DiveDoc: Structured Diagnosis of Diving Performance via Pose-Guided Action Parsing

Youbo Shao*
Tsinghua University
shao-yb23@mails.tsinghua.edu.cn

Zhuo Lin*
Tsinghua University
lin-z23@mails.tsinghua.edu.cn *

Abstract

*Video-based action quality assessment (AQA) is critical in both sports and fitness, yet many existing methods lack structured feedback and semantic interpretability. We propose **DiveDoc**, a pose-guided diagnostic framework that provides interpretable feedback on diving performance. Leveraging curated subsets of FineDiving and FineDiving-HM, DiveDoc integrates pose-aware encoding, semantic step segmentation, and contextual refinement to capture fine-grained action phases. It scores each modality and synthesizes the results into natural language feedback using a Large Language Model (LLM). Experiments show that DiveDoc outperforms the FineParser framework in both parsing and scoring accuracy, offering new insights into interpretable, multimodal AQA systems.*

1. Introduction

In recent years, video-based motion analysis has drawn growing interest from both the world of elite sports and everyday fitness. In competitive arenas like the Olympics and Winter Games, accurate assessment of athletic performance is essential—not only for scoring but also for understanding subtle movements that can separate champions from the rest. At the same time, ordinary users are turning to smart fitness platforms, such as Keep or Xiaomi Health, to track their workouts, improve posture, or recover from injuries with the help of visual diagnostics. Across both settings, there is a rising demand for systems that move beyond simple numerical scores and offer interpretable, context-aware feedback.

Meeting this demand requires models that understand not just how well an action is performed, but how it unfolds over time. Unlike classification tasks, movement diagnosis calls for reasoning over sequences of semantically distinct sub-actions—e.g., the takeoff, flight, and entry in diving—where subtle deviations in one phase can cascade into visible errors later. However, most existing vision-based systems rely on short clips or global video embeddings [8, 9, 13], limiting their ability to capture the internal temporal structure of actions. Recent work such as FineDiving[15] and FineParser[16] has highlighted the importance of procedure-aware analysis, where actions are parsed into consecutive steps to improve interpretability and alignment. Still, these systems primarily focus on step-level scoring and often lack explicit semantic grounding or diagnostic reasoning. Bridging this gap calls for fine-grained, human-centric action parsing—a foundation not only for scoring, but also for explaining why a mistake happened and how to correct it.

To address the need for interpretable, structured feedback in diving performance analysis, we present DiveDoc, a pose-guided diagnostic system that refines and extends fine-grained action parsing for human-centric understanding. Built upon the curated Fine Diving and FineDiving-HM datasets, DiveDoc is a structured diagnostic system designed to analyze complex diving actions and provide interpretable feedback. It consists of five main components: (1) a pose-aware visual encoder (PVE); (2) a semantic step segmenter (SSS); (3) a contextual visual refiner (CVR); (4) a scoring aggregator (SA); and (5) a diagnosis generator based on large language models (LLMs). Given a pair of query and exemplar videos, PVE first encodes each frame by integrating multi-scale visual features with 2D joint pose information. This enables the model to localize and emphasize human-centric action regions such as limb trajectories and body orientation, enhancing the reliability of spatial parsing. Next, SSS parses the encoded sequence into semantically meaningful sub-action steps—such as takeoff, flight, and entry—by identifying phase transitions via learned temporal attention. These steps allow finer alignment across actions with similar intent but differing motion speeds or styles. The contextual visual refiner (CVR) further enriches each step's representation by capturing static scene context and non-local cues, supporting robust matching under visual variance. The scoring aggregator (SA) then performs pairwise comparison between query and exemplar steps to estimate multi-aspect quality scores

** Equal contribution.

for each stage. Finally, all step-level scores, auxiliary metadata, and sub-action descriptors are fed into a large language model (LLM), which synthesizes the information into a natural language diagnosis. This diagnosis explains the strengths and weaknesses of the input performance, highlighting error sources and offering actionable suggestions—transforming DiveDoc from a scoring system into an intelligent assistant for movement analysis.

The contributions of this paper are summarized as follows:

- We present **DiveDoc**, a structured and interpretable diagnostic system for diving performance analysis, which integrates pose information, semantic segmentation, and language-based reasoning into a unified pipeline.
- We construct a refined subset of the FineDiving and FineDiving-HM dataset by selecting the most frequent and semantically consistent dive types, enriching each sample with synchronized pose annotations to enable robust pose-aware learning.
- We introduce a large language model-based diagnostic generator that transforms step-level action scores and metadata into natural language feedback, enabling interpretable, actionable, and human-aligned explanations of dive quality.

2. Related Work

2.1. Action Quality Assessment

Action Quality Assessment (AQA) aims to evaluate the execution quality of actions, particularly in sports contexts. Early approaches relied on handcrafted features and regression models [8, 9], which struggled to capture the nuanced dynamics of complex actions. Recent advancements have introduced deep learning techniques to address these challenges. Group-aware Contrastive Regression (CoRe) [20] reformulates AQA as a relative scoring task, leveraging pairwise comparisons to highlight subtle differences between performances. Multi-Stage Contrastive Regression (MCoRe) [1] segments actions into procedural stages, applying contrastive learning at each stage to enhance discriminative capability. Temporal Parsing Transformer [2] decomposes actions into temporal parts using learnable queries, facilitating fine-grained assessment. Tang et al. [13] introduced Uncertainty-aware Score Distribution Learning (USDL), which models the inherent ambiguity in judge scores by learning a probability distribution over possible scores, rather than predicting a single deterministic value. FineParser [16] introduces FineReg, a contrastive regression module aligning exemplar and query sequences for interpretable scoring. Many methods overlook or underuse pose information, limiting sensitivity to subtle motion nuances. DiveDoc addresses this by incorporating structured pose cues for finer performance discrimination.

