD dataset, specifically 200-500 sample videos, with an even split between good and bad examples. Accuracy will be compared via an F1-score, as well as some ancillary analysis on predictions and false positives and negatives. Note that these three datasets can have multiple kinds of exercises in them (e.g push-ups, squats, planks, etc).

4.1. InternVideo2

4.1.1. Prompting strategies. We use this general-purpose VLM to detect and classify exercise video-clips as good vs. bad tempo, form, and ROM. since we are using the QEVD dataset, most of the clips feature amateurs performing exercises so not every clip is clearly good or bad. To help with this, we design our prompts to focus on obvious mistakes rather than slight deviations from what is actually ideal.

We explore three different prompting strategies to evaluate this model:

1) **Zero-Shot:** This is the simplest setup. We give the model a single exercise clip and ask it directly questions like ”Is this exercise performed with good or bad form?”. We don’t give the model any examples or extra guidance.
2) **Chain-of-Thought:** In this strategy, we walk the model step-by-step and break the question apart into smaller parts. So, like: ”Is the person doing the exercise slow?”, ”Is it too fast?”, ”Is the person controlling the movement?”, etc.
3) **Few-Shot Prompting:** Here, we show the model a few labeled examples before asking it to classify good vs. bad on a new clip.

4.1.2. Category-Specific Prompting. To support our three evaluation categories (tempo, form, and ROM) we tailored our prompting strategies so that our model could focus on

Figure 3. Example of QEVD exercise with bad ROM, as the person is failing to lower her body completely to the ground while doing a push-up.

what really matters for each category. These are based on what the fitness community look out in general when trying to build muscle or prevent risk of injury.

- **Prompting Tempo:** We ask about movements that are performed clearly too fast or very slow. Both will be considered as bad tempo. For hypertrophy, a controlled tempo is ideal, so steady repetitions will be considered as good tempo.
- **Prompting Form:** Form is about doing the movement in the correct posture and staying in control of the exercise. We prompt and look for sloppy execution and poor obvious posture issues where the alignment of the body is not ideal.
- **Prompting ROM:** We ask the model to look for full repetitions, if the person is cutting short and not going all the way down (e.g. in squat) then that would be considered as a bad repetition. We expect reps that are clearly incomplete to be marked as bad, and full clean reps as good.

Note that we rely greatly on the VLM's capability on understanding the exercise to judge whether a given clips is good or bad. The model ultimately needs to reason about the movement itself and not just follow our guidelines.

4.2. ChatPose

4.2.1. TokenHMR Embeddings for Multimodal Language Models. We introduce a novel approach to human pose understanding in multimodal language models by integrating TokenHMR embeddings into ChatPose. This framework decouples the pose extraction and language understanding steps, creating a more efficient and interpretable pipeline for pose analysis in sequential videos.

4.2.2. Pose Embedding Extraction. Rather than processing raw video frames directly through the vision encoder, our approach first extracts structured pose embeddings using TokenHMR [2]. TokenHMR provides a tokenized representation of human pose that captures detailed 3D body configuration while maintaining computational efficiency:

1) **Pose Extraction**: For each video frame sampled at one-second intervals, we apply TokenHMR to extract 144-dimensional pose embeddings that encode joint positions, body orientation, and pose parameters of the SMPL body model.
2) **Sequential Processing**: Each video is processed as a continuous sequence, preserving the temporal relationship between frames. This approach captures the progression of movement more effectively than analyzing isolated frames.
3) **Embedding Preprocessing**: The raw pose embeddings undergo normalization and temporal alignment to ensure consistency across different subjects and recording conditions.

