WHAT SHOULD A STREAMING COACH REMEMBER? MOTION EVIDENCE ACROSS REPETITIONS AND TRANSITIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

A streaming coach is a vision-language model that watches workout video as it arrives and responds proactively. After it asks for a deeper squat, judging the next repetition requires the depth that prompted the correction; once the person moves on to lunges, that squat-specific guidance must no longer apply. It therefore needs movement evidence that stays available across repetitions and whose relevance changes at exercise transitions. Existing coaching models are trained on short single-exercise segments, and processing a longer stream does not itself keep earlier movement available for later feedback. We ask what a streaming coach should remember and propose StreamingCoach, which represents movement in the terms a coach reasons in and retains it independently of the feedback it generates: Form Tokens encode each chunk in anatomical joint coordinates grounded in person-centered visual features, and a gated Motion State is updated from them at every chunk, whether or not feedback is generated. On QEVD workouts processed as continuous streams without context reset at exercise boundaries, segment-trained coaches break down beyond a single exercise, and StreamingCoach improves reference-based feedback content and response timing over a streaming backbone fine-tuned under the same protocol, with both components contributing; on separately adapted probes it also recognizes the current exercise and counts repetitions substantially better. These results expose coaching demands that exercise-segment evaluation does not capture and turn what a streaming coach should remember into a testable question for streaming vision-language models.

1 INTRODUCTION

Someone exercising alone may not notice that their squats become shallower with each repetition. A streaming coach, a vision-language model that watches video as it arrives and responds proactively rather than when asked, should catch this trend in time to guide the next repetition. Identifying the exercise and evaluating the current pose are not enough: the coach must compare the current repetition with earlier ones, judge whether depth improved after a correction, and stop applying squat-specific advice once the person moves on to lunges. The first two decisions draw on earlier motion across repetitions; the third asks, at a transition, whether that motion still applies. Existing framings address neither. Live fitness coaching has been posed as deciding what to say and when to say it (Panchal et al., 2024), and streaming vision-language models can now watch video and respond as it unfolds (Chen et al., 2024; Qian et al., 2025; Xu et al., 2026), but what such a model retains of earlier motion, and whether it serves these decisions, has not been examined. *We ask what a streaming coach should remember across repetitions and transitions.*

Existing coaching models address this problem ***one exercise at a time***. The coaching model introduced with QEVD (Panchal et al., 2024) generates asynchronous counts and corrections, but its final training stage uses 30-second single-exercise segments, and training across a whole workout was left for future work. BioCoach (Ji et al., 2026) grounds streaming feedback in 3D kinematics and biomechanical analysis, but its context is the exercise in view, recomputed from a sliding window at each step. Neither carries motion beyond that window, so neither can compare a repetition with earlier ones or check whether a correction took hold; run on a whole workout, they rarely respond after the first exercise, and when they do, the feedback can address the wrong one (Fig. 1). GuideMe (Liu

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

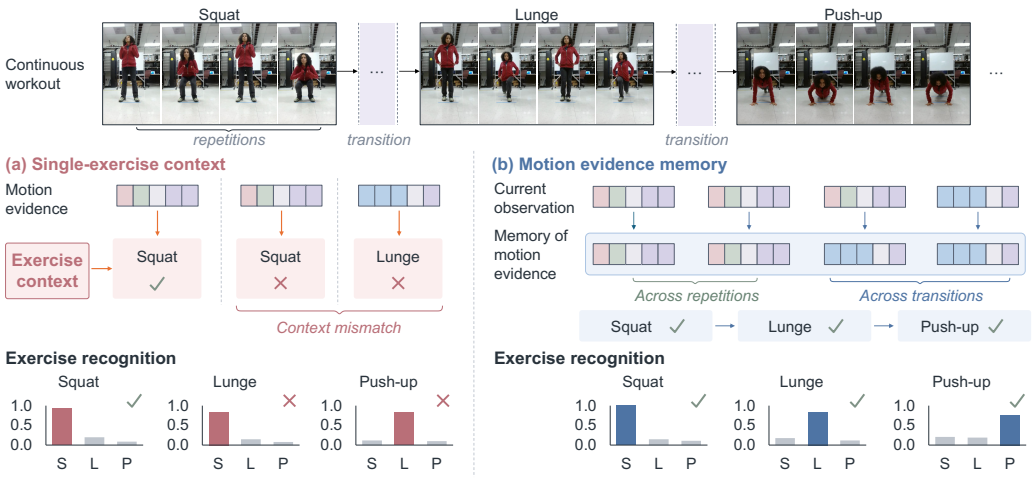

Figure 1: **What should a streaming coach remember?** Later repetitions call for comparison with earlier ones, a correction for checking whether it took effect, and a transition for setting squat-specific advice aside. **(a)** BioCoach, given the exercise as context, keeps applying it and misreads the exercise once the workout moves on. **(b)** StreamingCoach retains Form Tokens in a Motion State across repetitions and transitions and recognizes the exercise throughout without being told. Bars show exercise-recognition scores on this workout.

et al., 2026a) confirms that error detection and correction are the hard parts of streaming guidance, fitness included, but does not ask whether the difficulty lies in perceiving the current movement or in retaining the past. A coach for a whole workout needs a memory that outlives the window, which general-purpose streaming vision-language models are built to provide.

Such models can take the whole workout and respond based on their running context. *Processing, however, is not retention.* StreamingVLM (Xu et al., 2026) keeps a short window of visual tokens and a longer one of text; movement that leaves the window survives only through the text stream, yet a response such as "go deeper" records the correction, not the depth, and repetitions that draw no response leave no explicit record. Other assistants keep response-independent text records (Zhang et al., 2025c; Gong et al., 2026), compressed visual or latent memories (Qian et al., 2025; Zhang et al., 2026; Chen et al., 2026; Qu et al., 2026), or day-long memories (Jiang et al., 2026), but none represents movement as joints in comparable anatomical coordinates, and what they keep is chosen by novelty, query relevance, or words: a squat slightly shallower than the last is not novel, not queried, and not mentioned. Whether any retained representation preserves such small, ordered changes across repetitions, or lets go of exercise-specific evidence at a transition, is untested. The question is the evidence retained for coaching decisions, not the amount of video processed.

