Large Multimodal Models as Evidence-Grounded Personal Coaches

Anonymous ACL submission

Abstract

Giving useful exercise feedback requires more than detecting a wrong movement: a coach must identify the mistake, explain it in context, and ground the correction in understandable examples. We present **EviCoach**, a training-free framework that uses large multimodal models as personal coaches by treating expert demonstrations as visual evidence for feedback generation. Given a user-recorded exercise video, EviCoach identifies the exercise setting and likely form errors, retrieves expert demonstrations matched by exercise type, viewpoint, body shape proxy, and motion phase, and generates taxonomy-constrained error labels, natural-language coaching feedback, and optional visual corrections. To evaluate the framework, we build a benchmark of paired correct- and incorrect-form exercise videos with a trainer-verified error taxonomy. Experiments show that structured taxonomies and retrieved visual evidence make feedback more specific, actionable, and reliable than free-form prompting alone.

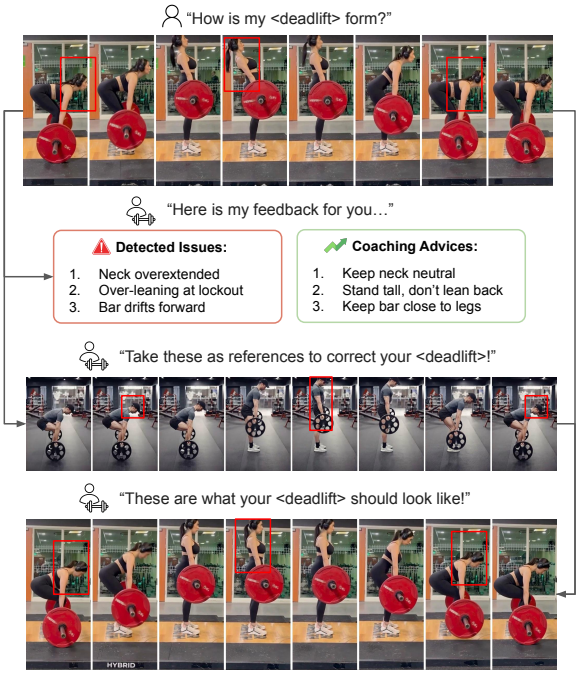

Figure 1: **EviCoach** turns exercise videos into evidence-grounded coaching feedback. Given a user-recorded clip, the system retrieves visually compatible expert demonstrations as visual evidence, diagnoses form errors, and generates natural-language coaching feedback, and personalized visual corrections.

1 Introduction

Effective exercise feedback requires more than detecting that a movement is wrong. A useful coach must identify the specific form error, explain it in context, and provide guidance that the user can apply to their own body and camera view. Prior work on fitness coaching and action-quality assessment has made progress in recognizing and scoring human performance from visual observations (Fieraru et al., 2021; Parmar et al., 2022; Ashutosh et al., 2025). However, feedback often remains textual, score-based, or pose-centered. A cue such as "keep your knees aligned with your toes" may be correct, but a novice may still not know what the corrected posture should look like during their own squat. This makes exercise coaching a challenging multimodal feedback generation problem that requires reasoning over motion, visual context, exercise-specific errors, and natural-language instructions.

Large vision-language models (LVLMs) offer a promising interface for this problem, as they can interpret exercise videos and generate corrective feedback in language. Recent datasets and systems have connected language with skilled human activities, egocentric and exocentric observations, and actionable feedback (Tang et al., 2023; Li et al., 2024b; Grauman et al., 2024). Yet open-ended LVLM feedback can be inconsistent, underspecified, or hallucinated when fine-grained physical errors are ambiguous (Li et al., 2023; Huang et al., 2024; Zhong et al., 2024). In embodied coaching, feedback should not merely sound plausible;

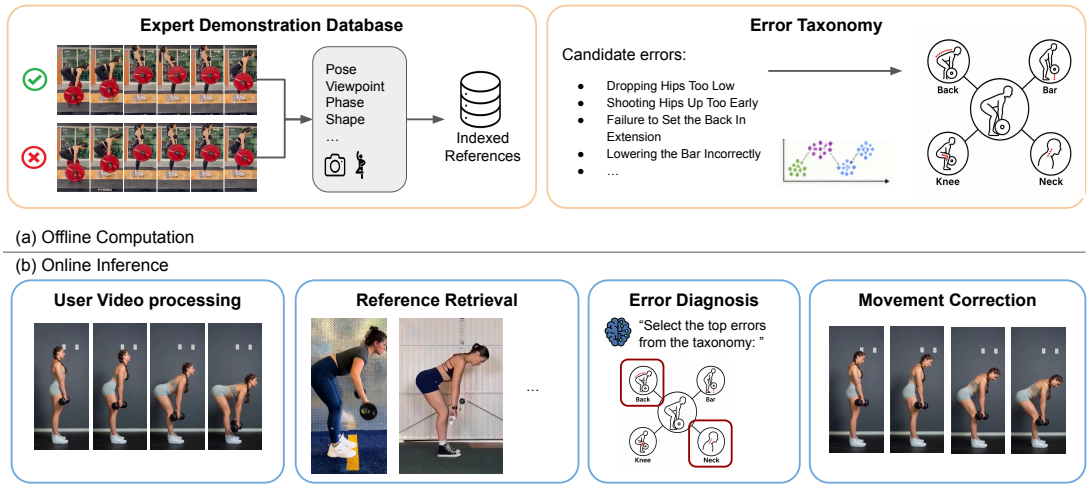

Figure 2: Overview of EviCoach. Offline, the system constructs an expert demonstration database and a trainer-verified error taxonomy. Online, given a user-recorded exercise video, EviCoach samples keyframes, retrieves visually compatible expert demonstrations as evidence, performs taxonomy-constrained error diagnosis with an LVLM, and generates grounded coaching feedback with optional personalized visual corrections.

it should be grounded in the observed motion, constrained by exercise-specific knowledge, and supported by concrete visual evidence.

We thus present **EviCoach**, a training-free framework that uses large multimodal models as personal coaches by treating expert demonstrations as visual evidence for feedback generation. As shown in Fig. 1, given a short user-recorded exercise video, EviCoach samples keyframes from one repetition, identifies the exercise setting and likely form errors, and retrieves expert demonstrations matched by exercise type, equipment, body-shape proxy, camera viewpoint, and motion phase. These retrieved references ground taxonomy-constrained error diagnosis, natural-language coaching feedback, and optional personalized visual corrections.

EviCoach relies on two structured components. First, we construct an exercise-specific error taxonomy by combining LVLM-assisted discovery of common form deviations with verification from an expert personal trainer. This converts open-ended feedback into structured multi-label diagnosis over trainer-verified categories. Second, we retrieve visually compatible expert demonstrations at inference time, which provide evidence for interpreting the user's motion and references for generating more specific and actionable feedback.

We evaluate EviCoach on a curated benchmark of paired correct- and incorrect-form exercise videos covering multiple strength-training exercises and error types. We study whether taxonomy constraints improve the reliability of structured error diagnosis, and whether grounding on retrieved visual evidence improves feedback quality and visual corrections over prompt-only or randomly retrieved baselines.

Our contributions are three-fold:

- We introduce **EviCoach**, a training-free framework for evidence-grounded multimodal coaching feedback from exercise videos.
- We construct a curated expert demonstration database and a trainer-verified taxonomy of common exercise-form errors, demonstrating that taxonomy constraints and retrieved visual evidence improve error diagnosis reliability.
- We instantiate retrieval-grounded feedback with personalized visual corrections and a verifier-editor refinement loop.