2.2. Pose Estimation Methods

Accurate human pose estimation is crucial for fine-grained action understanding. Traditional methods like OpenPose [3] and HRNet [12] have laid the groundwork, with OpenPose introducing Part Affinity Fields for real-time multi-person 2D pose estimation, and HRNet maintaining high-resolution representations throughout the network. MoveNet [14] offers a lightweight and fast solution suitable for mobile applications but may lack the precision required for complex actions. Recent advancements have leveraged Vision Transformers (ViTs) for pose estimation.ViTPose [18] uses a Vision Transformer for effective keypoint detection, offering scalability and flexibility. ViTPose++ [19] improves this with multi-scale features and better training, achieving strong results on benchmarks like MS COCO and MPII. Thus we adopt ViTPose++ in DiveDoc to extract pose features for a detailed understanding of diving actions.

2.3. Fine-grained Sports Datasets

Fine-grained sports datasets are essential for advancing detailed action understanding and quality assessment. FineGym [11] provides hierarchical annotations for gymnastics performances. MultiSports [6] offers spatio-temporal localization across various sports with dense annotations. BASKET [7] is a large-scale basketball video dataset focusing on fine-grained skill estimation. VideoBadminton [5] introduces fine-grained action categories in badminton. FineSports [17] is a multi-person hierarchical sports video dataset designed for fine-grained action understanding. In competitive diving, the FineDiving dataset [15] contains 3,000 videos with over 312,000 human-centric foreground masks, along with detailed annotations of action procedures. FineDiving-HM [16] further extends this by labeling divers' masks in the images of the FineDiving dataset. However, pose information is typically missing. Our dataset augments FineDiving and FineDiving-HM with reliable keypoint data to support pose-guided parsing and scoring.

3. Method

To generate structured diagnostic feedback for complex diving actions, we design **DiveDoc**, a pose-guided analysis framework that integrates spatial-temporal parsing with language-based reasoning. The system is trained to understand performance at both a visual and semantic level, enabling it to identify sub-action quality and articulate actionable feedback.

As shown in Figure 1, DiveDoc takes as input a pair of query and exemplar diving videos and processes them through a sequence of modular stages. We begin by extracting 2D pose sequences from each video, which are further encoded into temporal pose embeddings designed to highlight joint-level motion patterns across time. These

