Design Diagrams

- System Link: https://drive.google.com/file/d/1SMz2ORXjo8QXA-t7hJci-xpY6J-DbxRy/view?usp=sharing

- Container Link: https://drive.google.com/file/d/1IweWPpec4t-YwbWyhsoreErOy_DWmA7t/view?usp=sharing

- Component Link: https://drive.google.com/file/d/1BrN6rEG8mW-B3XhU6sVv-D-wOdgCdqmO/view?usp=sharing
- Tech Stack:
    - Frontend
        - React Native
        - HTML/CSS
    - Backend
        - AWS Lambda
        - AWS Fargate
        - AWS DynamoDB (NoSQL)
        - AWS S3
        - Python (Coding Language)
    - AI Component
        - Open AI GPT 4o-mini
        - Reinforcement Learning based on Human Feedback
    - Computer Vision Component
        - OpenCV
        - Google Media Pipe

Figma

- Figma Link: https://www.figma.com/design/CnSOLDBJ7UpBWs0WVctipw/Combined?t=IiZWShdWIErwvi6k-1

Design Trade

1. **Adaptive Workout Recommendation Engine**
    a. **Design Alternatives**
        i. Rule-Based Heuristic Engine: This approach uses predefined if–then rules to map past workout performance and form scores to future workout

recommendations. It is fast, deterministic, and highly explainable, but struggles to adapt to complex or unusual user progress patterns.

ii. Machine Learning (ML) Regression Model: This design uses a trained regression model to predict appropriate workout difficulty based on historical performance data. While it adapts over time and improves personalization, it requires model training and sacrifices some interpretability and response speed.

iii. Hybrid ML + Rule-Based Engine: This option combines rule-based logic for safety constraints and edge cases with an ML model for typical progression decisions. It provides strong personalization while preserving explainability and robustness at the cost of increased implementation complexity.

b. **Trade Study Evaluation:**

Criterion	Weight	Rule-Based	ML Regression	Hybrid
Accuracy	30	6	8	**9**
Computation Speed	25	9	7	**8**
Scalability	15	6	8	**9**
Ease of Integration	15	9	7	**8**
Explainability	15	10	5	**8**
Weighted Total	100	7.3	7.1	**8.4**

c. **Selected Design and Trade-Off Analysis: The Hybrid ML + Rule-Based Engine** was selected because it achieved the highest weighted score and best balanced accuracy, scalability, and explainability. By using rules for safety limits and edge cases while relying on machine learning for adaptive progression, the system delivers personalized recommendations without sacrificing transparency. Although this approach introduces greater implementation and maintenance complexity compared to a purely rule-based system, the gains in recommendation quality and long-term scalability justify the trade-off. This design also integrates cleanly with the backend database, CV subsystem, and AI chatbot, supporting both functional and non-functional system requirements.

2. **Contextualization Conversational Interface for Goals and Plan Setting**

a. **Design Alternatives:**

i. External LLM Chatbot: This approach uses a cloud-hosted large language model (e.g., OpenAI GPT-4 or Claude 3.5) accessed through secure API calls from the FastAPI backend. It provides highly accurate and natural conversational interactions but introduces recurring API costs and potential latency under heavy load.

ii. Local Small Model Interface: This design runs a lightweight, locally hosted LLM (e.g., Phi-3-mini or LLaMA 3 8B) on a server or local GPU to handle all conversational interactions. While it improves data privacy and reduces operational costs, it offers lower language understanding quality and requires additional infrastructure maintenance.

iii. Hybrid LLM Architecture: This option combines a local lightweight model for simple interactions (greetings, metric collection) with a cloud-based LLM for complex goal interpretation and contextual reasoning. It balances conversational quality, latency, and cost, but introduces additional system complexity through routing logic.

b. **Trade Study Evaluation:**

Criterion	**Weight**	**External LLM Chatbot**	**Local Small Model Interface**	**Hybrid LLM Architecture**
Response Latency	20%	7	9	8
Integration Complexity	15%	9	6	7
User Experience	25%	9	7	9
Personalization Accuracy	25%	10	7	9
Cost / API Efficiency	15%	5	9	8
Weighted total	**100**	**8.2**	**7.5**	**8.5**

c. **Selected Design and Trade-Off Analysis: The Hybrid LLM Architecture** was selected because it achieved the highest overall weighted score and best satisfied the competing demands of accuracy, responsiveness, and cost efficiency. By delegating simple, time-sensitive interactions to a local model and reserving cloud LLM calls for deeper contextual reasoning, the system maintains a natural conversational flow while minimizing latency and API expenses. Although this design introduces additional architectural complexity due to routing and fallback logic, the improved user experience and scalable cost structure outweigh the implementation overhead. This approach fully supports FR-2 by enabling accurate, personalized plan generation while ensuring seamless integration with the backend database and fitness-plan generation modules.