2 Related Work

2.1 Multimodal Human Activities Feedback

Automated feedback for skilled human activities has been studied in fitness coaching, action-quality assessment, and procedural activity understanding. Prior exercise-coaching systems recognize incorrect form and produce interpretable feedback from pose or body representations (Fieraru et al., 2021; Parmar et al., 2022), while recent datasets introduce language instructions, multi-view recordings, and egocentric/exocentric observations for multimodal activity understanding (Tang et al., 2023; Li et al., 2024b; Grauman et al., 2024). Beyond fitness, action-quality assessment evaluates fine-grained performance and rubric-informed correct-

ness in skilled activities (Xu et al., 2022, 2024; Zhang et al., 2024; Majeedi et al., 2024). Recent work further studies live coaching, actionable feedback generation, composite error recognition, and cross-sport feedback from video (Panchal et al., 2024; Ashutosh et al., 2025; Wang et al., 2024; Rai and Kovashka, 2026). EviCoach builds on this direction but focuses on grounded multimodal feedback generation: it produces structured error labels and coaching feedback using retrieved expert demonstrations as visual evidence and personalized corrections as user-facing explanations.

2.2 Grounding and Hallucination

Visual instruction tuning methods such as LLaVA and InstructBLIP enable models to follow natural-language prompts and solve diverse multimodal tasks (Liu et al., 2023; Dai et al., 2023). However, open-ended LVLM generation can hallucinate objects, attributes, actions, or relations that are not supported by the visual input. Prior work has proposed benchmarks and mitigation strategies for object hallucination (Li et al., 2023; Zhou et al., 2024; Liu et al., 2024), analyzed broader multimodal hallucination behaviors (Huang et al., 2024; Zhong et al., 2024), and explored retrieval-based decoding with visual references (Lee and Song, 2025). These issues are especially important for embodied coaching, where plausible but ungrounded advice may be misleading or unsafe. EviCoach instead performs structured multi-label diagnosis and uses visually matched references as evidence, making feedback more grounded and easier to evaluate.

2.3 Retrieval-Augmented Generation

Retrieval-augmented generation uses external examples to ground model outputs beyond parametric knowledge. This idea has been extended to multimodal settings, where retrieved image-text pairs, visual exemplars, or multimodal memories provide additional context for reasoning and generation (Yasunaga et al., 2023). Recent multimodal RAG benchmarks and systems emphasize visual evidence and source attribution beyond text alone (Hu et al., 2025; Ma et al., 2025). In visual generation, retrieved examples have been used to improve fidelity, controllability, and rare-concept synthesis (Blattmann et al., 2022; Sheynin et al., 2023; Chen et al., 2023), with recent work exploring fine-grained reference selection at generation time (Yuan et al., 2025; Shalev-Arkushin et al., 2026). Our work follows this broader principle, but uses retrieval for grounded feedback rather than visual fidelity alone: retrieved expert demonstrations serve as evidence for interpreting exercise form errors and visualizing corrected posture.

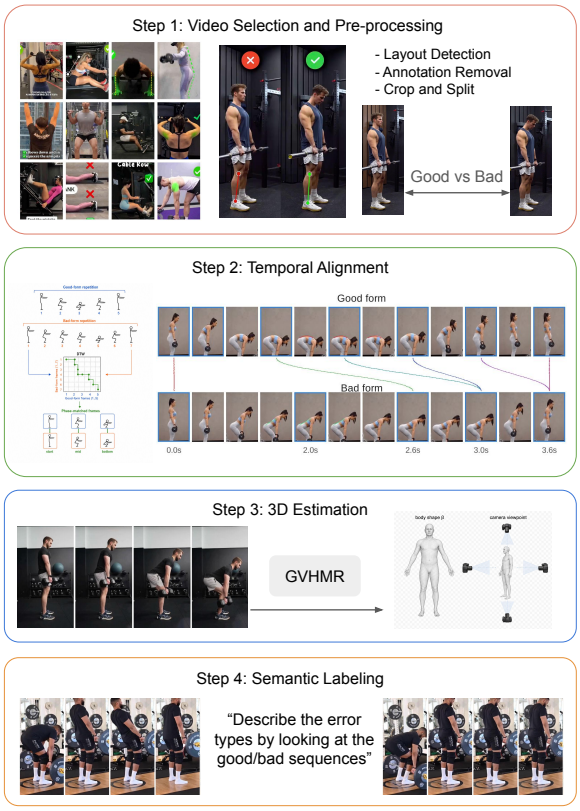

Figure 3: Expert demonstration database curation. Side-by-side correct/incorrect exercise videos are split into clips, temporally aligned into phase-matched frames, and processed to obtain body-shape proxy, viewpoint, and coarse error metadata. The resulting entries are indexed by exercise, equipment, phase, body-shape proxy, and camera angle for retrieval-grounded feedback.

3 Approach

Given a short exercise video, our goal is to generate grounded multimodal coaching feedback: structured form-error labels, actionable natural-language advice, and visual corrections. EviCoach addresses this by combining exercise-specific error taxonomies with visually matched expert demonstrations. We treat visual correction as a user-facing explanation layer, while the central task is structured and evidence-grounded feedback generation.

Fig. 2 summarizes the framework. Offline, EviCoach builds an expert demonstration database from correct-form videos and paired correct/incorrect comparisons, and constructs a trainer-verified error taxonomy. Online, it samples keyframes from a user video, retrieves visually compatible expert

demonstrations, performs taxonomy-constrained diagnosis with an LVLM, and generates grounded coaching feedback with optional personalized corrections.

3.1 Expert Demonstration Database

As shown in Fig. 3, we build an expert demonstration database to support evidence-grounded feedback. Each entry is indexed by exercise metadata, motion phase, body-shape proxy, and camera viewpoint, allowing retrieval of expert examples that are visually comparable to the user's motion.

3.1.1 Video Selection

We curate public short-form fitness videos that either (1) show side-by-side comparisons of correct and incorrect form for the same exercise or (2) show good-form-only repetitions of an exercise. The database covers 16 exercise categories, spanning diverse body-shape proxies, camera viewpoints, equipment types, and execution styles. For each video, we run DWPose (Yang et al., 2023) to localize the person and split side-by-side comparisons into correct-form and incorrect-form clips. We crop each clip around the full body and tag it with its exercise category and equipment type using an LVLM. Correct-form clips are used as expert demonstrations for retrieval, while paired correct/incorrect clips are used for taxonomy construction and evaluation.

3.1.2 Temporal Alignment

To compare correct and incorrect executions at corresponding movement stages, we detect exercise repetitions from the pose sequence and select one complete repetition from each clip. Since paired motions may differ in speed, we align their skeleton sequences using Dynamic Time Warping (DTW) (Sakoe and Chiba, 1978). After alignment, we sample 8 frames from each repetition, corresponding to approximately 2–3 seconds of motion. The sampled frame index is used as a discrete phase label, enabling retrieval of expert references that match the user's current movement stage.

3.1.3 3D Estimation

To support morphology- and viewpoint-aware retrieval, we run GVHMR (Shen et al., 2024) on the good-form clips to estimate SMPL-X parameters and camera pose. We use the SMPL-X shape parameters β (Pavlakos et al., 2019) as a body-shape proxy, and the estimated camera pose as a continuous viewpoint representation. Each reference frame is therefore indexed by exercise, equipment, phase, body-shape proxy, and camera angle.