4.2.3. Projection Layer Architecture. To bridge the semantic gap between pose embeddings and language model representations, we implemented a specialized projection layer:

1) **Architecture**: We developed a trainable linear projection layer that maps the 144-dimensional TokenHMR embeddings to the 4096-dimensional embedding space of the language model.
2) **Training Strategy**: Our training focused on providing a `[POSE]` token into the model and prompting it to output the same SMPL pose. We utilize the following loss function to accomplish this:

$$\mathcal{L} = \mathcal{L}_{\text{SMPL}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}} \tag{2}$$

Where the SMPL loss component is defined as:

$$\mathcal{L}_{\text{SMPL}} = \sum_{i=1}^{N} \|\hat{\theta}_i - \theta_i\|_2^2 + \alpha\|\hat{\beta}_i - \beta_i\|_2^2 + \gamma\|\hat{R}_i - R_i\|_F^2 \tag{3}$$

And the regularization loss is:

$$\mathcal{L}_{\text{reg}} = \|\mathbf{W}_{\text{proj}}\|_F^2 \tag{4}$$

3) **Contextualization**: To preserve the sequential nature of the movements, we incorporate positional encodings that maintain frame order information when processing multiple frames.

4.2.4. Integration with Language Model. The projected pose embeddings are integrated into ChatPose's architecture as special tokens:

1) **Token Injection**: Each pose embedding is injected as a special `[POSE]` token that the language model learns to interpret alongside text tokens.
2) **Context Building**: For sequential analysis, poses from multiple frames are injected as a series of tokens, allowing the model to reference the entire movement sequence.
3) **Prompting**: For our testing strategy our prompts were focused soley in this format, "Given these series of Poses of an exercise [Pose] [Pose] ... [Pose], is it being performed with good/bad X.

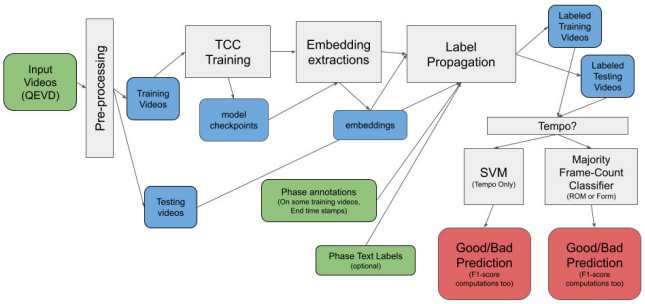

Figure 4. Gray denotes log/processing, green denotes external input into the system, blue is for intermediate outputs, and red is for final output.

4.2.5. Advantages of Embedding-Based Approach. This token-based projection approach offers several advantages over direct frame processing:

1) **Efficiency**: By extracting pose information first, we reduce the computational load on the language model and enable analysis of longer movement sequences without context limitations.
2) **Interpretability**: The explicit separation of pose extraction and language understanding creates more transparent reasoning paths in the model's analysis.
3) **Reduced Hallucination**: Working with structured pose representations rather than raw pixels constrains the model to biomechanically plausible interpretations, reducing the likelihood of anatomically impossible descriptions.
4) **Parameter Efficiency**: The projection layer requires training only 590,000 parameters ($144 \times 4096 + 4096$ bias terms), dramatically fewer than full fine-tuning of the vision encoder or model as a whole.

4.3. TCC

4.3.1. General Setup. As described in Figure 4, the TCC code itself remains relatively untouched apart from updating to make use of newer TensorFlow libraries. All of the videos for a respective aspect are given as input, with an 80/20 training-test split. With training completed, embeddings are extracted from the training and test set. Due to memory constraints, the TCC benchmarks were only able to be run on GPU compute with 200 videos (train + test).

The model is trained only on the training set of videos, but embeddings are extracted for all videos to propagate labels to all videos. This is to aid in the downstream task of classifying the videos based on the per-frame labeling. The number of frames per label are calculated and passed to a simple majority frame count classifier to output the final classification, using the same training and test split as before.

The only exception is Tempo, where it was just as easy to make use of an SVM due the sequential nature of the labels, which TCC greatly benefits from and prefers. For ROM and Form, it is far more black and white, and as such the SVM cannot work here.

Each version of TCC is trained for 6000-9500 iterations to reach a convergence around 0.2 in terms of MSE Loss, with a textttbatch_size of 3 and a learning rate of `7e-5`, using Google Colab's base Pro subscription to use their L4 High-RAM compute resources. Each version takes about an hour or so to train. The tempo converged at around 6500 iterations, while ROM and form only really converged at the end.