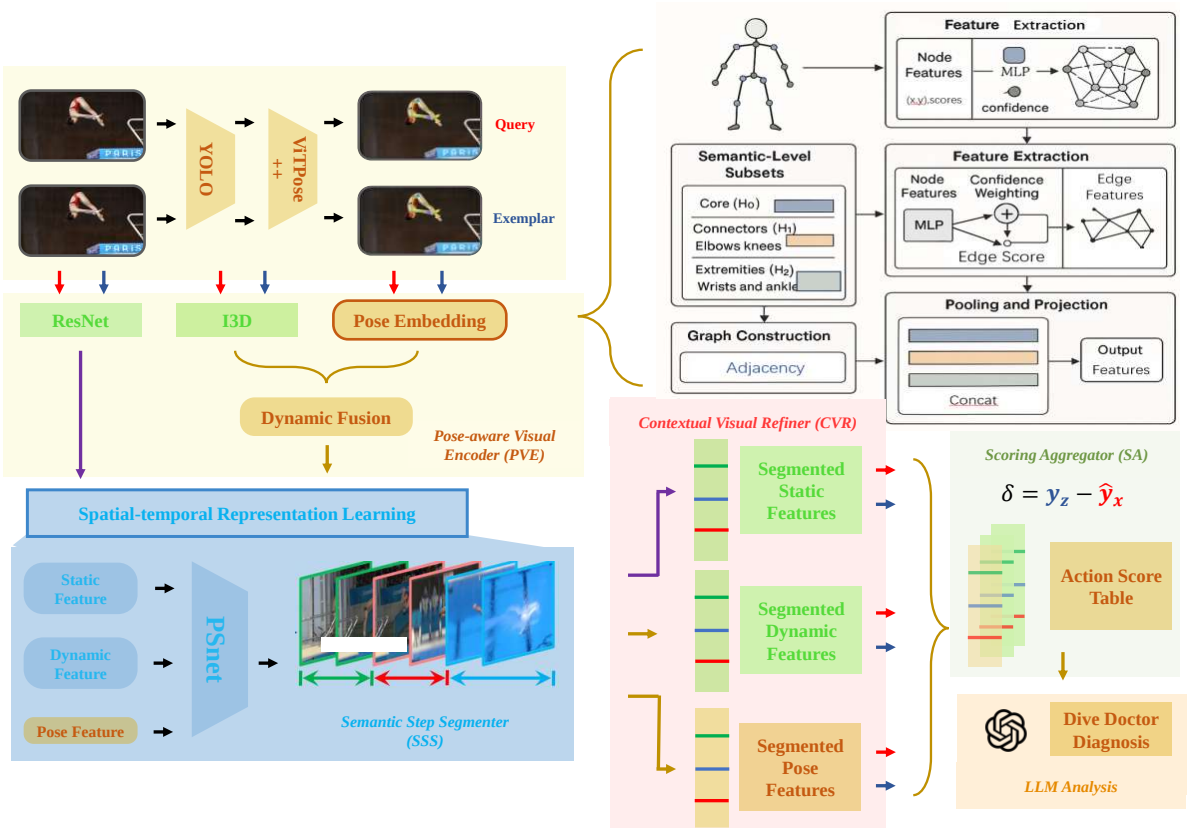

Figure 1. The architecture of the proposed **DiveDoc**. Given a pair of query and exemplar videos, we extract pose, static, and dynamic features via a combination of YOLOViT, ResNet, and I3D. These features are fused and segmented using PSNet into fine-grained temporal steps. For each step, a triplet of quality scores is computed across the three modalities and organized into an *Action Score Table*, termed the **DiveScoreCard**. This structured representation is passed to a large language model to generate interpretable diagnostic feedback.

embeddings are fused with RGB features through a spatial encoder to enhance action region representation. A temporal segmenter then decomposes the video into semantically meaningful sub-actions, enabling fine-grained comparisons across steps. Based on these aligned segments, a scoring module evaluates local quality, and the results are passed into a large language model (LLM), which produces interpretable natural language diagnoses.

Each component in this pipeline—pose embedding extraction, spatial-temporal representation learning, sub-action alignment, and feedback generation—is detailed in the following sections.

3.1. DiveDoc Framework

3.1.1. Pose Representation Learning

One key challenge in leveraging pose information is the uncertainty of the predicted keypoints, since they are generated by a model and may include noisy or inaccurate estimates. Fortunately, modern pose estimators such as ViTPose++ provide a confidence score for each keypoint, reflecting the reliability of the prediction. To take advantage of this, our pose embedding strategy jointly encodes both the keypoint coordinates and their confidence scores. We propose and compare three types of pose encoders that transform these inputs into compact 128-dimensional feature vectors.

NaiveMLP. This encoder treats the 17 keypoints as a flat vector of concatenated (x, y) coordinates and confidence scores, resulting in a (17×3)-dimensional input. After normalizing the coordinates to $[0, 1]$, the vector is passed through a two-layer MLP with a ReLU activation in between. The output is a single 128-dimensional embedding that summarizes the entire pose without modeling any spatial structure. This encoder is computationally efficient and serves as a strong baseline.

WeightedMLP. Similar to NaiveMLP, this encoder also takes the normalized (x, y) coordinates and confidence scores of 17 keypoints as input. However, instead of flattening the features directly, each keypoint is first processed independently through a shared MLP. The resulting features are then weighted by their corresponding confidence scores, allowing the encoder to emphasize reliable joints and suppress noisy ones. The weighted

features are flattened and projected through a final linear layer to produce the final embedding. This approach adds a level of robustness to the input quality without modeling spatial relationships.

HierarchicalEncoder. Inspired by [10], this encoder models the hierarchical structure of the human body. The 17 keypoints are grouped into three semantic subsets: **Core (H0)** including hips and shoulders, **Connectors (H1)** including elbows and knees, and **Extremities (H2)** including wrists and ankles. For each level, two adjacency graphs are constructed: a fully connected graph within the level (physical adjacency) and another connecting it to neighboring levels (cross-level adjacency). Node features are obtained by applying an MLP to the (x, y) coordinates and confidence scores, followed by confidence weighting. Edge features are computed using a fully vectorized EdgeConv operation, which aggregates differences between connected nodes. Node and edge features are pooled across each level and concatenated, and a final projection layer produces the 128-dimensional output. This encoder captures spatial dependencies and joint-level confidence hierarchically, yielding a rich representation of the pose structure.

Together, these encoders allow us to explore the trade-offs between architectural simplicity, confidence weighting, and structured spatial modeling in the context of pose-guided action quality assessment. An interesting finding during our development was the impact of implementation style on efficiency. Initially, our encoders were written in a straightforward but naïve manner using multiple for-loops, which resulted in extremely low GPU utilization and unacceptably long training times. We later realized the importance of adopting a fully vectorized coding style to take advantage of GPU parallelism. After refactoring our code accordingly, we observed a dramatic speed-up: the encoder's processing time improved by several hundred times, and overall training time was reduced by an order of magnitude.

3.1.2. Fine-grained Spatio-temporal Representation Parser

Given a query video X and an exemplar video Z, we aim to extract aligned, interpretable representations that capture the fine-grained spatio-temporal structure of diving actions. Our parser integrates pose, motion, and static appearance information, and consists of three major modules: a spatial encoder, a semantic step segmenter (SSS), and a contextual visual refiner (CVR). The process begins with frame-wise feature extraction and culminates in temporally segmented, multi-modal embeddings ready for scoring.

Spatial Encoder.