3.1.4 Coarse Semantic Labeling

To obtain initial semantic labels for form deviations, we use an LVLM to compare each paired correct- and incorrect-form clip. For each pair, the model identifies the major visible deviation in the incorrect execution relative to the correct demonstration. Some source videos contain instructional overlays, such as text labels, arrows, curves, or heatmaps, that explicitly indicate the demonstrated error. We use these overlays only as weak supervision for coarse labeling. After obtaining the labels, we remove visible annotations from the frames with an off-the-shelf image-editing model (Gemini Team, 2023), so that downstream retrieval and evaluation are not driven by textual or graphical hints.

3.2 Error Taxonomy

For each exercise category, we construct a compact taxonomy of common form errors to support structured diagnosis and evaluation. We first prompt an LVLM to propose candidate error types for each exercise, together with short descriptions of their visual symptoms. An expert personal trainer then reviews the candidates and removes errors that are not common, reliably visible, or actionable for coaching. We next normalize the coarse free-form labels obtained during database curation into this trainer-verified label space.

Since an LVLM may describe the same error with different surface forms, such as "knees collapsing inward" and "knee valgus," we embed both the coarse descriptions and taxonomy entries using CLIP text embeddings (Radford et al., 2021). Each coarse label is assigned to its nearest taxonomy entry within the same exercise category. A video may receive multiple labels when several deviations are visible. Taxonomy entries that are trainer-approved but not observed in the collected videos remain available as coaching categories at inference time, but are excluded from benchmark evaluation. This procedure converts noisy open-ended descriptions into a structured multi-label diagnosis space.

3.3 Error Diagnosis

3.3.1 Reference Retrieval

Given a user workout sequence, we retrieve a small set of visually compatible expert references from the demonstration database. We first identify the exercise and equipment type, then filter

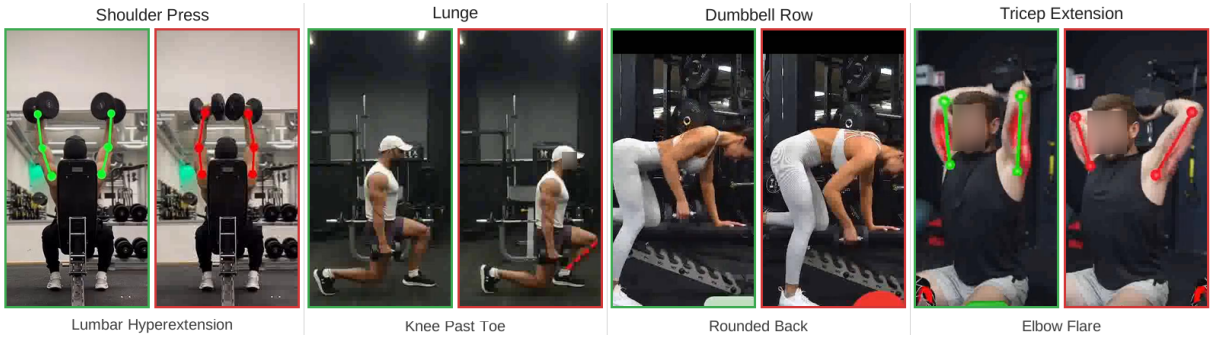

Figure 4: Gallery of paired examples from the expert demonstration database. Each example contains a correct-form frame and a corresponding incorrect-form frame from the same exercise, along with the associated coarse error type. We intentionally keep the original overlaid annotations in this figure to illustrate the weak supervision available during taxonomy construction; these annotations are removed and are not used during retrieval or evaluation.

candidate expert demonstrations by the same metadata. The remaining candidates are ranked by camera-viewpoint similarity and SMPL-X body-shape proxy distance. For each sampled user frame, we select the phase-aligned frame from the top-ranked expert videos. If fewer than the target number of references satisfy all filters, we relax the constraints and select the best remaining candidates. The retrieved top-three demonstrations serve as visually comparable good-form evidence for subsequent diagnosis and optional correction visualization, rather than as exact pose targets.

3.3.2 Retrieval-Grounded Error Tagging

After retrieving expert references, EviCoach performs taxonomy-constrained error diagnosis. Given the identified exercise and equipment type, we retrieve the corresponding exercise-specific taxonomy, including the error names and their visual descriptions. We then provide the LVLM with the taxonomy, the 8 sampled frames from the user's repetition, and the retrieved expert frames from the database. Using a short frame sequence rather than a single image allows the model to observe phase-dependent errors, while the retrieved references ground the diagnosis in visually comparable correct-form demonstrations.

We formulate diagnosis as multi-label classification over the taxonomy. For an exercise category with error set $\mathcal{C}_e = \{c_1, c_2, \ldots, c_K\}$, the model predicts a binary vector

$$\mathbf{y} = (y_1, y_2, \ldots, y_K), \quad y_k \in \{0, 1\},$$

where $y_k = 1$ indicates that error type c_k is present in the user sequence. This structured output constrains the LVLM to choose from trainer-verified coaching categories rather than producing unconstrained free-form descriptions, making the diagnosis more consistent and human-aligned.

3.4 Visual Feedback

Beyond structured diagnosis results and natural-language feedback, EviCoach produces corrected keyframes as a user-facing visual explanation. For each user frame, the system generates a frame-specific correction instruction from the diagnosed error labels and the current motion phase. A pre-trained multimodal image-editing model then receives the user frame, retrieved expert references, error labels, and instruction, and produces a corrected keyframe that aims to preserve the user's appearance, clothing, equipment, and scene context. The retrieved references serve as examples of proper form, while the instruction specifies the motion phase, error(s), and intended correction(s).

To improve the reliability of visualization, we use a lightweight verifier-editor loop. After the first edit, an LVLM checks whether the target error has been corrected by comparing the edited frame with the intended labels and retrieved references. If the correction is insufficient or overly aggressive, the system revises the previous output with a strengthened or softened instruction. We allow at most two refinement iterations per frame.

4 Experiments

4.1 Experimental Setup

We evaluate EviCoach on two exercise-video collections covering 16 strength-training exercises. The first collection contains 411 paired correct/incorrect comparison videos, where each video shows a correct-form demonstration and an incorrect-form execution of the same exercise. We use these paired videos for evaluation, including taxonomy coverage, error diagnosis, feedback quality, and optional visual correction. The second

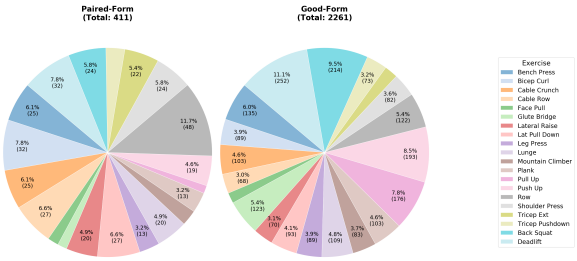

Figure 5: Distribution of the two video collections used in EviCoach. The paired correct/incorrect dataset is used for evaluation, while the good-form-only dataset builds the expert demonstration database for retrieval.

collection contains 2,261 good-form videos and is used to construct the expert demonstration database for retrieval. During evaluation, EviCoach retrieves references only from this good-form database, so the paired evaluation videos are never used as retrieval targets. The exercise distributions are illustrated in Fig. 5, and example pairs are shown in Fig. 4. We blur frontal faces in all figures and use the videos only for research-related analysis.

For each video, we extract one complete repetition and sample 8 keyframes as phase-aware visual inputs. We evaluate three aspects of the system: taxonomy-constrained error diagnosis, retrieval of visually compatible expert demonstrations, and retrieval-grounded feedback generation. For diagnosis, we report multi-label classification metrics, including precision, recall, macro and micro F1. For visual feedback, we compare against no-retrieval, random-reference, and weaker retrieval baselines, and evaluate whether matched references and the verifier-editor loop improve correction success, reference alignment, and preservation quality.