In light of this, we propose **StreamingCoach**, which bridges both gaps with one novel principle: represent movement in the terms a coach reasons in and retain it independently of the feedback it generates. Two novel components realize it. (i) ***Form Tokens*** make the representation comparable: a motion adapter expresses each incoming chunk as anatomically constrained joint angles and their derivatives, grouped by body region and grounded in the person's appearance, so that every repetition is placed in the same anatomical coordinates. Unlike motion tokens built for motion generation or generic video understanding, they describe how the current movement is executed in units where a slightly shallower squat is a measurable change rather than a few pixels. (ii) ***Motion State*** makes retention independent of response: a gated recurrent state integrates Form Tokens at every chunk in which valid motion is observed, including chunks that produce no feedback, and exposes the result to the decoder as State Tokens. It is never written from the coach's own feedback or from a query, so what it retains cannot depend on its responses or on what is asked, while the backbone's visual window and text context remain in place alongside it. We deliberately give the state no hand-designed repetition counter or exercise-indexed store: what it retains is learned rather than prescribed, and whether it retains the right evidence is left to the evaluation.

We evaluate on QEVD under the standard exercise-segment protocol and on full workouts of about 3.5 minutes spanning several exercises, processed as continuous streams without resetting context at exercise boundaries. Compared with segment-trained coaching models and a streaming backbone fine-tuned under the same protocol, StreamingCoach achieves superior performance in feedback

content and response timing and, on separately adapted probes, in exercise recognition and repetition counting, and both components contribute. In summary, the contributions of this work include:

◇ We pose *a novel formulation* of streaming coaching around what the coach should remember: retaining earlier movement across repetitions and setting it aside across transitions.

◇ We propose **StreamingCoach**, which has two novel components: Form Tokens, which make every repetition comparable in the same anatomical coordinates, and a Motion State that is updated from observed motion at every chunk, whether or not feedback is generated.

◇ We establish *a continuous-workout protocol* on QEVD that processes whole workouts without resetting context, where segment-trained coaches break down beyond one exercise, a fine-tuned streaming backbone recognizes the current exercise poorly, and StreamingCoach outperforms both.

2 Related Work

Fitness Assessment and Language-Based Coaching. Early systems assess execution against a reference: AIFit reconstructs 3D motion, segments repetitions, and compares each with an instructor to generate corrections (Fieraru et al., 2021), and PoseFix generates corrective text between two poses (Delmas et al., 2023). Recent work moves to vision-language models: FormCoach pairs user and reference clips for form correction (Zuo et al., 2025), CoachMe generates reference-based instructions for sport skills (Yeh et al., 2025), and ExpertAF produces expert actionable feedback from video and pose (Ashutosh et al., 2025). QEVD frames coaching as situated interaction with timestamped feedback, and its model learns when to observe and when to respond, trained on single-exercise segments (Panchal et al., 2024); BioCoach adds skeletal kinematics, cycle analysis, and biomechanical constraints (Ji et al., 2026). EgoExo-Fitness provides multi-exercise sequences with quality annotations (Li et al., 2024b); GuideMe benchmarks intervention in streaming task guidance, fitness included, and finds error correction the hardest part (Liu et al., 2026a); and PushupBench shows that VLMs count repetitions poorly (Li et al., 2026). These works assess or coach within an exercise; none studies what a coach must retain across repetitions and exercise transitions.

Streaming Video Understanding and Proactive Interaction. Streaming video models address two questions: when to respond and what to retain. For timing, VideoLLM-online decides per frame whether to respond (Chen et al., 2024), MMDuet and MMDuet2 learn timing through a video-text duet format and multi-turn reinforcement learning (Wang et al., 2025b; 2026), Dispider separates perception, decision, and reaction (Qian et al., 2025), StreamBridge and LiveStar gate responses of offline video LLMs (Wang et al., 2025a; Yang et al., 2025), LiveCC learns online commentary from timestamped transcripts (Chen et al., 2025), and ThinkStream reasons incrementally before responding (Liu et al., 2026b). For retention, StreamingVLM keeps a short visual window and longer text context (Xu et al., 2026); VideoStreaming, Flash-VStream, and StreamForest compress visual memory (Qian et al., 2024; Zhang et al., 2025b; Zeng et al., 2025); ReKV retrieves cached KV pairs on demand (Di et al., 2025); StreamChat keeps hierarchical memory across rounds (Xiong et al., 2025); TimeChat-Online and QueryStream prune tokens by redundancy or query relevance (Yao et al., 2025; Zhang et al., 2026); ProAssist, GROVE, and Memento write text records for long horizons (Zhang et al., 2025c; Gong et al., 2026; Jiang et al., 2026); and StreamTTT and LatentStream keep latent state through fast weights or query-conditioned evolution (Chen et al., 2026; Qu et al., 2026). These memories select by novelty, query relevance, or verbalization and target question answering or dialogue. We build on this line for timing; our state is written from observed movement at every chunk, independent of responses and queries, and is evaluated on continuous coaching.

Motion-Language Representations and Anatomical Grounding. Motion-language models connect movement with text: MotionGPT quantizes motion into a discrete vocabulary (Jiang et al., 2023), and LLaMo aligns motion features with language for instruction following (Li et al., 2025); MotionBench shows that video VLMs remain weak at fine-grained motion (Hong et al., 2025). These represent motion for generation or description, not for comparing one person's repetitions over time. Anatomically grounded reconstruction provides that basis: SKEL parameterizes body and skeleton with biomechanically meaningful joints (Keller et al., 2023), and HSMR estimates them from monocular images (Xia et al., 2025). Form Tokens build on these to express each repetition in the same anatomical coordinates, fused with person-centered visual evidence, and the Motion State carries them forward across repetitions and exercise changes.

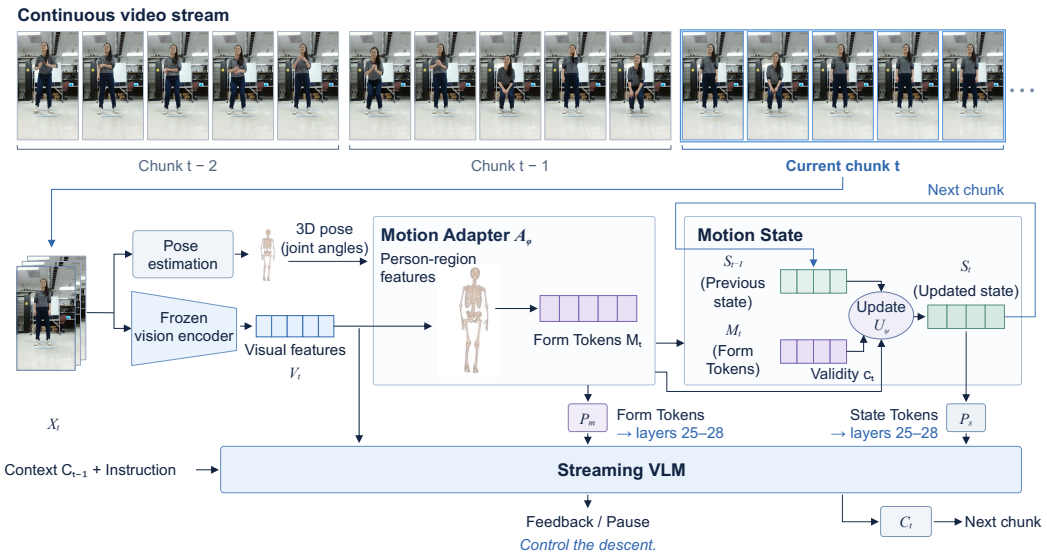

Figure 2: **StreamingCoach.** A motion adapter encodes each incoming chunk into Form Tokens, which update a gated recurrent Motion State. Both condition feedback generation alongside the backbone's visual window and text history. The state is written from observation only, is carried across exercise boundaries, and continues to update in chunks that produce no feedback.