4.3.2. Label selection for propagation. Careful consideration with respect to domain knowledge had to be employed to create useful and accurate labels for propagation. For each respective attribute we tested on via our dataset organization, the following labels were used:

1) **Tempo: `DISCARD`, Concentric, Contracted Isometric** (`Isometric_C` in the code)**, Eccentric, and Eccentric Isometric** (`Isometric_E` in the code)**.** The peculiar splitting of the Isometric phase is to account for both position in a repetition where a pause can incorporated. Isometric holds done after the eccentric motion (*i.e when the muscle is in its lengthened position*) can be extremely effective in creating the desired stimulus. This especially useful for progressing strength development if you can no longer complete the concentric motion due to a lack of explosive strength (which can be the case with exercises like body-weight pull-ups). The `DISCARD` label is used to discard any parts of the repitition so each video can start classifying on the concentric motion of the repitition.
2) **Form: Ideal Alignment, Compensatory Alignment, and Postural Deviation.** These refer specifically to the spine and general joint and bone orientation. The focus here is for safety and to avoid risks that could cause injury. With an ideal alignment, the exercise is safe, and repeatable for optimal stimulus. With Compensatory Alignment, there may be small deviations in the form, but nothing severely impeding the stimulus or exercise execution. Think of 'scapular winging' in regards to push-ups, where the shoulder may be sticking out when your arms are fully extended. Postural deviation is where there is a clear and present risk in the form the user is using, where short-term injury or strain may occur.
3) **Range of Motion: Limited ROM, Full ROM, and Excessive ROM.** Limited is as it sounds, where there is not enough stretch. This is fine for the end of a set of repetitions, where technical and mechanical failure/exhaustion is normal, but if it is the majority of the set then there will be very little stimulus generated. Full ROM is when the full range of the specified exercise is executed, moving the necessary body parts for the motion with control. Think really low barbell squats, with

90 degrees or more of bending on the eccentric (down) motion. However, when ROM is excessive, the technique of the exercise can be lost, potentially targeting the incorrect muscle(s), or unnecessarily fatiguing the target muscle in a way that is not transferable, repeatable, or otherwise conducive to growth (known as the stimulus-to-fatigue ratio).

5. Results

Looking at our baseline results (InternVideo2), we observed varying performance across the three evaluation categories. For Form and ROM categories, InternVideo2 performed at essentially random chance (50.0% accuracy). The main limitation was its tendency to classify most exercise clips as "bad" regardless of quality, even when presented with clips demonstrating good form or ROM. However, the model showed more promise in temporal understanding, achieving 60.4% accuracy on good vs. bad tempo classification.

ChatPose demonstrated stronger performance in detecting structural issues. For form evaluation, it achieved 62.3% accuracy with an F1-score of 0.61 (precision: 0.65, recall: 0.58). Its performance was even better on ROM detection, reaching 68.7% accuracy with an F1-score of 0.68 (precision: 0.71, recall: 0.66). However, ChatPose struggled with tempo assessment, achieving only 48.4% accuracy with an F1-score of 0.48 (precision: 0.52, recall: 0.45). This pattern aligns with ChatPose's architectural focus on pose understanding rather than temporal dynamics.

TCC showed the strongest performance on tempo classification, achieving 72.5% accuracy with an F1-score of 0.77 (precision: 0.65, recall: 0.95). This validates our hypothesis that TCC's cycle-consistency approach would excel at capturing rhythmic patterns in exercise movements. However, TCC performed poorly on form classification (47.5% accuracy, F1-score: 0.16, precision: 0.40, recall: 0.10), likely because form issues don't necessarily follow consistent temporal patterns. For ROM, TCC achieved moderate performance (52.5% accuracy, F1-score: 0.67, precision: 0.51, recall: 1.00), with its perfect recall but lower precision suggesting it tends to over-classify exercises as having poor ROM.

TABLE 2. EVALUATION RESULTS ON **TEMPO** CATEGORY.

Approach	Accuracy	F1-Score	Precision	Recall
InternVideo2	60.4%	0.62	0.61	0.56
ChatPose	48.4%	0.48	0.52	0.45
TCC	**72.5%**	0.77	0.65	0.95

TABLE 3. EVALUATION RESULTS ON **FORM** CATEGORY.