4.2 Error Diagnosis

4.2.1 Taxonomy Coverage

We first evaluate whether the trainer-verified taxonomy covers the form deviations observed in the paired correct/incorrect evaluation videos. For each incorrect-form clip, we collect coarse free-form error descriptions from the paired comparison and retrieve the nearest trainer-verified taxonomy entry within the same exercise category using CLIP text embeddings. This nearest-neighbor match is treated only as a candidate mapping. We manually annotate each candidate mapping as valid or invalid based on whether the taxonomy entry preserves the intended coaching meaning.

Across 16 exercises and 411 paired evaluation videos, we obtain 192 unique free-form error descriptions and 118 trainer-verified taxonomy entries. To compute coverage, we measure the fraction of unique descriptions whose candidate mapping is manually judged valid. The taxonomy covers 98.4% of descriptions under micro averaging and 96.2% under macro averaging across exercise categories, with a median exercise-level coverage of 97.9%. These results suggest that the taxonomy provides a compact but expressive label space for structured coaching feedback. The most frequent concepts in the collected descriptions, such as knee alignment, back posture, hip position, depth, and torso angle, are represented in the final taxonomy. This supports using the taxonomy as the label space for the diagnosis experiments in Section 4.2.2.

We further visualize the semantic structure of the taxonomy for one representative exercise. Fig. 6 shows the UMAP projection of coarse semantic labels from the *Deadlift* category and their assignments to trainer-verified taxonomy labels. Four of the twelve candidate labels are either redundant with other error categories or lack sufficient visual support in the data, and are therefore removed to keep the final taxonomy compact and observable.

Table 1: LVLM diagnosis benchmark under the same taxonomy-constrained setting. Higher is better for all.

Model	Prec.	Rec.	Macro-F1	Micro-F1
Gemini 3.1 Pro	0.69	0.76	0.70	0.71
GPT-5.5	**0.82**	**0.88**	**0.82**	**0.86**
Claude Opus 4.7	0.73	0.77	0.75	0.75
Qwen 3.5 Plus	0.52	0.59	0.54	0.55
LLaVA 72B	0.28	0.20	0.21	0.26

4.2.2 LVLM Diagnosis Benchmark

We benchmark different modern LVLMs under the same taxonomy-constrained diagnosis setting, including proprietary frontier models, *Gemini 3.1 Pro* (Google DeepMind, 2026), *GPT-5.5* (OpenAI, 2026), and *Claude Opus 4.7* (Anthropic, 2026), a commercial multimodal model *Qwen 3.5 Plus* (Qwen Team, 2026), and the open-source *LLaVA-OneVision 72B* (Li et al., 2024a). Each model receives the exercise-specific taxonomy, 8 sampled keyframes from one repetition, and the retrieved expert references, and predicts a multi-label error vector over the taxonomy. We evaluate against human-verified multi-label annotations for each clip using precision, recall, macro F1, and micro F1. Metric details are defined in the Appendix.

Tab. 1 shows that GPT-5.5 achieves the highest precision, recall, macro-F1, and micro-F1 un-

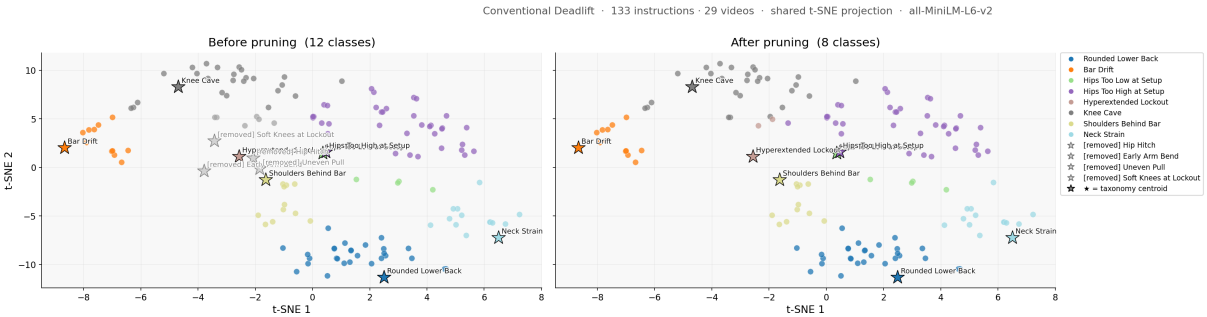

Figure 6: Visualization of error taxonomy curation for the DEADLIFT category. LVLM-generated coarse error descriptions are first matched to nearest trainer-verified taxonomy labels in the original embedding space, and the candidate mappings are manually verified. Labels that are redundant or lack sufficient visual support in the paired data are removed from the final taxonomy. For visualization, the embeddings are projected to 2D with UMAP. Small points denote coarse descriptions, while large stars indicate expert-defined taxonomy centers.

Table 2: Ablation of diagnosis prompting and temporal context. The default setting uses the exercise-specific taxonomy and 8 sampled keyframes from one repetition.

Setting	Prec.	Rec.	Macro-F1	Micro-F1
Free-form	0.35	0.39	0.35	0.37
1-frame + Tax.	0.65	0.70	0.64	0.68
8-frame + Tax.	**0.82**	**0.88**	**0.82**	**0.86**

der the same taxonomy-constrained setting. The gap is especially large compared with smaller or open-source models, suggesting that fine-grained exercise-form diagnosis remains challenging and benefits from stronger multimodal reasoning.

4.2.3 Ablation Studies

We next evaluate whether structured taxonomy prompting and temporal context improve error diagnosis with the best-performing LVLM, GPT-5.5. We compare our default setting with two ablations: (1) free-form diagnosis, where the model predicts open-ended error descriptions that are post-hoc mapped to the nearest taxonomy entry using CLIP text embeddings, and (2) single-frame diagnosis, where the model receives the same taxonomy but only one sampled frame instead of the full 8-frame sequence. Table 2 shows that taxonomy-constrained prompting substantially improves diagnosis over free-form prompting, raising macro-F1 from 0.35 to 0.64 with a single frame. Using 8 sampled keyframes further improves performance across all metrics, reaching 0.82 macro-F1 and 0.86 micro-F1, suggesting that temporal context helps capture phase-dependent form errors.

4.3 Retrieval-Grounded Visual Feedback

We evaluate whether retrieved expert demonstrations improve user-facing coaching feedback and optional visual corrections. To isolate the effect of retrieval, all methods use the same input frames and target error labels for correction. We compare five settings: *No retrieval*, which generates feedback and corrections from the user frames and predicted labels only; *Random references*, which uses randomly selected good-form examples; *Exercise-only references*, which retrieves correct-form examples from the same exercise category without phase, viewpoint, or body-shape-proxy matching; *EviCoach w/o verifier*, which uses the full retrieval strategy but performs a single editing pass; and *EviCoach*, which additionally applies the verifier–editor loop to refine under- or over-corrections.

4.3.1 Quantitative Results

We evaluate visual correction along two axes: whether the target form error is corrected, and whether the edited image preserves the original appearance and scene. For correction quality, we report (1) human correction success rate, where annotators judge whether the edited frame fixes the target error while maintaining a plausible exercise pose, and (2) LLM correction success, where GPT-5.5 judges the edited frame under the same criterion. We further report two reference-alignment metrics: ΔCLIP-Ref, the increase in CLIP similarity between the corrected frame and the retrieved expert references relative to the input frame, and ΔPose-Ref, the relative reduction in normalized 2D pose distance to the retrieved expert references. Finally, we measure preservation using appearance similarity between the input and corrected person crops, and background similarity between the input and corrected non-person regions.