3 METHOD

3.1 PROBLEM FORMULATION

A streaming coach observes a workout in video chunks and decides whether and how to respond after each chunk. Let X_t denote the video chunk at step t, and let P_t contain its pose estimates, person boxes, and validity masks; t indexes chunks rather than repetitions or exercises. In the continuous-workout setting, an instruction lists the planned exercises in order but does not identify the current exercise, which the coach must infer from the observations. We omit the instruction from the notation. Given observations through step t and prior responses, the coach generates a token sequence Y_t containing either feedback or the backbone's pause token, autoregressively from $p_\theta(Y_t \mid X_{\le t}, P_{\le t}, Y_{<t})$. In addition to the backbone's visual window and text history, the coach maintains a Motion State $S_t \in \mathbb{R}^{K_s \times d}$, a deterministic function of the observation history. This state is reset only at the start of a stream and persists across chunks, including exercise transitions.

3.2 OVERVIEW

StreamingCoach realizes the principle of Sec. 1 with two components: a motion adapter A_ϕ that produces Form Tokens, and a gated recurrent update U_ψ that maintains the Motion State (Fig. 2). Formally,

$$(M_t, c_t) = A_\phi(X_t, P_t), \qquad S_t = U_\psi(S_{t-1}, M_t, c_t), \tag{1}$$

where $M_t \in \mathbb{R}^{K_m \times d}$ encodes how movement is performed in chunk t, and $c_t \in [0, 1]$ is the fraction of valid pose samples. Both enter the decoder through learned projections, M_t as Form Tokens and S_t as State Tokens (Sec. 3.5). Generated feedback enters the text history, but only observations update S_t, so movement in chunks without feedback can still inform later responses.

3.3 FORM TOKENS: ENCODING CURRENT EXECUTION

Form Tokens encode each video chunk in a shared set of anatomical coordinates, grounded in person-centered visual features (Fig. 3a). This provides a consistent basis for comparing repetitions without segmenting them or assigning one token to each.

Anatomical Dynamics. Joint angles of a biomechanical skeleton (Keller et al., 2023) are estimated in each frame (Xia et al., 2025), so that repetitions, within and across people, are expressed in the

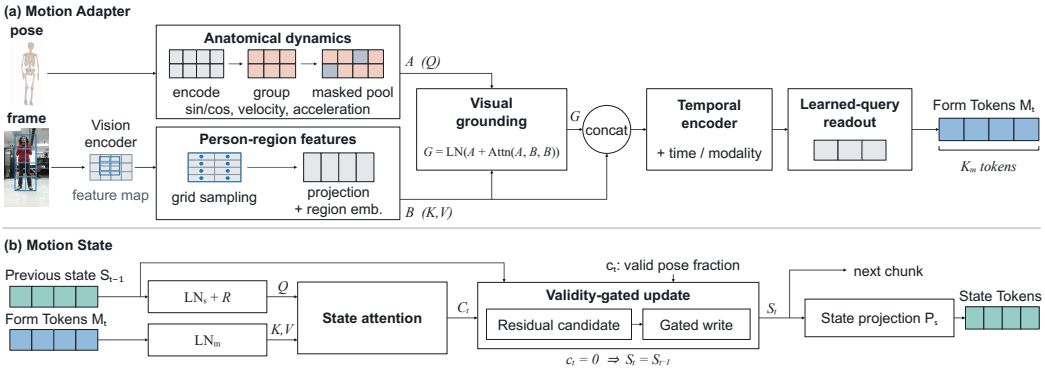

Figure 3: **Form Tokens and Motion State.** **(a)** The motion adapter encodes anatomical configuration and dynamics (sine/cosine, velocity, acceleration), pools them by body group, grounds them in person-region visual features through attention, and compresses the result into K_m Form Tokens with learned queries. **(b)** The Motion State attends from its slots to the incoming Form Tokens, forms a residual candidate, and writes it through a validity-modulated gate; with $c_t = 0$ the state is preserved exactly. The updated state is projected into State Tokens and carried to the next chunk.

same joint coordinates. For pose sample i and coordinate j, we encode configuration, velocity, and acceleration:

$$f_{i,j} = [\sin q_{i,j}, \cos q_{i,j}, \dot{q}_{i,j}, \ddot{q}_{i,j}],$$
$$\dot{q}_{i,j} = r\ \mathrm{wrap}(q_{i,j} - q_{i-1,j}), \qquad \ddot{q}_{i,j} = r(\dot{q}_{i,j} - \dot{q}_{i-1,j}), \tag{2}$$

where r is the pose sampling rate and $\mathrm{wrap}(a) = \mathrm{atan2}(\sin a, \cos a)$; the periodic encoding avoids discontinuities at angle boundaries. Derivatives are backward differences between adjacent source samples and are set to zero whenever a sample in their stencil is invalid; the two preceding source samples are carried across chunk boundaries, so only the first samples of a stream lack derivatives. Each coordinate is embedded, and the embeddings are averaged within each anatomical group $\mathcal{G}_g$:

$$e_{i,j} = \mathrm{LN}(\mathrm{GELU}(W_f f_{i,j} + b_f + u_j)), \qquad a_{i,g} = \frac{1}{|\mathcal{G}_g|} \sum_{j \in \mathcal{G}_g} e_{i,j} + v_g, \tag{3}$$

where u_j and v_g are learned coordinate and group embeddings, and LN denotes layer normalization. Embedding each sample before temporal pooling lets velocity and acceleration contribute to the pooled tokens. Validity-aware pooling masks invalid samples and aggregates the remaining samples within each visual time bin k, yielding one anatomical token per group in $A_{t,k}$.