We follow FineParser in adopting an I3D-based backbone $B(\cdot)$ to encode visual snippets into feature maps. Transposed convolutions are inserted before pooling layers to preserve spatial resolution. For each video $X = \{X_i\}_{i=1}^{N}$, we compute:

$$\mathbf{M}_i^{(j)} = \text{Upsample}(B_j(X_i)), \quad j = 1, 2, 3, 4$$

where B_j denotes the j-th stage feature block. A fused mask is obtained as:

$$\mathbf{M}_i^{\text{fuse}} = \text{Conv3d}(\text{Concat}(\{\mathbf{M}_i^{(j)}\}))$$

The final visual embedding is gated by this mask:

$$\mathbf{X}_V = B(X) \odot \sigma(\mathbf{M}^{\text{fuse}})$$

where σ is the sigmoid function.

Semantic Step Segmenter (SSS).

To enable fine-grained temporal parsing of the action sequence, we introduce a semantic step segmenter that decomposes the video into interpretable sub-actions. The segmenter is adapted from PSNet and jointly attends to dynamic visual features and pose embeddings.

Given the fused features $\mathbf{X}_V$ and pose embeddings $\mathbf{P}$, we concatenate them along the channel dimension and feed them into a temporal transition prediction head $S(\cdot)$ that estimates the probability of each frame being a step transition point for each sub-action k:

$$S(\mathbf{X}_V, \mathbf{P})[t, k] \in \mathbb{R}$$

To locate the actual transition timestamp $\hat{t}_k$, we perform a local argmax search within a predefined temporal window $(T_{k-1}, T_k]$:

$$\hat{t}_k = \arg \max_{t \in (T_{k-1}, T_k]} S(\mathbf{X}_V, \mathbf{P})[t, k]$$

This process segments the video into $L' + 1$ consecutive steps $\{(\hat{t}_{l-1}, \hat{t}_l)\}_{l=1}^{L'+1}$, enabling downstream modules to operate on aligned semantic phases such as take-off, flight, or entry. Compared to prior video-level approaches, our segmenter preserves step boundaries and supports localized quality analysis.

Contextual Visual Refiner (CVR).

To enhance each sub-action segment with contextual appearance cues, we apply a ResNet-34 encoder $T(\cdot)$ to RGB frames and perform average pooling within each step segment:

$$X_S^{(l)} = \text{Proj}(T(X)[\hat{t}_{l-1} : \hat{t}_l])$$

where $\text{Proj}(\cdot)$ is a learned projection layer. This results in a static embedding $X_S^{(l)}$ for each step l.

By the end of this parsing stage, we obtain step-wise multi-modal features: dynamic ($\mathbf{X}_V^{(l)}$), static ($X_S^{(l)}$), and pose ($\mathbf{P}^{(l)}$), which are jointly used for scoring and diagnostic generation.

3.1.3. Diagnosis-Oriented Scoring and Feedback Generation

To support interpretable diagnosis, we design a modular scoring pipeline that produces independent quality estimates from three feature modalities—dynamic, static, and pose—at each semantic step. These scores form the basis for the DiveScoreCard, a structured diagnostic representation that is later converted to natural language feedback by a large language model.

Per-modality Scoring.

For each segmented sub-action step $l \in \{1, ..., L'+1\}$, we use three modality-specific encoders to produce individual quality scores:

$$\Delta y_l^{\text{dyn}} = R_{\text{dyn}}(\mathbf{X}_V^{(l)}), \quad \Delta y_l^{\text{stat}} = R_{\text{stat}}(X_S^{(l)}),$$
$$\Delta y_l^{\text{pose}} = R_{\text{pose}}(\mathbf{P}^{(l)})$$

where R_{dyn}, R_{stat}, and R_{pose} are separate MLPs trained to regress the quality of motion, appearance, and body configuration respectively.

The final predicted score is computed as a weighted sum of step-level scores and the reference score from the exemplar:

$$\hat{y}_X = \sum_{l=1}^{L'+1} \lambda_l \left(\Delta y_l^{\text{dyn}} + \Delta y_l^{\text{stat}} + \Delta y_l^{\text{pose}} \right) + y_Z$$

where λ_l is a learnable weight for the l-th step and y_Z is the ground-truth score of the exemplar video.

DiveScoreCard and LLM Diagnosis.

Rather than outputting a single scalar score, we organize the per-step, per-modality scores into a structured diagnostic table called the **DiveScoreCard**. This table includes, for each semantic step, the estimated quality along the dynamic, static, and pose dimensions. Formally, the DiveScoreCard is represented as:

$$\text{DiveScoreCard} = \left\{ \left(\Delta y_l^{\text{dyn}}, \Delta y_l^{\text{stat}}, \Delta y_l^{\text{pose}} \right) \right\}_{l=1}^{L'+1}$$

This interpretable score matrix, together with auxiliary metadata (e.g., dive type, difficulty degree), is used to prompt a large language model. Our prompt describes the scoring mechanism in natural language and presents exemplar-based deltas across modalities. For each query video, the prompt includes the example score (from the best exemplar), per-modality deltas, and the final predicted score. The LLM is then asked to identify potential weaknesses in each modality and suggest targeted improvements. This design allows the system to emulate the reasoning process of an expert coach and produce segment-level feedback grounded in quantitative assessment.

3.2. Training and Inference

Training Objective.