Table 3 shows that matched expert references substantially improve visual correction quality over prompt-only and random-reference baselines. Adding the verifier–editor loop further increases

Table 3: Visual correction evaluation. Correction success measures whether target form errors are fixed, judged by human annotators and GPT-5.5. ΔCLIP and ΔPose measure improvement toward retrieved expert references in CLIP feature space and normalized 2D pose space. Appearance and background similarity measure preservation with respect to the original input frame.

Method	Correction		Reference Align.		Preservation	
	Human CS (%) ↑	LLM CS (%) ↑	ΔCLIP (× 0.01) ↑	ΔPose (%) ↑	App. Sim ↑	BG Sim ↑
No Retrieval	38.2	33.4	6	7.8	**0.91**	0.93
Random Refs	29.1	35.9	9	10.4	0.89	**0.95**
Exercise-Only Refs	46.5	57.7	21	21.6	0.87	0.94
EviCoach w/o Verifier	78.5	88.2	24	27.9	0.85	0.93
EviCoach	**92.0**	**94.1**	**29**	**32.9**	0.88	0.92

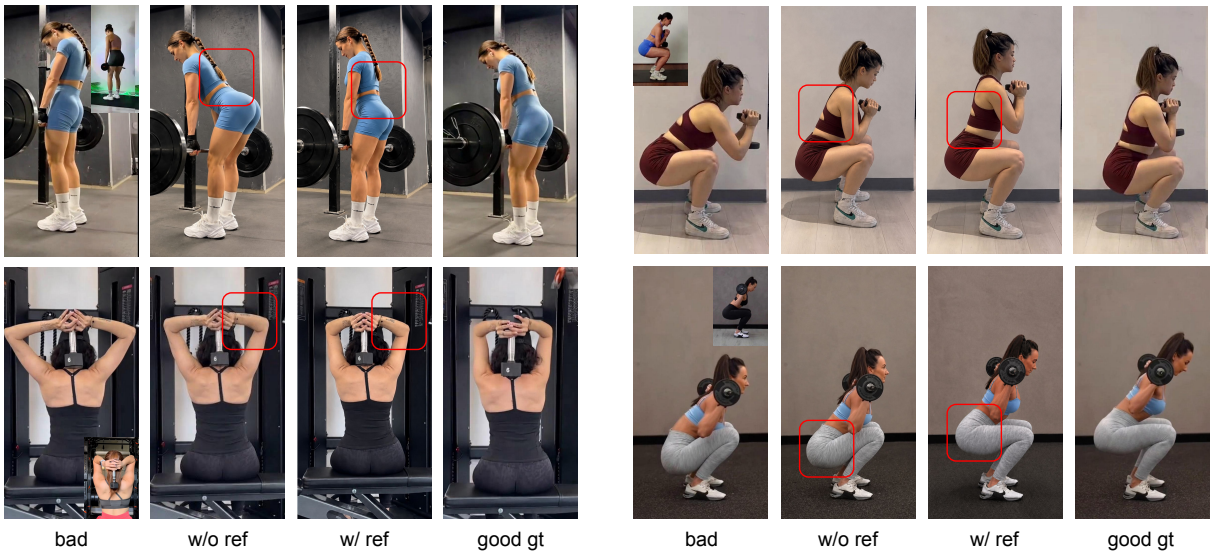

Figure 7: Qualitative examples of retrieval-grounded visual feedback. We show the user input, retrieved expert demonstrations, visual corrections generated with and without references, and the paired correct-form frame. Matched references provide visual evidence for phase-specific coaching cues and plausible personalized corrections.

both human and LLM correction success, while also improving alignment with retrieved expert demonstrations in CLIP and pose space. The preservation metrics remain comparable across methods, indicating that stronger correction does not substantially degrade the user's appearance or background. Overall, these results suggest that retrieval-grounded visual evidence helps the model produce more accurate and visually actionable corrections while preserving the original scene. The verifier–editor loop further improves correction precision by refining edits that are overly weak or aggressive.

4.3.2 Qualitative Analysis

Figure 7 shows representative examples of visual corrections. Without retrieved references, the model can often generate a plausible corrected image, but the magnitude of the edit is less precise. For example, in squats, it may adjust the squat depth or straighten the back, but still fail to make the thighs approximately parallel to the ground or align the spine consistently; in some cases, the correction is either too weak or overly aggressive. In contrast, EviCoach uses visually compatible demonstrations to provide concrete targets for phase-specific corrections. This makes the corrected previews more controlled and better aligned with the intended feedback.

5 Conclusion

We presented **EviCoach**, a training-free framework for evidence-grounded multimodal exercise coaching. Given a short user-recorded exercise video, EviCoach performs taxonomy-constrained error diagnosis with an LVLM, retrieves matched expert demonstrations as visual evidence, and generates structured coaching feedback with optional personalized visual corrections. Our experiments show that structured taxonomies and retrieved visual evidence make feedback more reliable and visually concrete than free-form prompting alone.

Limitations

EviCoach currently focuses on a limited set of common strength-training exercises and depends on the coverage of the expert demonstration database. When the user's exercise, equipment, camera view, body shape, or error type is poorly represented, retrieval may provide less useful evidence. The system also depends on the reliability of pretrained components for pose estimation, body-shape estimation, LVLM diagnosis, and image editing, and may fail under occlusion, unusual viewpoints, loose clothing, poor lighting, or complex equipment interactions. Although visual corrections are intended only as user-facing explanations, the editing model may still introduce artifacts, over-correct posture, or change irrelevant visual details. Future work includes expanding the benchmark to more exercises and populations, improving uncertainty estimation for unsafe or ambiguous feedback, incorporating stronger temporal reasoning for full-repetition analysis, and developing on-device or hybrid web deployment for interactive coaching.

References

Anthropic. 2026. Introducing Claude Opus 4.7. https://www.anthropic.com/news/claude-opus-4-7. Accessed: 2026-05-24.

Kumar Ashutosh, Tushar Nagarajan, Georgios Pavlakos, Kris Kitani, and Kristen Grauman. 2025. ExpertAF: Expert actionable feedback from video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13582–13594.

Andreas Blattmann, Robin Rombach, Kaan Oktay, Jonas Müller, and Björn Ommer. 2022. Retrieval-augmented diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 15309–15324.

Wenhu Chen, Hexiang Hu, Chitwan Saharia, and William W. Cohen. 2023. Re-Imagen: Retrieval-augmented text-to-image generator. In *International Conference on Learning Representations (ICLR)*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36.

Mihai Fieraru, Mihai Zanfir, Silviu Cristian Pirlea, Vlad Olaru, and Cristian Sminchisescu. 2021. 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.

Gemini Team. 2023. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Google DeepMind. 2026. Gemini 3.1 Pro Model Card. https://deepmind.google/models/model-cards/gemini-3-1-pro/. Accessed: 2026-05-24.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, and 1 others. 2024. Ego-Exo4D: Understanding skilled human activity from first- and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19383–19400.

Wenbo Hu, Jia-Chen Gu, Zi-Yi Dou, Mohsen Fayyaz, Pan Lu, Kai-Wei Chang, and Nanyun Peng. 2025. MRAG-Bench: Vision-centric evaluation for retrieval-augmented multimodal models. In *International Conference on Learning Representations*.