Visual Grounding. We sample features on a spatial grid within the person box from an intermediate vision-encoder map, then project them and add region embeddings to obtain visual tokens $B_{t,k}$. The anatomical tokens attend to these visual tokens:

$$G_{t,k} = \mathrm{LN}\left(A_{t,k} + \mathrm{Attn}(A_{t,k}, B_{t,k}, B_{t,k})\right). \tag{4}$$

Modality embeddings distinguish the anatomical and visual streams. The grounded tokens $G_{t,k}$ and visual tokens $B_{t,k}$ are concatenated; sinusoidal temporal embeddings are added, the tokens from all bins of the chunk are encoded by a Transformer, and K_m learned queries compress the result into M_t.

3.4 Motion State: Retaining Evidence Across Repetitions and Transitions

The Motion State is a fixed set of K_s slots updated once per chunk; at the start of a stream it is initialized to a learned state S_0 (Fig. 3b). The update uses neither repetition segmentation nor an exercise-boundary detector, and nothing is cleared at a transition: evidence from earlier chunks can influence later decisions throughout the stream, and whether the state retains useful history without producing stale feedback is left to the evaluation. Each update forms an attention-based candidate from the previous state and incoming tokens, then applies a validity-modulated gate. With $\bar{S}_{t-1} = \mathrm{LN}_s(S_{t-1})$, $\bar{M}_t = \mathrm{LN}_m(M_t)$, and slot-role embeddings R,

$$C_t = \mathrm{Attn}(\bar{S}_{t-1} + R, \bar{M}_t, \bar{M}_t), \qquad D_t = \tanh\left(\mathrm{MLP}([\bar{S}_{t-1}; C_t])\right),$$
$$\widetilde{S}_t = \mathrm{LN}_u(\bar{S}_{t-1} + D_t), \qquad g_t = c_t\, \sigma\big(W_g[S_{t-1}; \widetilde{S}_t; c_t] + b_g\big), \tag{5}$$

$$S_t = \begin{cases} S_{t-1}, & c_t = 0, \\ \mathrm{LN}_u\left(\bar{S}_{t-1} + g_t \odot (\widetilde{S}_t - \bar{S}_{t-1})\right), & c_t > 0. \end{cases} \tag{6}$$

For $c_t > 0$, the gate blends the normalized previous state $\bar{S}_{t-1}$ with the candidate $\widetilde{S}_t$ elementwise before normalization; c_t is broadcast across slots and channels. When $c_t = 0$, the previous state is preserved exactly. During training, write dropout replaces c_t with $\hat{c}_t = z_t c_t$ inside U_ψ only, with $z_t \in \{0, 1\}$ drawn independently per chunk; the Form Tokens of a dropped chunk still reach the decoder. This decouples the number of updates from the number of elapsed chunks, reducing the usefulness of the update count as a timing cue. The updated state is projected into State Tokens for the current response and carried to the next chunk, even when no feedback is emitted.

3.5 Decoder Interface

Form and State Tokens occupy reserved placeholder positions in the prompt. Before the last four of the 28 decoder layers, which are the layers trained in the second stage, we add at each reserved position a residual equal to the projected Form or State Token minus the corresponding placeholder input embedding, which cancels the placeholder embedding's direct term in the residual stream; each projection includes a learned per-slot offset.

This injects motion information without altering the backbone's visual-token pipeline. Before inserting the current State Tokens, we remove the previous set from the key–value cache, leaving only the latest State Tokens available for direct attention. For each chunk, the decoder generates feedback followed by the backbone's native pause token $\langle\mathrm{pause}\rangle$, or the pause token alone; $\langle\mathrm{pause}\rangle$ ends every turn and hands control back to the stream, and a turn consisting of $\langle\mathrm{pause}\rangle$ alone is silent.A pause-only turn withholds feedback while observation and state updates continue. Responses enter the text history, and the backbone manages its visual and text cache.

3.6 Training

Training proceeds in two stages. We first fine-tune the streaming backbone with $\mathcal{L}_{\mathrm{LM}}$ alone and without the motion inputs, updating all language-model parameters while the vision encoder stays frozen; the motion modules are then added to this backbone and trained with the full objective below.

Language Objective. Timestamped annotations supervise response timing and content. Each supervised turn targets annotated feedback followed by the pause token, or the pause token alone; both sequences end with the end-of-turn token. We minimize cross-entropy across all target tokens:

$$\mathcal{L}_{\mathrm{LM}} = -\frac{1}{N} \sum_{t \in \mathcal{T}} \sum_{j=1}^{L_t^*} \log p_\theta(y_{t,j}^* \mid X_{\le t}, P_{\le t}, Y_{<t}^*, y_{t,<j}^*), \tag{7}$$

where $\mathcal{T}$ is the set of supervised turns and N is the total number of target tokens. The first target token distinguishes feedback from a pause; subsequent tokens supervise content and termination. In the second stage, chunks within the tolerance window of an annotated feedback time, other than the chunk the feedback is assigned to, receive no language target, so that near-correct timing is not penalized. During training the model conditions on the annotated response history, whereas at inference it conditions on its own generated history.

Motion Supervision. The language objective can be partly satisfied by language priors and response history alone, without using the motion inputs. In the second stage, we add two auxiliary heads, discarded at inference, to encourage the representations to retain motion information. For chunk t, $s_t \in \mathbb{R}^{4J}$ collects, for each of the J coordinates, the means of $\sin q$ and $\cos q$ and the mean and standard deviation of $\log(1 + |\dot{q}|)$ over valid samples. A head h_m predicts the current chunk's s_t from pooled Form Tokens M_t. A head h_s predicts the same statistics from the previous state S_{t-1}, encouraging the state to retain information predictive of subsequent movement:

$$\mathcal{L}_{\mathrm{pose}} = \frac{1}{|\mathcal{T}_m| + |\mathcal{T}_s|} \Big(\sum_{t \in \mathcal{T}_m} \mathrm{MSE}\left(h_m(M_t), s_t\right) + \sum_{t \in \mathcal{T}_s} \mathrm{MSE}\left(h_s(S_{t-1}), s_t\right) \Big), \tag{8}$$

where $\mathcal{T}_m$ contains chunks that receive gradients (the supervised window and up to eight preceding chunks), and $\mathcal{T}_s \subseteq \mathcal{T}_m$ contains those whose predecessor is also in $\mathcal{T}_m$. The second-stage objective

is $\mathcal{L} = \mathcal{L}_{\mathrm{LM}} + \lambda_{\mathrm{pose}}\mathcal{L}_{\mathrm{pose}}$, with $\lambda_{\mathrm{pose}} = 0.1$. Both heads regularize the representations toward motion content; neither prescribes which history the state keeps.

Windows and Optimization. In the second stage, we train in a single pass over a randomly selected subset of supervision windows. Context preceding the window is processed without language gradients, while gradients through the state reach up to eight preceding chunks. The adapter, state module, projections, and language adaptation are trained jointly on top of the fine-tuned backbone; the vision encoder and output head remain frozen. For streaming coaching, we update the upper four decoder layers and final normalization. For the recognition and form-correction transfer experiments, we use low-rank adapters on all linear layers of the language model.