3. **User Fitness Dashboard**
   a. **Design Alternatives**
      i. Card-Based Modular Subsystem Design: A single-page dashboard where each subsystem (workout snapshot, progress visualizations, AI chat, profile access) is represented as an individual card, allowing centralized viewing and navigation from one screen.

      ii. Mock Web Browser Design: A tab-based interface that mimics a traditional web browser, where each subsystem opens in its own tab and loads independently.

      iii. Deterministic AI Response Design: A fully AI-driven interaction model where users access subsystems and functionality exclusively through conversational prompts to an AI coach.

   b. **Trade Study Evaluation**

	Usability	Ease of Implementation	Performance	Reliability	Weighted Total
Card Based Modular Subsystem Design	8.5	6.5	7.0	7.0	7.525
Mock Web Browser Design	9.0	6.0	5.0	8.0	7.35
Deterministic AI Response Design	6.0	4.0	5.5	3.0	4.975

   c. **Selected Design and Trade-Off Analysis**

   The Card-Based Modular Subsystem Design was selected as the superior solution because it achieved the highest weighted score and best balanced usability, performance, and reliability. While the mock web browser design benefitted from strong user familiarity, it introduced performance risks due to independent loading and increased navigation overhead. The AI-driven design, although innovative, posed significant reliability and usability risks due to AI hallucinations and higher implementation complexity. Overall, the card-based approach best supports quick access, centralized interaction, and low-latency transitions, aligning closely with the dashboard's role as the primary entry point to the system.

4. CV Form Analysis Subsystem

a. Design Alternatives

i. Full Cloud Raw Video: Upload the full raw workout video to the cloud where the backend runs pose estimation and temporal/fault analysis end-to-end using heavier, high-accuracy models. This maximizes precision and robustness but increases privacy risk and ongoing compute/storage cost because raw video must be handled server-side.

ii. Downsampled/Trimmed Video to Cloud: Upload a reduced version of the video (downsampled frames, shorter clips, or trimmed segments) to cut bandwidth and inference cost while still running full analysis in the cloud. This improves throughput and cost relative to Alt A, but accuracy can drop and it still requires sending raw footage (even if reduced).

iii. Two-Stage Cloud Pipeline: Upload raw video only long enough to extract pose keypoints/motion features, then permanently delete the raw footage and perform rep segmentation, fault detection, and scoring from the keypoints. This preserves near-best accuracy and robustness while substantially improving privacy and keeping costs scalable through a modular pipeline.

b. Trade Study Evaluation

Criterion Weight	Weight	Alt A: Full Cloud Raw Video	Alt B: Downsampled Video	Alt C: Two-Stage Cloud Pip
Pose & Fault Accuracy	**25**	**9**	**8**	**9**
Privacy & Data Minimization	**20**	**6**	**7**	**9**
Infrastructure Cost	**15**	**6**	**8**	**8**
Batch Throughput	**10**	**7**	**8**	**8**
Robustness	**10**	**9**	**8**	**9**
Maintainability	**10**	**8**	**8**	**9**
Upload Friendliness	**5**	**6**	**8**	**7**

Platform Integration	**3**	**8**	**8**	**8**
Footprint	**2**	**10**	**10**	**9**
Weighted Total Score	**100%**	**74.2%**	**78.8%**	**82.0%**

c. Selected Design and Trade-Off Analysis

Alternative C achieved the best overall balance across the project’s priorities. It maintained high pose/fault accuracy and robustness comparable to the fully cloud-based approach, while significantly improving privacy by minimizing raw video retention and shifting downstream analysis to keypoints. Compared to Alt A, it reduces exposure of identifiable user footage and lowers long-term storage burdens, directly supporting the data-minimization and cost constraints. Compared to Alt B, it avoids the accuracy loss associated with aggressive downsampling while still controlling compute and throughput. The main trade-off is added backend complexity, but the modular design improves maintainability and makes future exercise/fault additions easier without requiring mobile app changes.