Wen Huang, Hongbin Liu, Minxin Guo, and Neil Zhenqiang Gong. 2024. Visual hallucinations of multimodal large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*.

Jihoon Lee and Min Song. 2025. Retrieval visual contrastive decoding to mitigate object hallucinations in large vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8200–8219, Vienna, Austria. Association for Computational Linguistics.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024a. LLaVA-OneVision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Yuan-Ming Li, Wei-Jin Huang, An-Lan Wang, Ling-An Zeng, Jing-Ke Meng, and Wei-Shi Zheng. 2024b. EgoExo-Fitness: Towards egocentric and exocentric full-body action understanding. In *Computer Vision – ECCV 2024*, volume 15078 of *Lecture Notes in Computer Science*, pages 363–382. Springer.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2024. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *International Conference on Learning Representations (ICLR)*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36.

Xueguang Ma, Shengyao Zhuang, Bevan Koopman, Guido Zuccon, Wenhu Chen, and Jimmy Lin. 2025. VISA: Retrieval augmented generation with visual source attribution. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30154–30169, Vienna, Austria. Association for Computational Linguistics.

Abrar Majeedi, Viswanatha Reddy Gajjala, Satya Sai Srinath Namburi GNVV, and Yin Li. 2024. RICA^2: Rubric-informed, calibrated assessment of actions. In *Computer Vision – ECCV 2024*, volume 15121 of *Lecture Notes in Computer Science*, pages 143–161. Springer.

OpenAI. 2026. Introducing GPT-5.5. https://openai.com/index/introducing-gpt-5-5/. Accessed: 2026-05-24.

Sunny Panchal, Apratim Bhattacharyya, Guillaume Berger, Antoine Mercier, Cornelius Böhm, Florian Dietrichkeit, Reza Pourreza, Xuanlin Li, Pulkit Madan, Mingu Lee, Mark Todorovich, Ingo Bax, and Roland Memisevic. 2024. What to say and when to say it: Live fitness coaching as a testbed for situated interaction. In *Advances in Neural Information Processing Systems*, volume 37. Datasets and Benchmarks Track.

Paritosh Parmar, Amol Gharat, and Helge Rhodin. 2022. Domain knowledge-informed self-supervised representations for workout form assessment. In *Computer Vision – ECCV 2022*, volume 13698 of *Lecture Notes in Computer Science*, pages 105–123. Springer.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. 2019. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985.

Qwen Team. 2026. Qwen3.5: Towards native multimodal agents. https://qwen.ai/blog?id=qwen3.5. Accessed: 2026-05-24.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Arushi Rai and Adriana Kovashka. 2026. Generalizing sports feedback generation by watching competitions and reading books: A rock climbing case study. In *Proceedings of the 2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 8136–8145.

Hiroaki Sakoe and Seibi Chiba. 1978. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49.

Rotem Shalev-Arkushin, Rinon Gal, Amit H. Bermano, and Ohad Fried. 2026. ImageRAG: Dynamic image retrieval for reference-guided image generation. In *International Conference on Learning Representations (ICLR)*.

Zehong Shen, Huaijin Pi, Yan Xia, Zhi Cen, Sida Peng, Zechen Hu, Hujun Bao, Ruizhen Hu, and Xiaowei Zhou. 2024. World-grounded human motion recovery via gravity-view coordinates. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11.

Shelly Sheynin, Oron Ashual, Adam Polyak, Uriel Singer, Oran Gafni, Eliya Nachmani, and Yaniv Taigman. 2023. kNN-diffusion: Image generation via large-scale retrieval. In *International Conference on Learning Representations (ICLR)*.

Yansong Tang, Jinpeng Liu, Aoyang Liu, Bin Yang, Wenxun Dai, Yongming Rao, Jiwen Lu, Jie Zhou, and Xiu Li. 2023. FLAG3D: A 3d fitness activity dataset with language instruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22106–22117.

Shunli Wang, Shuaibing Wang, Dingkang Yang, Mingcheng Li, Haopeng Kuang, Xiao Zhao, Liuzhen Su, Peng Zhai, and Lihua Zhang. 2024. CPR-Coach: Recognizing composite error actions based on single-class training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18782–18792.

Jinglin Xu, Yongming Rao, Xumin Yu, Guangyi Chen, Jie Zhou, and Jiwen Lu. 2022. FineDiving: A fine-grained dataset for procedure-aware action quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2949–2958.

Jinglin Xu, Sibo Yin, Guohao Zhao, Zishuo Wang, and Yuxin Peng. 2024. FineParser: A fine-grained spatio-temporal action parser for human-centric action quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14628–14637.

Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. 2023. Effective whole-body pose estimation with two-stages distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 4210–4220.

Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Richard James, Jure Leskovec, Percy Liang, Mike Lewis, Luke Zettlemoyer, and Wen-Tau Yih. 2023.

Retrieval-augmented multimodal language modeling. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pages 39755–39769. PMLR.

Huaying Yuan, Ziliang Zhao, Shuting Wang, Shitao Xiao, Minheng Ni, Zheng Liu, and Zhicheng Dou. 2025. FineRAG: Fine-grained retrieval-augmented text-to-image generation. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING)*, pages 11196–11205, Abu Dhabi, UAE. Association for Computational Linguistics.

Shiyi Zhang, Sule Bai, Guangyi Chen, Lei Chen, Jiwen Lu, Junle Wang, and Yansong Tang. 2024. Narrative action evaluation with prompt-guided multimodal interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18430–18439.

Weihong Zhong, Xiaocheng Feng, Liang Zhao, Qiming Li, Lei Huang, Yuxuan Gu, Weitao Ma, Yuan Xu, and Bing Qin. 2024. Investigating and mitigating the multimodal hallucination snowballing in large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2024. Analyzing and mitigating object hallucination in large vision-language models. In *International Conference on Learning Representations (ICLR)*.

A Artifact Documentation and Intended Use

We use and create several research artifacts, including an expert demonstration database, an exercise-form error taxonomy, preprocessing and retrieval code, LVLM prompts, and evaluation annotations.

Created artifacts. The created artifacts include: (1) a trainer-verified taxonomy of common form errors for 16 strength-training exercises; (2) derived metadata for expert demonstrations, including exercise category, equipment type, motion phase, viewpoint estimate, body-shape proxy, and coarse error labels; (3) prompts for taxonomy construction, diagnosis, feedback generation, visual correction, and LLM-based evaluation; and (4) code for preprocessing, retrieval, evaluation, and metric computation. The taxonomy supports structured exercise-form diagnosis and evaluation, while the retrieval metadata supports evidence-grounded feedback using visually compatible correct-form demonstrations as references rather than exact pose targets.

Release plan. We plan to release the code, prompts, evaluation scripts, and trainer-verified error taxonomy. We will also release derived metadata and aggregate statistics needed to reproduce the taxonomy construction, retrieval protocol, and evaluation procedures, subject to privacy and platform-use constraints. We will not redistribute original videos, downloaded video files, raw frames, or creator-identifying media.

Source videos. The source videos are public short-form fitness videos collected from online platforms and used only for research-related analysis. Although publicly viewable, the copyright and privacy interests remain with the original creators and platforms. Therefore, the released artifact will not include copies of videos or extracted raw frames. If source URLs are released, they will be provided only as an optional URL manifest for videos that remain publicly available and where such sharing is permitted by platform terms. Users should access source videos only through the original platforms or authorized interfaces, respect content removals and privacy-setting changes, and not redistribute downloaded copies.