4 EXPERIMENTS

We evaluate on QEVD (Panchal et al., 2024) in the continuous-workout setting of Sec. 3.1 and under the original exercise-segment protocol, with two separately adapted probes of exercise identity and repetition progress, and on Fit3D (Fieraru et al., 2021) for form correction of individual repetitions.

4.1 SETUP

Datasets. QEVD-FIT-COACH (Panchal et al., 2024) provides complete workouts covering 25 exercises with timestamped feedback: 148 training workouts (874 segments, 4,408 feedback events) and 74 benchmark workouts of six consecutive exercises each (444 segments, of which 398, with reference feedback, are scored; 1,993 events), with no shared videos. For Fit3D (Fieraru et al., 2021), we use the publicly released subset of the FormCoach annotations (Zuo et al., 2025): 470 user–reference repetition pairs over 10 exercises with expert feedback that specifies a correction or states that none is needed. StreamingCoach is adapted on 263 pairs, validated on 76, and tested on 89 from a held-out subject; the remaining 42 pairs are excluded because their reference repetition is performed by a held-out subject.

Settings. In the continuous-workout setting, each benchmark workout is one stream: the instruction lists the planned exercises in order without indicating which is underway, and neither the language context nor the Motion State is reset at exercise boundaries, so the coach must keep responding, notice when the activity changes, and relate its guidance to the exercise being performed. All systems receive the same instructions and process the full stream; responses are aligned with the reference exercise intervals for scoring. In the exercise-segment setting, each of the 398 segments starts from a fresh context, and the instruction names the exercise.

Probes. The *exercise probe* shows the central 6 s of each segment as an isolated clip with a fresh context and asks for the exercise name in an open question, without the planned-exercise list; answers are mapped to the 25 classes by name normalization and synonym matching. The *count probe* is posed at each of the 132 moments at which the reference coach reports a count: the model reports the completed repetitions, scored by MAE against the reference announcements. StreamingCoach, its ablations, and StreamingVLM are adapted to each probe with low-rank adapters on the training split (Sec. 3.6); FitCoach and BioCoach take no free-form queries and are probed through their own feedback (Appendix B), marked † and not directly comparable.

Models and metrics. StreamingCoach uses the Qwen2.5-VL-7B streaming backbone of StreamingVLM (Xu et al., 2026) with HSMR joint angles (Xia et al., 2025). Baselines are QEVD-fine-tuned video-language models, StreamingVLM fine-tuned on QEVD (Appendix A), and the coaching systems FitCoach[1] and BioCoach (Ji et al., 2026). Content is scored on temporally aligned prediction–reference pairs by METEOR (Banerjee & Lavie, 2005), ROUGE-L (Lin, 2004), BERTScore (Zhang et al., 2020), and, in the segment setting, an LLM judge (LLM-Acc., 1–5, judged by LLaMA-3-70B-Instruct); timing is measured by the temporal F-score (T-F), the harmonic mean of temporal precision and recall, where a prediction matches a reference feedback within a 3-s window. Higher values are better, except for count MAE.

Table 1: Continuous-workout evaluation on QEVD without context reset at exercise boundaries. Probe columns are separately adapted tasks; † marks values probed through each method's own feedback, which precludes direct comparison. *w/o Motion State* removes the Motion State; *w/o direct Form Tokens* keeps the recurrent update but withholds Form Tokens from the decoder. Best in bold, second best underlined.

	Feedback content			Timing	Probes (adapted)	
Method	METEOR↑	ROUGE-L↑	BERTScore↑	T-F↑	Exercise Probe Acc.↑	Count MAE↓
FitCoach (Panchal et al., 2024)	0.106	0.056	0.850	0.066	23.8†	8.27†
BioCoach (Ji et al., 2026)	0.109	0.061	0.853	0.068	20.1†	7.61†
StreamingVLM (Xu et al., 2026)	0.122	0.102	0.858	0.606	25.6	3.29
w/o ***Motion State***	0.126	0.104	0.859	<u>0.609</u>	<u>82.2</u>	1.52
w/o ***direct Form Tokens***	<u>0.130</u>	<u>0.106</u>	<u>0.863</u>	0.608	81.7	<u>1.21</u>
StreamingCoach	**0.132**	**0.107**	**0.864**	**0.613**	**82.4**	**1.08**

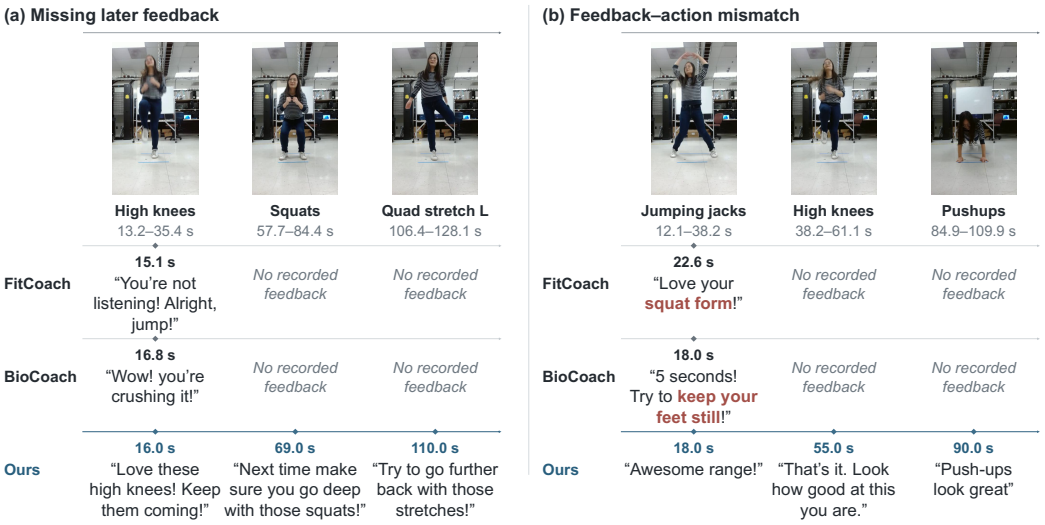

Figure 4: **Continuous-workout coaching, qualitative.** **(a)** FitCoach and BioCoach respond early, then produce no feedback in later segments. **(b)** During jumping jacks, FitCoach praises squat form and BioCoach tells the participant to keep their feet still. In these examples StreamingCoach keeps responding and its feedback matches the ongoing exercise. Time ranges mark exercise intervals; timestamps mark displayed responses.