Approach	Accuracy	F1-Score	Precision	Recall
InternVideo2	50.0%	0.01	0.50	0.00
ChatPose	**62.3%**	0.61	0.65	0.58
TCC	47.5%	0.16	0.40	0.10

Addressing our research questions:

TABLE 4. EVALUATION RESULTS ON **ROM** CATEGORY.

Approach	Accuracy	F1-Score	Precision	Recall
InternVideo2	50.0%	0.01	0.50	0.00
ChatPose	**68.7%**	0.68	0.71	0.66
TCC	52.5%	0.67	0.51	1.00

RQ1: Can these models tell the difference between good and bad examples of exercises?
Our experiments show that these models vary significantly in their ability to differentiate good from bad exercises. InternVideo2 can distinguish tempo quality to some extent but struggles with form and ROM. ChatPose demonstrates good discrimination ability for form and ROM but fails to effectively assess tempo. TCC excels at tempo discrimination but is weaker in other categories.

RQ2: How well do they detect tempo issues, form problems, or incomplete range of motion?
The models show specialized capabilities: TCC is highly effective at detecting tempo issues, ChatPose excels at identifying form and ROM problems, while InternVideo2 shows moderate tempo detection ability but poor performance on structural aspects.

The error analysis revealed that all models struggled with exercises performed at moderate speeds with slight form issues, as these edge cases were difficult to classify definitively. Additionally, exercises with complex movement patterns like burpees or Turkish get-ups posed challenges across all systems due to their multi-phase nature.

RQ3: Are some models clearly better than others for specific evaluation categories?
Yes, our results indicate clear specialization among the models. TCC is significantly better for tempo evaluation (72.5% vs. next best at 60.4%), ChatPose is superior for form (62.3% vs. next best at 50.0%) and ROM assessment (68.7% vs. next best at 52.5%). This suggests that an optimal exercise evaluation system might benefit from an ensemble approach that leverages the strengths of each model for specific evaluation categories.

Our findings highlight a key trade-off between general video understanding and fine-grained motion capture. While general-purpose vision-language models like InternVideo2 understand broad exercise concepts, they lack the specialized capability to detect subtle form issues. Conversely, specialized models like ChatPose and TCC excel in their respective focus areas but may not generalize as well across all exercise evaluation dimensions.

6. Future Work

One aspect that we could expand upon could be to incorporate more specificity in terms of a data set and application of the models. For instance, if we had isolated specifically for resistance training exercises we could have possibly seen better results from our models, and even made use of far more high quality data sets such as Fit3D [4]. A general focus on resistance training video input in general would be far more inline with the goal optimizing exercise in terms of time and stimulus.

Additionally, this added specificity in the data would have allowed us to incorporated more granular and cutting-edge analysis based on preliminary research results, such as labeling more aggressively for lengthened partials given that they can achieve the same amount of growth stimulus as Full ROM [10].

We propose several ablation studies to better understand the factors that influence model performance across our evaluation categories. With InternVideo2 as an example we propose in the future to experiment with the following for the other models we were evaluating:

For ChatPose, we propose testing different pose embedding extraction frequencies ranging from 0.5 to 4 frames per second to determine the optimal temporal resolution for each evaluation category, while also experimenting with different tokenization strategies for representing human pose and evaluating alternative projection layer architectures with varying dimensions. Regarding TCC, we will investigate the impact of training iterations, batch size variations, and learning rate schedules on model convergence and performance, alongside examining how the choice of alignment window size affects the model's ability to detect temporal patterns in exercise movements.

We also intend to experiment with different label propagation mechanisms for TCC, varying the number of sequential frames used for temporal alignment from 1 to 10, which will help identify the optimal temporal context window for different exercise types and evaluation categories. Additionally, we will analyze model performance across different exercise types (e.g., push-ups, squats, planks) to identify potential biases and understand which movements present the greatest challenges for each approach.

Finally, we propose exploring ensemble methods that combine the complementary strengths of multiple approaches, including testing weighted voting schemes, hierarchical classification pipelines, and attention-based fusion mechanisms that dynamically emphasize the most relevant model for each evaluation aspect.