Intended use and restrictions. The released artifacts are intended for research on multimodal feedback generation, exercise-form understanding, retrieval-augmented LVLM systems, and grounded

coaching evaluation. They are not intended for medical, rehabilitation, or professional training advice. Users should not use the artifacts to identify, profile, or contact individuals appearing in source videos, or to train systems for biometric identification, surveillance, profiling, or body-shape inference. The body-shape proxy is used only to retrieve visually comparable demonstrations and should not be interpreted as a demographic or identity attribute. We will provide a takedown contact for requests to remove source references or derived metadata associated with specific videos.

B Evaluation Metrics

Visual correction metrics. For each input frame x_i, the system produces a corrected frame $\hat{x}_i$ conditioned on target error labels E_i and retrieved expert references $R_i = \{r_{i1}, \ldots, r_{iK}\}$. We evaluate corrections with six metrics: human correction success, LLM correction success, CLIP-reference improvement, pose-reference improvement, appearance preservation, and background preservation.

Correction success. Human annotators judge whether the edited frame corrects each target error while preserving a plausible exercise pose:

$$s_{i,e}^{\text{human}} = \mathbb{1}[\text{error } e \text{ is judged corrected}],$$

$$\text{CS}^{\text{human}}(i) = \frac{1}{|E_i|} \sum_{e \in E_i} s_{i,e}^{\text{human}}.$$

The final score is averaged over all evaluated frames:

$$\text{CS}^{\text{human}} = \frac{1}{N} \sum_{i=1}^{N} \text{CS}^{\text{human}}(i).$$

When multiple annotators are used, we aggregate their binary judgments by majority vote. We compute LLM correction success analogously using GPT-5.5 as an automatic judge, with the input frame, corrected frame, and target error labels as input. Human correction success is our primary correction-quality metric, while LLM correction success is used as a scalable proxy.

CLIP-reference improvement. To measure whether the edited image becomes visually closer to correct-form demonstrations, we compute CLIP similarity to retrieved expert references. Let $\phi_{\text{CLIP}}(\cdot)$ denote the CLIP image encoder. We define the CLIP similarity between two images as

$$s_{\text{CLIP}}(a, b) = \cos(f_{\text{CLIP}}(a), f_{\text{CLIP}}(b)),$$

where f_{CLIP} denotes the CLIP image encoder. The reference-alignment score is then computed as

$$\text{CLIPRef}(z_i, R_i) = \frac{1}{K} \sum_{r \in R_i} s_{\text{CLIP}}(z_i, r).$$

We report the improvement after editing:

$$\Delta\text{CLIPRef}(i) = \text{CLIPRef}(\hat{x}_i, R_i) - \text{CLIPRef}(x_i, R_i).$$

Pose-reference improvement. We compute 2D pose distance using detected body keypoints normalized by hip center and torso length. Let $P(z)$ denote the normalized pose of image z, and let $\mathcal{J}$ be the set of visible joints shared by two images:

$$D_{\text{pose}}(a, b) = \frac{1}{|\mathcal{J}|} \sum_{j \in \mathcal{J}} \|P(a)_j - P(b)_j\|_2.$$

The distance from image z_i to retrieved references is

$$D_{\text{ref}}(z_i, R_i) = \frac{1}{K} \sum_{r \in R_i} D_{\text{pose}}(z_i, r).$$

We report the reduction in distance after editing:

$$\Delta\text{PoseRef}(i) = D_{\text{ref}}(x_i, R_i) - D_{\text{ref}}(\hat{x}_i, R_i).$$

Higher values indicate that the corrected pose moves closer to retrieved expert demonstrations.

Appearance and background preservation. Appearance preservation measures whether the corrected frame preserves the user's clothing and visual appearance while changing the target posture. We crop the person region from the input and corrected frames, mask the background when a person mask is available, and compute feature similarity:

$$\text{AppSim}(i) = \cos\left(\phi_{\text{app}}(\text{crop}(x_i)), \phi_{\text{app}}(\text{crop}(\hat{x}_i))\right).$$

Background preservation compares the non-person regions of the input and corrected frames. Given a dilated person mask M_i, the background mask is $B_i = 1 - \text{Dilate}(M_i)$:

$$\text{BGSim}(i) = \cos\left(\phi_{\text{bg}}(B_i \odot x_i), \phi_{\text{bg}}(B_i \odot \hat{x}_i)\right).$$

Both metrics compare the corrected frame against the original input frame, not the expert reference.

C Potential Risks, Limitations, and Responsible Use

Potential risks. EviCoach is a research framework for studying grounded multimodal exercise

feedback, not a substitute for professional coaching, medical advice, or rehabilitation guidance. An LVLM may misdiagnose an error, miss an unsafe movement, or generate plausible but inappropriate advice for a user's body, equipment, load, injury history, or training goal. Visual corrections may introduce artifacts, over-correct posture, or depict movements that look plausible but are unsafe to imitate. These risks are especially important for high-load exercises, users with injuries, and videos with occlusion, unusual viewpoints, poor lighting, loose clothing, or incomplete body visibility.

Limitations. Our benchmark covers 16 common strength-training exercises and focuses on visible form deviations in short exocentric videos. The taxonomy does not cover all exercises, populations, training goals, equipment setups, or injury-related constraints. EviCoach also depends on the coverage of the expert demonstration database; retrieval may be weak when the user's exercise type, equipment, viewpoint, motion phase, body-shape proxy, or error type is poorly represented. The system further relies on pretrained components for pose estimation, body-shape proxy estimation, LVLM diagnosis, retrieval, image editing, and LLM-based evaluation, so errors in any component can propagate to the final feedback.

Dataset and evaluation limitations. The source videos are public short-form fitness videos, which may over-represent certain camera styles, body types, exercise settings, creator demographics, and platform-specific presentation styles. Although we blur faces in figures and do not release raw videos or frames, the original source videos may still contain identifiable people, platform watermarks, or creator-specific context. Since we do not redistribute the original videos, exact reproduction of the raw visual dataset may be limited by link rot, content removal, or platform access changes. Our evaluation measures consistency with trainer-verified taxonomy labels and retrieved references, but does not guarantee biomechanical correctness, injury prevention, or long-term coaching effectiveness.

Responsible use. EviCoach outputs should be presented as assistive feedback with uncertainty, not authoritative exercise prescriptions. Systems based on this framework should disclose that generated advice may be incomplete or incorrect, encourage users to consult qualified coaches or medical professionals for high-risk exercises or injury-related concerns, and avoid confident feedback when the video is ambiguous or outside the taxonomy. Released artifacts should not be used to identify people in source videos, infer demographic or biometric attributes, or train systems for surveillance, profiling, or body-shape classification.

D Use of AI Assistants

AI systems are used in two ways in this work. First, large vision-language models and image-editing models are part of the proposed framework: they are used for taxonomy construction, exercise and equipment identification, error diagnosis, feedback generation, visual correction, and automatic evaluation, as described in the main paper. These uses are central to the design and are reported with model names, prompts, and evaluation protocols.

Second, the authors used AI assistants during the research and writing process for limited support, including brainstorming, code assistance, grammar polishing, LaTeX editing, and wording refinement. The authors reviewed and verified the resulting text, code, prompts, experimental setup, and scientific claims. AI assistants were not treated as factual sources; citations, model descriptions, experimental results, and responsible-research statements were checked by the authors.

E Human Annotation Protocol

We use human annotation for three parts of the study: expert taxonomy review, validation of taxonomy mappings, and evaluation of visual correction.