4.2 Continuous-Workout Coaching

Processing a whole workout without resets exposes failures that the segment protocol cannot show (Tab. 1). FitCoach and BioCoach, trained on single-exercise segments and designed to be told which exercise is being performed, reach T-F of only 0.066 and 0.068; after the first exercise, they respond in only 0% and 0% of segments, and 89% of their responses name another exercise; Fig. 4 shows both failure modes, feedback that stops after the early segments and feedback for the wrong exercise. StreamingVLM fine-tuned on QEVD keeps responding throughout and is the strongest baseline; StreamingCoach scores higher on every content metric and in timing (METEOR 0.132 vs. 0.122, T-F 0.613 vs. 0.606; higher in 50 and 63 of the 74 workouts), and in the examples of Fig. 4, its feedback matches the ongoing exercise, although it is never told which one is underway. The probes measure what the representation carries, not how the coach uses it while coaching: under the same protocol, StreamingCoach recognizes the exercise in 82.4% of segments compared to 25.6%, and has a MAE of 1.08 compared to 3.29.

The ablations remove one motion input at a time. Removing the Motion State while keeping Form Tokens (w/o Motion State) lowers every content metric and raises count MAE from 1.08 to 1.52; withholding direct Form Tokens while keeping the state (w/o direct Form Tokens) changes them less (METEOR 0.130, MAE 1.21). Probe accuracy is essentially unchanged in both (82.2% and

[1]Called Stream-VLM in Panchal et al. (2024); we use the name of its public release to avoid confusion with StreamingVLM.

Table 2: Exercise-segment evaluation on QEVD. In contrast to Tab. 1, each single-exercise segment is coached independently, from a fresh context and with the exercise named in the instruction.

	Feedback content				Timing
Method	METEOR↑	ROUGE-L↑	BERTScore↑	LLM-Acc.↑	T-F↑
Socratic-LLaMA-2-7B (Zeng et al., 2022)	0.094	0.071	0.860	2.17	0.500
Video-ChatGPT (Maaz et al., 2024)	0.108	0.093	0.863	2.33	0.500
LLaMA-VID (Li et al., 2024a)	0.106	0.090	0.860	2.30	0.500
FitCoach (Panchal et al., 2024)	0.127	0.112	0.863	2.45	0.560
BioCoach (Ji et al., 2026)	0.129	**0.122**	**0.864**	2.56	0.544
StreamingCoach	**0.131**	0.121	**0.864**	**2.57**	**0.597**

Table 3: Form-correction feedback on Fit3D with FormCoach annotations.

Method	METEOR↑	ROUGE-L↑	BERTScore↑
VideoLLaMA3-7B (Zhang et al., 2025a)	0.153	0.151	0.860
Qwen2.5-VL-7B (Bai et al., 2025)	0.155	0.162	0.870
GPT-4.1 (OpenAI, 2025)	**0.217**	0.193	0.873
StreamingCoach	0.215	**0.211**	**0.881**

81.7%), so it is reached without the state, and the gain over StreamingVLM there is attributable to the anatomical motion input. The state thus improves counting and content even when Form Tokens are present; whether this comes from retained history or from its encoding of the current chunk cannot be settled by ablation alone.

4.3 Exercise-Segment Coaching and Form Correction

Under the original protocol (Tab. 2; zero-shot models, all below the fine-tuned ones, are in Appendix Table 4), where each segment covers a single exercise of about 30 s, named in the instruction, retained movement has less to add: StreamingCoach's content scores are on par with BioCoach (higher METEOR and LLM-Acc., lower ROUGE-L, and equal BERTScore), and the clear margin is in timing (T-F 0.597 vs. 0.560). Fit3D tests whether Form Tokens adapt to form correction, where feedback must state how a specific repetition deviates from correct execution (Tab. 3). StreamingCoach is adapted on 263 pairs. StreamingCoach outperforms open-source models of the same scale on all three metrics and is comparable to GPT-4.1 (higher ROUGE-L and BERTScore, slightly lower METEOR), so under this protocol, anatomical motion evidence supports fine-grained form feedback beyond the QEVD setting.

5 Conclusion

Coaching across repetitions requires relating current form to earlier movement; coaching across transitions requires updating which earlier evidence remains relevant. We studied these demands with **StreamingCoach**, which combines anatomically grounded Form Tokens with a Motion State updated from observed movement at every chunk, including those that produce no feedback. On continuous QEVD workouts, StreamingCoach obtains higher reference-based feedback and timing scores than a streaming backbone fine-tuned under the same protocol. Qualitative examples illustrate missing or exercise-mismatched feedback from segment-trained coaches, and separately adapted probes show stronger exercise recognition and repetition counting. The ablations indicate that both motion pathways contribute to the measured outcomes; whether the state uses observations beyond the current chunk remains to be isolated by intervention. Coaching thus offers a concrete test of what movement evidence a streaming AI assistant retains for later decisions.

Limitations. Our protocol covers workouts without repeated sets and extends directly to them; the probes characterize the representation, and interventions on state writes can test its use during coaching; video-grounded rating would complement the reference-based metrics; and neither component depends on the backbone's tokenization, so both apply to other streaming models.

Reproducibility Statement

Sec. 4.1 specifies the datasets, splits, evaluation settings, and metrics. Sec. 3 describes the model, while the appendix provides implementation and training details, probe protocols, the LLM-judge prompt, and the temporal matching rule. Code, configuration files, evaluation scripts, and trained model weights will be released upon publication.

AI Use Statement

Large language models (Claude and ChatGPT) served as writing and discussion aids: editing the manuscript, surveying related work, and providing feedback on the framing. All methods, experiments, and analyses were designed, implemented, and verified by the authors, who made every research and reporting decision and take full responsibility for the paper's content.

Ethics Statement

We used previously released fitness-video datasets and annotations and recruited no new participants. Exercise videos may contain identifiable individuals; their use and any redistribution must follow the datasets' access terms. Feedback generated by StreamingCoach may be inaccurate, and performance may vary with pose-estimation quality, camera view, and participant characteristics. The system is a research prototype and should not be relied upon as medical advice or as a substitute for qualified supervision where incorrect guidance could cause harm.

References

Kumar Ashutosh, Tushar Nagarajan, Georgios Pavlakos, Kris Kitani, and Kristen Grauman. ExpertAF: Expert actionable feedback from video. In *CVPR*, 2025.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 2005.

Joya Chen, Zhaoyang Lv, Shiwei Wu, Kevin Qinghong Lin, Chenan Song, Difei Gao, Jia-Wei Liu, Ziteng Gao, Dongxing Mao, and Mike Zheng Shou. Videollm-online: Online video large language model for streaming video. In *CVPR*, 2024.