Given a pair of query and exemplar videos (X, Z), the DiveDoc framework is optimized by minimizing a composite loss function over three modules:

$$\mathcal{L} = \mathcal{L}_{\text{PVE}} + \mathcal{L}_{\text{SSS}} + \mathcal{L}_{\text{Score}}$$

where: - $\mathcal{L}_{\text{PVE}}$ supervises the prediction of spatial attention masks from the pose-aware visual encoder; - $\mathcal{L}_{\text{SSS}}$ supervises temporal segmentation of sub-action phases by the semantic step segmenter; - $\mathcal{L}_{\text{Score}}$ supervises per-modality regression for each semantic step.

The mask loss $\mathcal{L}_{\text{PVE}}$ is a focal loss [17] defined as:

$$\mathcal{L}_{\text{PVE}} = \sum \text{FocalLoss}(p(\mathbf{M}_{j,i}))$$

where $p(\mathbf{M}_{j,i})$ represents the predicted mask probability for pixel (j, i).

The segmentation loss $\mathcal{L}_{\text{SSS}}$ encourages accurate step transition prediction and is given by:

$$\mathcal{L}_{\text{SSS}} = -\sum_t \Big(p_k(t) \log S(\mathbf{X}_V, \mathbf{P})[t, k] + (1 - p_k(t)) \log(1 - S(\mathbf{X}_V, \mathbf{P})[t, k]) \Big)$$

where $p_k(t)$ is a binary indicator function denoting whether frame t corresponds to the k-th sub-action transition.

The scoring loss $\mathcal{L}_{\text{Score}}$ is defined as the mean squared error between the predicted score $\hat{y}_X$ and ground-truth score y_X:

$$\mathcal{L}_{\text{Score}} = \|\hat{y}_X - y_X\|_2^2$$

Inference Procedure.

During inference, given a query video X, we select E exemplar videos $\{Z_j\}_{j=1}^E$ from the training set that share the same action type. For each query-exemplar pair (X, Z_j), DiveDoc computes per-modality step-wise deltas and predicts a relative score. The final prediction is obtained through multi-exemplar voting:

$$\hat{y}_X = \frac{1}{E} \sum_{j=1}^{E} \left(F(X, Z_j, y_{Z_j}; \Theta) + y_{Z_j} \right)$$

where $F(\cdot)$ is the DiveDoc inference function and Θ represents all learnable parameters.

This inference pipeline ensures that temporal parsing, modality-specific deltas, and the DiveScoreCard are used in tandem to generate both accurate numeric predictions and interpretable diagnostic feedback.

4. Experiments

4.1. Experimental Setup

Dataset.

We observed that the original FineDiving dataset exhibits a significant long-tailed distribution. Among the 3,000

videos spanning 52 action types, more than half of the action types have fewer than 25 videos, and one-third of the action types have five or fewer videos. This imbalance negatively impacts the effectiveness of the FineParser framework and slows down the training process, as it requires pairwise comparisons between query and example data. To address this issue and reduce computational overhead, we selected the top 5 action types with the highest occurrence counts in the dataset. This selection retains nearly half (1,317 out of 3,000) of the original data while significantly accelerating the training speed.

Pose Extraction Pipeline.

To build the DiveDoc dataset, we utilize a pose estimation pipeline to analyze human poses in competitive diving videos. The pipeline processes frames from the FineDiving[15] and FineDiving-HM[16] datasets, incorporating mask information to isolate the diver's position. It integrates object detection and pose estimation models to extract keypoints and their corresponding confidence scores for each frame. The key steps in the pose analysis process are as follows:

- **Masking and Preprocessing:** To focus exclusively on the diver, a binary mask from FineDiving-HM is applied to each frame in FineDiving, setting pixels outside the mask to zero. This ensures that only the diver's region is processed.
- **Object Detection:** We utilize the YOLOv8n [4] model to detect bounding boxes for the "person" class in each frame. The model is configured with a confidence threshold of 0.1 and limited to one detection per frame. This threshold is carefully chosen to ensure the diver is consistently detected before entering the water and ignored once submerged.
- **Pose Estimation:** For each detected bounding box, we employ the ViTPose++-huge model [19], one of the most advanced keypoint detection models, to estimate 17 keypoints corresponding to human body joints (e.g., nose, shoulders, elbows, wrists, hips, knees, and ankles). Each keypoint is accompanied by a confidence score indicating the reliability of the detection. The results for each frame are stored as either None (if no diver is detected) or a dictionary containing the keypoints and their confidence scores.

After processing through this pipeline, each video is transformed into a structured dataset containing per-frame keypoints, confidence scores, and bounding boxes, which serve as the foundation for subsequent tasks. Processing over 300,000 frames takes approximately five hours, highlighting the extensive data preparation required for our experiments.

Implementation Details.

We implement our framework based on the official FineParser codebase, using PyTorch, FineDiving and FineDiving-HM dataset. All experiments are conducted on a single NVIDIA A100 GPU, and each run takes approximately 8 hours to complete.

We adopt I3D pre-trained on Kinetics-400 as the backbone for visual encoding, and use ResNet-34 for static feature extraction. The frame length per sample is set to 96, and we split each video into 9 snippets with 16 frames each. We follow previous works [15, 16] and set the number of semantic steps L' as 3. The weighting factors λ_l used in final score regression are set to $\{3, 5, 2\}$.

