مروری جامع بر ارزیابی کیفیت فعالیت‌های انسان مبتنی بر ویدئو

مرجان مزروعی[۱]، دانشجوی دکتری، احسان فضل ارثی[۲*]، استادیار، عابدین واحدیان[۳]، دانشیار، احد هراتی[۴]، دانشیار

[۱] گروه مهندسی کامپیوتر – دانشکده مهندسی – دانشگاه فردوسی مشهد – مشهد – ایران – marjan.mazruei@mail.um.ac.ir

[۲] گروه مهندسی کامپیوتر – دانشکده مهندسی – دانشگاه فردوسی مشهد – مشهد – ایران – fazlersi@um.ac.ir

[۳] گروه مهندسی کامپیوتر – دانشکده مهندسی – دانشگاه فردوسی مشهد – مشهد – ایران – vahedian@um.ac.ir

[۴] گروه مهندسی کامپیوتر – دانشکده مهندسی – دانشگاه فردوسی مشهد – مشهد – ایران – a.harati@um.ac.ir

چکیده

ارزیابی کیفیت فعالیت[۱] (AQA)، حوزه‌ای مهم و در حال رشد در بینایی ماشین، بر توسعه روش‌های خودکار و عینی[۲] برای سنجش میزان درستی اجرای فعالیت‌ها و سطح مهارت در ویدئوها تمرکز دارد. کاربردهای متنوع آن در زمینه‌های ورزشی، مراقبت‌های پزشکی، تولیدات صنعتی و سایر سناریوهای نوظهور، توجه پژوهشگران را به خود جلب کرده است. با وجود پیشرفت‌های چشمگیر، نیاز به یک مرور جامع و نظام‌مند برای یکپارچه‌سازی دانش پراکنده و تعیین اولویت‌های پژوهشی آتی همچنان وجود دارد. در این مرور نظام‌مند، با بهره‌گیری از روش‌شناسی استاندارد کیچنهام[۳]، تعداد ۱۰۰ مقاله مرتبط برای تحلیل نهایی گزینش و بررسی شده است. حوزه AQA از پژوهش‌های پایه به سمت رویکردهای ریزدانه[۴]، چندوجهی[۵]، تعمیم‌پذیر و چند وظیفه‌ای تکامل یافته است. علاوه بر این، روندهای نوین پژوهشی مانند یادگیری مداوم، یادگیری خودنظارتی و سامانه‌های هوش مصنوعی قابل توضیح[۶]، به ویژه رویکردهای عصبی-نمادین، نقشی محوری در ارائه بازخورد شفاف و کاربردی ایفا می‌کنند. این مرور نگاهی جامع به جنبه‌های گوناگون این حوزه شامل بررسی سیستماتیک روش‌ها، مجموعه‌داده‌های معیار، معیارهای ارزیابی عملکرد، چالش‌های موجود و مسیرهای تحقیقاتی آینده ارائه می‌دهد. هدف اصلی آن فراهم‌سازی مرجعی ارزشمند برای پژوهشگران تازه‌وارد و متخصصان باتجربه جهت تسهیل مطالعات بعدی و هدایت پیشرفت‌های آتی است.

واژه‌های کلیدی: ارزیابی کیفیت فعالیت، ارزیابی مهارت، درک ویدئو، بینایی ماشین، یادگیری عمیق، بازنمایی ویژگی.

* نویسنده مسئول، **fazlersi@um.ac.ir**

[۱] Action Quality Assessment (AQA)

[۲] objectively

[۳] Kitchenham

[۴] fine-grained

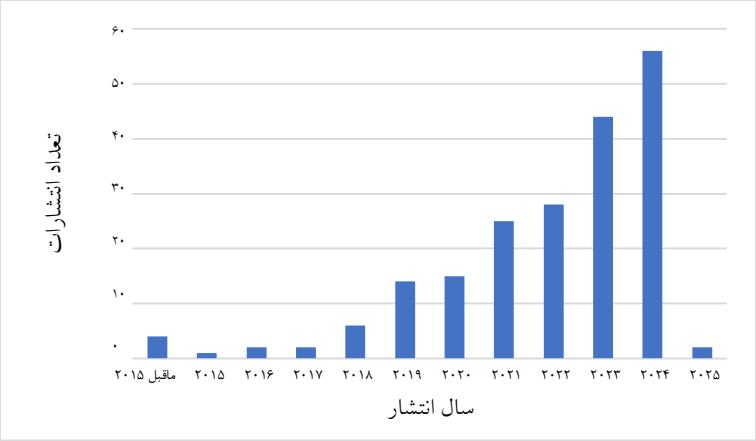
[۵] multimodal

[۶] Explainable AI (XAI)