Joya Chen, Ziyun Zeng, Yiqi Lin, Wei Li, Zejun Ma, and Mike Zheng Shou. Live: Learning video llm with streaming speech transcription at scale. In *CVPR*, 2025.

Joya Chen, Zeyun Zhong, and Mike Zheng Shou. StreamTTT: Reconciling real-time perception and long-term memory in streaming vlms. *arXiv preprint arXiv:2608.13416*, 2026.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *NeurIPS*, 2023.

Ginger Delmas, Philippe Weinzaepfel, Francesc Moreno-Noguer, and Grégory Rogez. Posefix: correcting 3d human poses with natural language. In *ICCV*, 2023.

Shangzhe Di, Zhelun Yu, Guanghao Zhang, Haoyuan Li, Hao Cheng, Bolin Li, Wanggui He, Fangxun Shu, and Hao Jiang. Streaming video question-answering with in-context video kv-cache retrieval. In *ICLR*, 2025.

Mihai Fieraru, Mihai Zanfir, Silviu Cristian Pirlea, Vlad Olaru, and Cristian Sminchisescu. AIFit: Automatic 3D human-interpretable feedback models for fitness training. In *CVPR*, 2021.

Sitong Gong, Caixin Kang, Tianyu Yan, Guo Chen, Bo Zheng, Kaipeng Zhang, Yunzhi Zhuge, Xiang Ruan, Huchuan Lu, and Yifei Huang. GROVE: Growing and reasoning over temporally stratified memory from streaming video experience. *arXiv preprint arXiv:2608.02392*, 2026.

Wenyi Hong, Yean Cheng, Zhuoyi Yang, Weihan Wang, Lefan Wang, Xiaotao Gu, Shiyu Huang, Yuxiao Dong, and Jie Tang. Motionbench: Benchmarking and improving fine-grained video motion understanding for vision language models. In *CVPR*, 2025.

Yuyang Ji, Yixuan Shen, Shengjie Zhu, Yu Kong, and Feng Liu. From 3D pose to prose: Biomechanics-grounded vision–language coaching. In *CVPR*, 2026.

Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *NeurIPS*, 2023.

Hongxiang Jiang, Zengrui Ge, Guo Chen, Qixiong Wang, Jile Jiao, Xuetao Feng, Yuan Wang, and Yan Wang. Memento: Toward an all-day proactive assistant for ultra-long streaming video. In *ICLR*, 2026.

Marilyn Keller, Keenon Werling, Soyong Shin, Scott Delp, Sergi Pujades, C Karen Liu, and Michael J Black. From skin to skeleton: Towards biomechanically accurate 3d digital humans. *TOG*, 2023.

Lei Li, Sen Jia, Jianhao Wang, Zhongyu Jiang, Feng Zhou, Ju Dai, Tianfang Zhang, Zongkai Wu, and Jenq-Neng Hwang. Human motion instruction tuning. In *CVPR*, 2025.

Shengzhi Li, Jiarun Chen, Karun Sharma, Jiaqi Su, and Shichao Pei. Pushupbench: Your vlm is not good at counting pushups. *arXiv preprint arXiv:2604.23407*, 2026.

Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In *ECCV*, 2024a.

Yuan-Ming Li, Wei-Jin Huang, An-Lan Wang, Ling-An Zeng, Jing-Ke Meng, and Wei-Shi Zheng. Egoexo-fitness: Towards egocentric and exocentric full-body action understanding. In *ECCV*, 2024b.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *EMNLP*, 2024.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 2004.

Fang Liu, Jinpeng Chen, Ke Xu, Yuhao Liu, Huankang Guan, Xudong Lu, Yang Bo, Gerhard Hancke, Rui Liu, and Rynson WH Lau. GuideMe: Multi-domain task guidance and intervention in streaming video. In *ECCV*, 2026a.

Zikang Liu, Longteng Guo, Handong Li, Ru Zhen, Xingjian He, Ruyi Ji, Xiaoming Ren, Yanhao Zhang, Haonan Lu, and Jing Liu. Thinking in streaming video. In *ECCV*, 2026b.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *ACL*, 2024.

OpenAI. Introducing GPT-4.1 in the API. `https://openai.com/index/gpt-4-1/`, 2025.

Sunny Panchal, Apratim Bhattacharyya, Guillaume Berger, Antoine Mercier, Cornelius Böhm, Florian Dietrichkeit, Reza Pourreza, Xuanlin Li, Pulkit Madan, Mingu Lee, et al. What to say and when to say it: Live fitness coaching as a testbed for situated interaction. *NeurIPS*, 2024.

Rui Qian, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Shuangrui Ding, Dahua Lin, and Jiaqi Wang. Streaming long video understanding with large language models. In *NeurIPS*, 2024.

Rui Qian, Shuangrui Ding, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. Dispider: Enabling video llms with active real-time interaction via disentangled perception, decision, and reaction. In *CVPR*, 2025.

Hongyu Qu, Guangming Yao, Ling Xing, Xiaobin Hu, Rongxing Ding, Guibin Zhang, Fan Zhang, Yi Yuan, Xiangbo Shu, and Shuicheng Yan. Beyond retrieval: Progressive latent memory evolution for streaming video understanding. *arXiv preprint arXiv:2609.04131*, 2026.

Haibo Wang, Bo Feng, Zhengfeng Lai, Mingze Xu, Shiyu Li, Weifeng Ge, Afshin Dehghan, Meng Cao, and Ping Huang. StreamBridge: Turning your offline video large language model into a proactive streaming assistant. In *NeurIPS*, 2025a.

Yueqian Wang, Xiaojun Meng, Yuxuan Wang, Jianxin Liang, Jiansheng Wei, Huishuai Zhang, and Dongyan Zhao. Videollm knows when to speak: Enhancing time-sensitive video comprehension with video-text duet interaction format. In *EMNLP*, 2025b.

Yueqian Wang, Songxiang Liu, Disong Wang, Nuo Xu, Guanglu Wan, Huishuai Zhang, and Dongyan Zhao. Mmduet2: Enhancing proactive interaction of video mllms with multi-turn reinforcement learning. In *ICLR*, 2026.

Yan Xia, Xiaowei Zhou, Etienne Vouga, Qixing Huang, and Georgios Pavlakos. Reconstructing humans with a biomechanically accurate skeleton. In *CVPR*, 2025.

Haomiao Xiong, Zongxin Yang, Jiazuo Yu, Yunzhi Zhuge, Lu Zhang, Jiawen Zhu, and Huchuan Lu. Streaming video understanding and multi-round interaction with memory-enhanced knowledge. In *ICLR*, 2025.