Pose estimation is performed as a preprocessing step using YOLOv8n for bounding box detection and ViTPose++-huge for 2D keypoint extraction. These keypoints and their confidence scores are stored and used for downstream pose embedding modules.

We use the Adam optimizer with an initial learning rate of 1e-3 and no weight decay. The batch size is set to 4 for training and 2 for testing. We train the model for 100 epochs, with a learning rate decay factor of 0.1 applied after a 10-epoch warm-up. Batch normalization layers are frozen throughout training. Following [16], we use 75% of the dataset for training and 25% for testing. During inference, we employ a multi-exemplar voting strategy with 10 reference samples per query.

Evaluation Metrics.

We evaluate our method across three major components: action score prediction, temporal parsing, and spatial mask generation.

Action Quality Assessment. We use Spearman's rank correlation coefficient (ρ), ℓ_2 error (L2), and relative ℓ_2 distance (Rℓ_2) to measure the consistency and accuracy of predicted scores compared to ground-truth annotations.

Temporal Action Parsing. To assess the alignment of predicted sub-action segments with ground truth, we report average temporal Intersection over Union (tIoU) at thresholds 0.5 and 0.75, following [15].

Spatial Action Parsing. We evaluate predicted action masks using standard pixel-wise metrics, including Intersection-over-Union (IoU), F1 score, F2 score, accuracy, and recall. These metrics quantify both region-level and boundary-level segmentation quality.

Baselines.

We compare our proposed DiveDoc system with two representative baselines from prior work and several variants of our own embedding modules to evaluate the effectiveness of pose-guided scoring.

FineParser [16] is a state-of-the-art method for diving action quality assessment. It serves as our primary baseline, using spatial-temporal parsing and contrastive regression without explicit pose information.

Naive Embedding denotes our simplest variant, where pose features are averaged across frames and used as a global auxiliary embedding for scoring.

Weighted Embedding introduces keypoint-wise confidence weighting into the pose encoder, allowing the

model to downweight unreliable joints based on detection confidence.

Hierarchical Encoder is our full variant, which captures both joint-level and temporal structure via a hierarchical attention mechanism. This variant corresponds to the final implementation of DiveDoc used in our main results.

4.2. Overall Results

Pose Extraction Quality.

We first validate the effectiveness of our pose estimation pipeline, which serves as a foundation for downstream modules. As illustrated in Figure 2, the YOLOv8n detector reliably localizes the diver across frames, and the ViTPose++-huge model generates consistent keypoints with high precision. Even in challenging scenarios involving motion blur or body contortion, most keypoints remain correctly located, and uncertain detections are downweighted via confidence scores. This contributes to robust pose embedding and mitigates the risk of propagating noisy joint estimates.

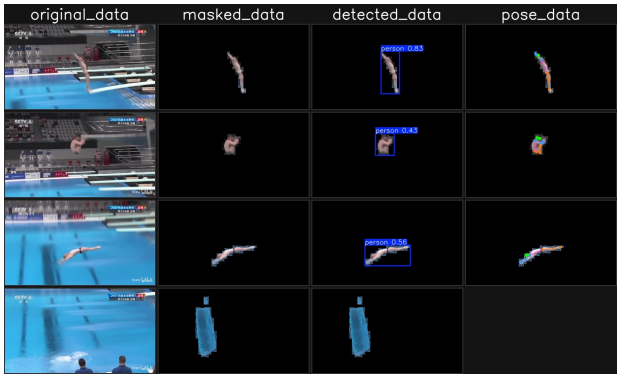

Figure 2. Visualization of the pose extraction pipeline. The fourth column shows ViTPose++-huge predictions, with only minor errors (e.g., foot misplacement in row 4 due to motion blur).

Spatial Mask Prediction.

To assess whether the introduction of pose features affects the spatial parsing module, we visualize predicted action masks generated by our framework. As shown in Figure 3, the system accurately identifies diver regions while suppressing background noise, demonstrating that our fusion strategy preserves spatial alignment quality. This confirms the compatibility of pose and appearance features in supporting interpretable region-level understanding.

Numerical Score Prediction.

As shown in Table 1, both the Weighted MLP and our final DiveDoc model outperform the FineParser (FP) baseline and the naive pose variant across most metrics.

The Weighted MLP achieves the best performance in correlation ($\rho = 0.8974$), regression error (L2 = 24.32), and relative ℓ_2 distance (RL2 = 0.529), indicating the effectiveness of incorporating pose confidence into the scoring mechanism. It also yields the highest F1 and F2 scores and the best recall, showing that reliable pose attention contributes to improved spatial focus and step sensitivity.

Meanwhile, DiveDoc achieves the highest tIoU@0.5 and tIoU@0.75, confirming that the addition of hierarchical temporal reasoning improves temporal alignment of sub-actions. It also achieves the best IoU score, showing stronger mask consistency across frames.

Overall, both variants provide notable improvements over the FineParser baseline. Weighted MLP focuses on accurate per-modality regression, while DiveDoc prioritizes structured temporal understanding and interpretable spatial parsing, with consistent top-tier results across dimensions.

Diagnosis Results.

To demonstrate the interpretability and semantic diversity of our DiveDoc diagnostic module, we present a case study comparing three query dives against the top exemplar (score: 91.2). For each query, we compute modality-specific deltas across dynamics, statics, and pose, then generate natural language feedback based on the resulting DiveScoreCard.