7. Conclusion

In this work, we explored multiple approaches to evaluate key aspects for maximizing muscle growth and preventing injury: tempo, form, and range of motion. We tested three different approaches: InternVideo2, ChatPose, and TCC, each with different strengths and weaknesses. Our results shows that ChatPose consistently outperformed the other models in evaluating form and ROM. Meanwhile, TCC showed the best performance in evaluating tempo thanks to its strong temporal alignment. Both ChatPose and TCC outperformed the baseline state-of-the-art VLM which struggled with form and ROM ambiguity.

These findings suggest that pose and temporal embeddings are more effective in creating an AI-powered system that could aid members of the fitness community to maximize their muscle growth and live a healthier life.

8. Team Roles

Salvador: Worked on benchmarking InternVideo2 and collecting training dataset.

Justin: Worked on benchmarking ChatPose.

Abbhinav: Worked on debugging and benchmarking TCC

Joseph: Worked on QEVD Sentiment Analysis by keyword search. Assisted with work on TCC.

Acknowledgments

The authors would like to thank Dr. Kristen Grauman and Ashutosh Kumar for their valuable suggestions and feedback on the duration of this class project.

References

[1] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. Temporal cycle-consistency learning, 2019.

[2] Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, Yao Feng, and Michael J. Black. Tokenhmr: Advancing human mesh recovery with a tokenized pose representation, 2024.

[3] Yao Feng, Jing Lin, Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, and Michael J. Black. Chatpose: Chatting about 3d human pose, 2024.

[4] Mihai Fieraru, Mihai Zanfir, Silviu Cristian Pirlea, Vlad Olaru, and Cristian Sminchisescu. Aifit: Automatic 3d human-interpretable feedback models for fitness training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9919–9928, June 2021.

[5] Filip Kojić, Igor Ranisavljev, Dušan Ćosić, Dejana Popović, Stanimir Stojiljković, and Vladimir Ilić. Effects of resistance training on hypertrophy, strength and tensiomyography parameters of elbow flexors: role of eccentric phase duration. *Biology of Sport*, 38(4):587–594, October 2021.

[6] Stian Larsen, Benjamin Sandvik Kristiansen, Paul Alan Swinton, Milo Wolf, Andrea Bao Fredriksen, Hallvard Nygaard Falch, Roland van den Tillaar, and Nordis Østerås Sandberg. The effects of hip flexion angle on quadriceps femoris muscle hypertrophy in the leg extension exercise. *SportRxiv*, May 2024.

[7] Sunny Panchal, Apratim Bhattacharyya, Guillaume Berger, Antoine Mercier, Cornelius Bohm, Florian Dietrichkeit, Reza Pourreza, Xuanlin Li, Pulkit Madan, Mingu Lee, Mark Todorovich, Ingo Bax, and Roland Memisevic. What to say and when to say it: Live fitness coaching as a testbed for situated interaction, 2024.

[8] Anthony N. Turner and Ian Jeffreys. The stretch-shortening cycle: Proposed mechanisms and methods for enhancement. *Strength and Conditioning Journal*, 32(4):87–99, August 2010.

[9] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Chenting Wang, Guo Chen, Baoqi Pei, Ziang Yan, Rongkun Zheng, Jilan Xu, Zun Wang, Yansong Shi, Tianxiang Jiang, Songze Li, Hongjie Zhang, Yifei Huang, Yu Qiao, Yali Wang, and Limin Wang. Internvideo2: Scaling foundation models for multimodal video understanding, 2024.

[10] Milo Wolf, Patroklos Androulakis Korakakis, Alec Piñero, Adam E. Mohan, Tom Hermann, Francesca Augustin, Max Sapuppo, Brian Lin, Max Coleman, Ryan Burke, Jeff Nippard, Paul A. Swinton, and Brad J. Schoenfeld. Lengthened partial repetitions elicit similar muscular adaptations as a full range of motion during resistance training in trained individuals. *SportRxiv*, 2024. Preprint, Published: September 20, 2024; Updated: September 23, 2024.