Ruyi Xu, Guangxuan Xiao, Yukang Chen, Liuning He, Kelly Peng, Yao Lu, and Song Han. Streamingvlm: Real-time understanding for infinite video streams. In *ICLR*, 2026.

Zhenyu Yang, Kairui Zhang, Yuhang Hu, Bing Wang, Shengsheng Qian, Bin Wen, Fan Yang, Tingting Gao, Weiming Dong, and Changsheng Xu. Livestar: Live streaming assistant for real-world online video understanding. In *NeurIPS*, 2025.

Linli Yao, Yicheng Li, Yuancheng Wei, Lei Li, Shuhuai Ren, Yuanxin Liu, Kun Ouyang, Lean Wang, Shicheng Li, Sida Li, et al. Timechat-online: 80% visual tokens are naturally redundant in streaming videos. In *ACM MM*, 2025.

Wei-Hsin Yeh, Yu-An Su, Chih-Ning Chen, Yi-Hsueh Lin, Calvin Ku, Wen-Hsin Chiu, Min-Chun Hu, and Lun-Wei Ku. CoachMe: Decoding sport elements with a reference-based coaching instruction generation model. In *ACL*, 2025.

Andy Zeng, Maria Attarian, Brian Ichter, Krzysztof Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, et al. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv preprint arXiv:2204.00598*, 2022.

Xiangyu Zeng, Kefan Qiu, Qingyu Zhang, Xinhao Li, Jing Wang, Jiaxin Li, Ziang Yan, Kun Tian, Meng Tian, Xinhai Zhao, et al. Streamforest: Efficient online video understanding with persistent event memory. In *NeurIPS*, 2025.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025a.

Hang Zhang, Xin Li, and Lidong Bing. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. In Yansong Feng and Els Lefever (eds.), *EMNLP*, 2023.

Haoji Zhang, Yiqin Wang, Yansong Tang, Yong Liu, Jiashi Feng, and Xiaojie Jin. Flash-vstream: Efficient real-time understanding for long video streams. In *ICCV*, 2025b.

Kairui Zhang, Zhenyu Yang, Bing Wang, Shengsheng Qian, and Changsheng Xu. Querystream: Advancing streaming video understanding with query-aware pruning and proactive response. In *ICLR*, 2026.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In *ICLR*, 2020.

Yichi Zhang, Xin Luna Dong, Zhaojiang Lin, Andrea Madotto, Anuj Kumar, Babak Damavandi, Joyce Chai, and Seungwhan Moon. Proactive assistant dialogue generation from streaming egocentric videos. In *EMNLP*, 2025c.

Yuanhan Zhang, Bo Li, Haotian Liu, Yong Jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. LLaVA-NeXT: A strong zero-shot video understanding model, 2024. Online. Available: `https://llava-vl.github.io/blog/2024-04-30-llava-next-video/`.

Xiaoye Zuo, Nikos Athanasiou, Ginger Delmas, Yiming Huang, Xingyu Fu, and Lingjie Liu. Formcoach: Lift smarter, not harder. *arXiv preprint arXiv:2508.07501*, 2025.

A TRAINING DETAILS

Stage 1. StreamingVLM (Qwen2.5-VL-7B) is fine-tuned with $\mathcal{L}_{\mathrm{LM}}$ alone and without motion inputs on 7,087 windows of 4 s (stride 2 s) cut from the 148 QEVD training workouts. All language-model parameters, including embeddings and the output head, are updated, and the vision encoder is frozen. Videos are sampled at 16 frames per 1-s chunk with at most 117,600 pixels per frame, and the text sink and sliding window hold 512 tokens each. We train for 20 steps of 128 windows (2,560 windows) with AdamW (learning rate 10^{-5}, cosine decay after 6 warm-up steps, no weight decay), bf16 precision, and seed 42.

Stage 2. Supervision windows are cut from 141 training workouts; the remaining 7 are held out. Each window supervises four consecutive 1-s chunks (stride four chunks), and all preceding chunks of the workout are processed as context without language gradients. Of the resulting 5,637 windows, 640 (11.4%) are sampled uniformly with seed 42 and used in a single pass, with one window per GPU on four GPUs (160 steps).

The motion adapter, the state module, and both projections are trained with learning rate 10^{-4}; decoder layers 24–27 and the final normalization use 10^{-5}; all other parameters stay frozen. We use AdamW with weight decay 0.01, cosine decay with 5% warm-up, gradient clipping at 1.0, $\lambda_{\mathrm{pose}} = 0.1$, write-dropout probability 0.15, and state backpropagation through eight chunks. The model uses $K_m = K_s = 8$ tokens and hidden width 256, takes visual features from vision-encoder layer 23, and applies a tolerance window of ± 1.5 s around each annotated feedback time. Video sampling matches Stage 1, and the visual window holds the current chunk and three preceding chunks, as at inference.

Write dropout (Sec. 3.4) sets $z_t = 0$ with probability 0.15, sampled independently for every chunk of every training window. A dropped chunk leaves the Motion State unchanged, while its Form Tokens still reach the decoder. Write dropout is active whenever the motion modules are trained, including the transfer experiments, and is disabled at inference.

Baselines. StreamingVLM fine-tuned on QEVD follows the Stage-1 recipe and budget (20 steps of 128 windows). For the full-workout setting, it is trained on 8,719 windows of 4 s cut from the training workouts and paired with the ordered exercise plan. FitCoach and BioCoach are evaluated with their released checkpoints without further training, and zero-shot models are not trained. Probe variants without Form Tokens or Motion State use the same data, adapters, schedule, and selection rule as the full model.

B PROBING FITCOACH AND BIOCOACH

FitCoach and BioCoach do not take free-form queries, so both probes are read from their own feedback. For the exercise probe, the exercise name is removed from their instruction and the first exercise named in their feedback within the segment is taken as the answer; a segment with no named exercise is scored as incorrect. For the count probe, the most recent count reported in their feedback before the probe moment is used, and a probe with no reported count is scored as an error equal to the reference count.

Table 4: Zero-shot video-language models under the exercise-segment protocol, from Panchal et al. (2024); T-F is not reported there for zero-shot models.

Method	METEOR↑	ROUGE-L↑	BERTScore↑	LLM-Acc.↑
InstructBLIP (Dai et al., 2023)	0.047	0.040	0.839	1.56
Video-LLaVA (Lin et al., 2024)	0.057	0.025	0.847	2.16
Video-ChatGPT (Maaz et al., 2024)	0.098	0.078	0.850	1.91
Video-LLaMA (Zhang et al., 2023)	0.101	0.077	0.859	1.29
LLaMA-VID (Li et al., 2024a)	0.100	0.079	0.859	2.20
LLaVA-NeXT (Zhang et al., 2024)	0.104	0.078	0.858	2.27