Figure 4 shows the full diagnostic result for **Query 1**. The motion delta is substantially negative (-53.4), while static and pose deltas are mildly positive. This suggests that although the athlete's posture and visual appearance are relatively stable, the dynamic aspects—such as timing, rotation speed, or entry angle—are significantly weaker. The generated feedback advises the athlete to focus on improving motion fluency, especially during entry and aerial transitions.

To assess the semantic robustness of our LLM-based diagnosis, we further analyze two additional cases:

Query 2 (Pred: 81.5, GT: 86.4) yields a moderate dynamic delta (-29.8) and strong pose delta ($+0.55$). The LLM highlights body control and alignment as strengths, while suggesting improvements in synchronization during phase transitions. The feedback emphasizes refinement rather than overhaul, using different phrasing from Query 1.

Query 3 (Pred: 76.9, GT: 81.6) shows a large dynamic penalty (-43.7), with solid static and pose scores. The generated advice emphasizes rhythm and aerial execution quality, again with distinct language and focus compared to the previous two examples.

These cases illustrate not only the accuracy of our scoring model, but also the LLM's capacity to produce diverse, context-sensitive feedback grounded in structured diagnostic input.

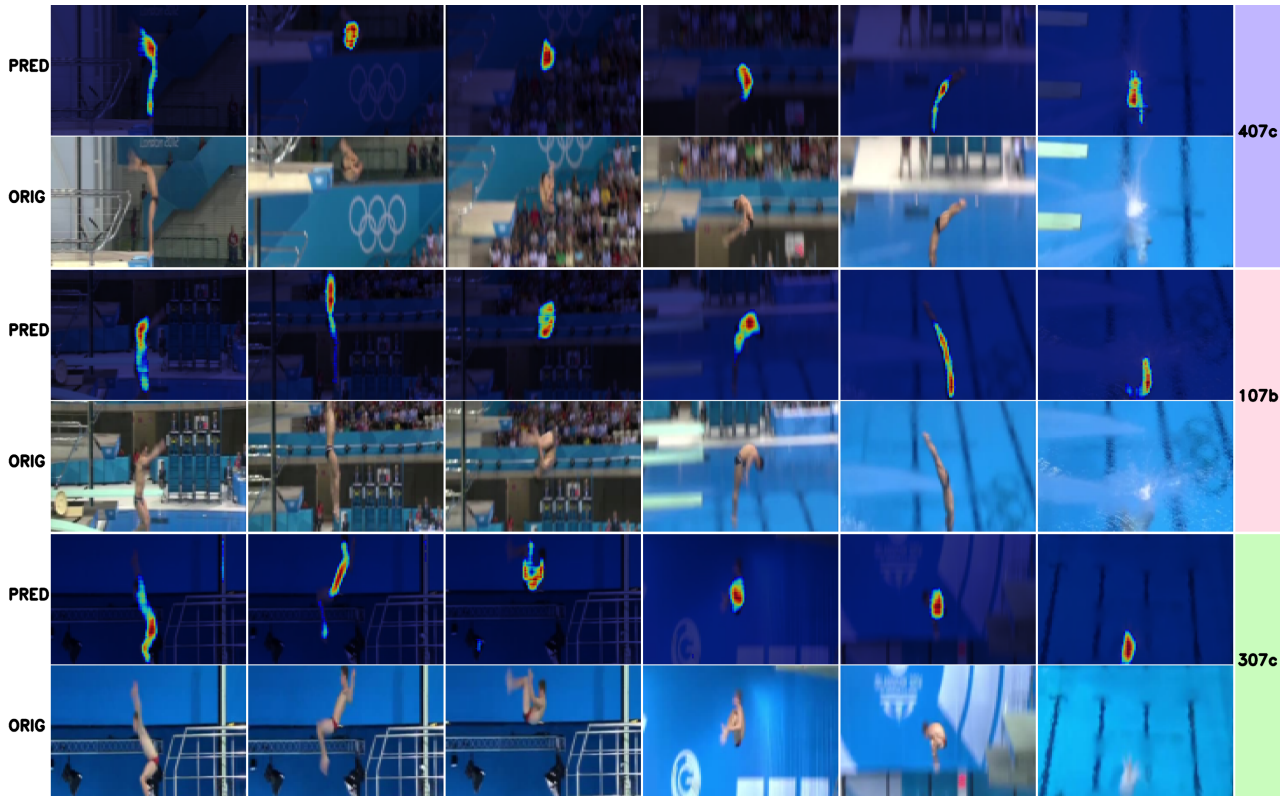

Figure 3. Visualization of predicted target action masks. Despite the integration of pose embeddings, our system maintains precise spatial focus, validating the robustness of the region parsing module.

Table 1. Comparison of DiveDoc and baselines. *RL2*, *IoU*, *F1*, *F2*, and *Recall* are scaled by (×100). Arrows indicate better direction.