Expert taxonomy review. The initial taxonomy was generated with LVLM assistance and reviewed by an expert personal trainer with 4 years of strength-training coaching experience. For each exercise category, the trainer reviewed candidate error names and visual descriptions, removing categories that were uncommon, not visually identifiable from short exocentric videos, redundant, or not actionable for coaching. The trainer also edited names and descriptions for clarity.

Taxonomy mapping validation. For taxonomy coverage analysis, we collected 192 unique free-form error descriptions from paired correct/incorrect exercise videos. Each description was automatically matched to the nearest trainer-verified taxonomy entry within the same exercise category using CLIP text embeddings. Three annotators then validated these candidate mappings. For each

candidate pair, annotators saw the free-form description, the proposed taxonomy label, and the taxonomy label description, and marked the mapping as *valid*, *invalid*, or *uncertain*. A mapping was counted as covered only when it was marked *valid*; final coverage was computed by majority vote.

Visual correction evaluation. For visual correction evaluation, six annotators judged whether edited keyframes successfully corrected the target form error while preserving a plausible exercise pose. Each item contained the original user frame, the target error label(s), and the generated correction. Annotators focused on the target error rather than unrelated exercise-style differences. A correction was marked successful if the target error was visibly improved or fixed and the resulting pose remained physically plausible. It was marked unsuccessful if the target error remained, the edit was too weak or overly aggressive, the pose was implausible, or irrelevant visual changes made the feedback unreliable. We report majority-vote.

Annotation instructions. For visual correction success, annotators were instructed:

> Given the original exercise frame, the target form error, and the edited frame, decide whether the edited frame successfully corrects the target error. Mark *yes* if the target error is visibly corrected and the edited pose remains plausible. Mark *no* if the target error is still present, the correction is too weak, the correction is overly aggressive, or the edited pose appears unrealistic or unsafe. Ignore small visual artifacts unless they affect the correctness or reliability of the correction.

For taxonomy mapping validation, annotators were instructed:

> Given a free-form error description and a proposed taxonomy label, decide whether the taxonomy label captures the same coaching meaning. Mark *valid* if the taxonomy label would lead to substantially the same coaching feedback. Mark *invalid* if it describes a different error. Mark *uncertain* if the description is ambiguous or lacks enough information.

Annotators and ethics. Annotators were students with at least one year of fitness practice experience. They were all within the research institute and participated voluntarily. They were informed that their labels would be used for research on multimodal exercise feedback generation. The annotation task involved judging exercise videos or edited frames collected from public online sources; annotators did not perform exercises themselves and were not asked to provide personal health or fitness information.

F Prompt Templates

We provide the main prompt templates used in EviCoach. Curly brackets denote variables filled at inference time, such as sampled frames, retrieved references, taxonomy entries, and target error labels. We omit API-specific wrappers and formatting tokens that do not affect model behavior.

F.1 Taxonomy-Constrained Diagnosis

```
You are an exercise-form analysis assistant.

Input:
- Exercise: {exercise_name}
- Equipment: {equipment_type}
- User frames: {8_sampled_user_frames}
- Retrieved correct-form references: {
    retrieved_reference_frames}
- Error taxonomy:
{taxonomy_entries}

Task:
Identify which taxonomy errors are visible in
    the user sequence.
Use the retrieved references as correct-form
    visual evidence.
Only select errors from the provided taxonomy.
Do not invent new error categories.
If the evidence is weak or ambiguous, do not
    select the error.

Output JSON:
{
  "exercise": "...",
  "equipment": "...",
  "present_errors": [
    {
      "error_name": "...",
      "evidence": "brief observable visual cue
    ",
      "frame_indices": [...]
    }
  ],
  "absent_or_uncertain_errors": [...]
}
```

F.2 Free-form diagnosis baseline.

```
You are an exercise-form analysis assistant.

Input:
- Exercise: {exercise_name}
- Equipment: {equipment_type}
- User frames: {8_sampled_user_frames}
```

```
- Retrieved correct-form references: {
    retrieved_reference_frames}

Task:
Describe the visible form errors in the user
    sequence relative to the correct-form
    references.
Use short, concrete error descriptions.
Do not use a predefined taxonomy.

Output JSON:
{
  "free_form_errors": [
    {
      "description": "...",
      "evidence": "brief observable visual cue
    ",
      "frame_indices": [...]
    }
  ]
}
```

F.3 Reference-grounded coaching feedback.

```
Input:
- Exercise: {exercise_name}
- Target errors: {diagnosed_error_labels}
- User frames: {sampled_user_frames}
- Retrieved correct-form references: {
    retrieved_reference_frames}
- Taxonomy descriptions: {error_descriptions}

Task:
Generate concise coaching feedback grounded in
     the observed user frames and retrieved
    references.
For each error, include:
1. what is wrong,
2. why it matters for form,
3. how to correct it,
4. what to look for in the references.

Avoid medical claims. If the visual evidence
    is ambiguous, state uncertainty.

Output JSON:
{
  "detected_issues": [...],
  "coaching_cues": [...],
  "reference_grounded_explanation": "...",
  "uncertainties": [...]
}
```

F.4 Frame-specific correction instruction.

```
Input:
- Exercise: {exercise_name}
- Current motion phase: {
    phase_index_or_description}
- Target error labels: {target_errors}
- User frame: {user_frame}
- Retrieved correct-form references: {
    phase_aligned_references}

Task:
Write a concise image-editing instruction that
     corrects only the target form errors in
    this frame.
Use the retrieved references as examples of
    proper form.
Preserve the person's appearance, clothing,
    equipment, camera view, and background.
Do not change unrelated body parts unless
    needed for the correction.
Do not add text, arrows, labels, or new
    objects.

Output:
{
  "edit_instruction": "..."
}
```

F.5 Verifier-editor refinement.

```
Input:
- Original user frame: {user_frame}
- Edited frame: {edited_frame}
- Target error labels: {target_errors}
- Retrieved correct-form references: {
    references}
- Previous edit instruction: {
    previous_instruction}

Task:
Check whether the edited frame correctly fixes
     the target errors while preserving a
    plausible exercise pose and the original
    scene.

Classify the edit as one of:
- "success": target errors are corrected and
    the pose is plausible.
- "under_corrected": target errors remain or
    the edit is too weak.
- "over_corrected": the edit changes the pose
    too much or creates an unnatural form.
- "artifact_or_drift": the edit changes
    clothing, equipment, background, identity/
    appearance, or adds artifacts.

If the edit is not successful, write a revised
     instruction.
Use a stronger instruction for under-
    correction and a softer or more
    constrained instruction for over-
    correction or drift.

Output JSON:
{
  "status": "...",
  "brief_evidence": "...",
  "revise": true/false,
  "revised_instruction": "..."
}
```

F.6 LLM correction success judge.

```
You are evaluating an edited exercise frame.

Input:
- Original frame: {input_frame}
- Edited frame: {corrected_frame}
- Target error labels: {target_errors}
- Error descriptions: {
    target_error_descriptions}
```

```
Task:
For each target error, decide whether the
     edited frame successfully corrects the
     error.
Mark "yes" only if the target error is visibly
      improved or fixed and the resulting pose
     remains physically plausible.
Mark "no" if the error remains, the correction
      is too weak, the correction is overly
     aggressive, the pose is unrealistic or
     unsafe-looking, or unrelated visual
     changes make the edit unreliable.
Ignore minor image artifacts unless they
     affect the correction judgment.

Output JSON:
{
  "per_error_judgment": [
    {
      "error_name": "...",
      "corrected": true/false,
      "brief_reason": "..."
    }
  ],
  "overall_success": true/false
}
```