Model	ρ (↑)	*L2* (↓)	*RL2* (↓,×100)	*tIoU@0.5* (↑)	*tIoU@0.75* (↑)	*IoU* (↑,×100)	*F1* (↑,×100)	*F2* (↑,×100)	*Acc* (↑)	*Recall* (↑,×100)
FP	0.8910	27.72	0.603	0.992	0.939	15.84	26.88	19.34	0.942	22.44
Naive	0.8918	25.90	0.563	0.996	0.945	14.97	24.28	17.24	0.958	20.80
Weighted	0.8974	24.32	0.529	0.996	0.942	15.87	27.21	19.66	0.959	23.63
DiveDoc	0.8926	27.88	0.607	0.997	0.955	15.94	25.41	18.23	0.959	21.58

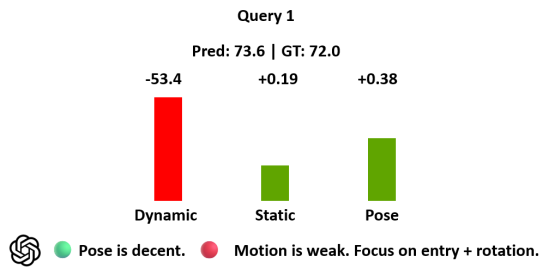

Figure 4. DiveDoc diagnostic visualization for Query 1. Modality-specific deltas relative to the exemplar reveal severe motion degradation and moderate pose/static quality. The corresponding feedback provides targeted, interpretable training suggestions.

5. Conclusion

We proposed **DiveDoc**, a pose-guided diagnostic framework that integrates fine-grained action parsing with language-based reasoning to deliver interpretable feedback on diving performance. Our experiments demonstrate that DiveDoc surpasses the FineParser framework in both score prediction and temporal segmentation, while producing context-aware natural language feedback. A key limitation is its reliance on highly annotated, domain-specific data, which may hinder generalization to other sports. One possible future direction is to more deeply integrate LLMs by designing the pipeline around their strengths and exploring fine-tuning strategies that incorporate expert feedback to generate more precise and domain-adapted advice.

References

[1] Qi An, Mengshi Qi, and Huadong Ma. Multi-stage contrastive regression for action quality assessment. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4110–4114. IEEE, 2024. 2

[2] Yang Bai, Desen Zhou, Songyang Zhang, Jian Wang, Errui Ding, Yu Guan, Yang Long, and Jingdong Wang. Action quality assessment with temporal parsing transformer. In *European conference on computer vision*, pages 422–438. Springer, 2022. 2

[3] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):172–186, 2019. 2

[4] Glenn Jocher, Ayush Chaurasia, and Jirka Qiu. Ultralytics yolov8. https://github.com/ultralytics/ultralytics, 2023. Accessed: 2025-05-22. 6

[5] Qi Li, Tzu-Chen Chiu, Hsiang-Wei Huang, Min-Te Sun, and Wei-Shinn Ku. Videobadminton: a video dataset for badminton action recognition. In *2024 IEEE International Conference on Big Data (BigData)*, pages 1387–1392. IEEE, 2024. 2

[6] Yixuan Li, Lei Chen, Runyu He, Zhenzhi Wang, Gangshan Wu, and Limin Wang. Multisports: A multi-person video dataset of spatio-temporally localized sports actions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13536–13545, 2021. 2

[7] Yulu Pan, Ce Zhang, and Gedas Bertasius. Basket: A large-scale video dataset for fine-grained skill estimation. *arXiv preprint arXiv:2503.20781*, 2025. 2

[8] Paritosh Parmar and Brendan Tran Morris. Learning to score olympic events. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 20–28, 2017. 1, 2

[9] Hamed Pirsiavash, Carl Vondrick, and Antonio Torralba. Assessing the quality of actions. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13*, pages 556–571. Springer, 2014. 1, 2

[10] Mengshi Qi, Hao Ye, Jiaxuan Peng, and Huadong Ma. Action quality assessment via hierarchical pose-guided multi-stage contrastive regression. *arXiv preprint arXiv:2501.03674*, 2025. 4

[11] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. Finegym: A hierarchical video dataset for fine-grained action understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2616–2625, 2020. 2

[12] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5693–5703, 2019. 2

[13] Yansong Tang, Zanlin Ni, Jiahuan Zhou, Danyang Zhang, Jiwen Lu, Ying Wu, and Jie Zhou. Uncertainty-aware score distribution learning for action quality assessment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9839–9848, 2020. 1, 2

[14] TensorFlow Team. Movenet: Ultra fast and accurate pose detection model. https://www.tensorflow.org/hub/tutorials/movenet, 2021. Accessed: 2025-05-22. 2

[15] Jinglin Xu, Yongming Rao, Xumin Yu, Guangyi Chen, Jie Zhou, and Jiwen Lu. Finediving: A fine-grained dataset for procedure-aware action quality assessment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2949–2958, 2022. 1, 2, 6

[16] Jinglin Xu, Sibo Yin, Guohao Zhao, Zishuo Wang, and Yuxin Peng. 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*, pages 14628–14637, 2024. 1, 2, 6

[17] Jinglin Xu, Guohao Zhao, Sibo Yin, Wenhao Zhou, and Yuxin Peng. Finesports: A multi-person hierarchical sports video dataset for fine-grained action understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21773–21782, 2024. 2

[18] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in neural information processing systems*, 35:38571–38584, 2022. 2

[19] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose++: Vision transformer for generic body pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(2):1212–1230, 2023. 2, 6

[20] Xumin Yu, Yongming Rao, Wenliang Zhao, Jiwen Lu, and Jie Zhou. Group-aware contrastive regression for action quality assessment. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7919–7928, 2021. 2