۱. مقدمه

امروزه تکنیک‌های یادگیری عمیق و شبکه‌های عصبی کانولوشنی در حوزه‌های مختلف پردازش تصویر، عملکرد بسیار موفقی نشان داده‌اند [۱] و [۲] و [۳]. درک ویدئو نیز از موضوعات بنیادین در حوزه بینایی ماشین به شمار می‌رود و در این میان، ارزیابی کیفیت فعالیت انسانی در ویدئو به‌عنوان یکی از شاخه‌های نوظهور، جایگاه ویژه‌ای در میان پژوهش‌های اخیر یافته است. طی دهه گذشته، این حوزه از یک زمینه تخصصی محدود به عرصه‌ای پویا و در حال گسترش تبدیل شده است. افزایش قابل توجه تعداد پژوهش‌ها و رویکردهای پیشنهادی نوین در سال‌های اخیر نشان‌دهنده رشد چشمگیر توجه جامعه علمی به این موضوع است (شکل (۱)). هدف اصلی در AQA، سنجش میزان مهارت، صحت اجرا و میزان انطباق فعالیت‌ها با معیارهای استاندارد است؛ امری که در نهایت به تولید بازخوردهای دقیق و کاربردی برای بهبود عملکرد انسان منجر می‌شود. سیستم‌های AQA در طیف گسترده‌ای از حوزه‌ها کاربرد دارند. یکی از شاخص‌ترین آن‌ها، **تحلیل عملکرد ورزشی** است. در این حوزه، AQA با تحلیل داده‌محور از اجرای ورزشکاران، امکان شناسایی خطاهای فنی، ارزیابی شدت آن‌ها و ارائه پیشنهادهای اصلاحی دقیق را فراهم می‌کند [۴] و [۵]. برای نمونه، ارزیابی اجرای شیرجه یا حرکات نمایشی اسکیت می‌تواند به‌جای قضاوت‌های ذهنی داوران، با استفاده از معیارهای عینی و خودکار انجام گیرد و سوگیری‌های انسانی را کاهش دهد. در حوزه **پزشکی و توان‌بخشی** نیز AQA نقش مهمی ایفا می‌کند. این سیستم‌ها ابزارهایی کارآمد برای پایش از راه دور بیماران، ارزیابی پیشرفت درمان و تحلیل میزان دقت در اجرای تمرین‌های فیزیوتراپی فراهم می‌سازند. افزون بر این، در آموزش جراحی، تحلیل خودکار حرکات و تعامل با ابزارهای پزشکی، معیاری عینی برای سنجش مهارت و بهبود فرآیند آموزش است. کاربردهای AQA به حوزه **صنعت و توسعه نیروی کار** نیز گسترش یافته است [۶]. در این زمینه، از آن برای ارزیابی مهارت‌های عملی کارگران، نظارت بر رعایت استانداردهای ایمنی در خطوط تولید و بهبود همکاری میان انسان و ربات استفاده می‌شود. پیشرفت‌های اخیر در یادگیری عمیق و بینایی ماشین، دقت ارزیابی و قابلیت انطباق این سیستم‌ها را به میزان چشمگیری افزایش داده است. فراتر از حوزه‌های شناخته‌شده‌ای مانند ورزش و پزشکی، AQA در سال‌های اخیر به حوزه‌های نوینی همچون **موسیقی** (مانند ارزیابی کیفیت نواختن پیانو) و **فعالیت‌های روزمره زندگی** نیز راه یافته است. این گسترش حوزه‌های کاربردی نشان‌دهنده پتانسیل بالای AQA در سنجش کیفیت عملکرد انسانی در محیط‌های متنوع است.

با توجه به رشد سریع و پراکندگی منابع در این زمینه، انجام یک مرور نظام‌مند برای بررسی جامع وضعیت پژوهش‌های AQA ضرورت دارد. در این مطالعه، به‌منظور تضمین دقت و قابلیت بازتولید نتایج، از چارچوب مرور نظام‌مند کیچنهام استفاده شده است. این روش‌شناسی که یکی از رویکردهای پذیرفته‌شده در علوم کامپیوتر محسوب می‌شود، شامل مراحل اصلی مانند تدوین پرسش‌های پژوهشی، انتخاب مقاله‌ها، استخراج داده‌ها و تحلیل نتایج است. هدف این مرور، ترسیم چشم‌اندازی جامع از وضعیت فعلی حوزه AQA، تحلیل مقایسه‌ای روش‌ها، شناسایی شکاف‌های تحقیقاتی و ارائه مسیرهای بالقوه برای پژوهش‌های آینده است. ساختار مقاله به‌گونه‌ای تنظیم شده است که نمایی جامع و منسجم از ارزیابی کیفیت فعالیت انسانی مبتنی بر بینایی ارائه دهد. در بخش دوم، مفاهیم و تعاریف بنیادی معرفی می‌شود. بخش سوم به بررسی روش‌ها و الگوریتم‌های AQA از رویکردهای کلاسیک تا مدل‌های یادگیری عمیق اختصاص دارد. در بخش چهارم، مجموعه‌داده‌ها و معیارهای ارزیابی متداول مرور می‌شوند. بخش پنجم به چالش‌ها و محدودیت‌های اصلی اختصاص یافته و بخش ششم مسیرهای آینده پژوهش را مطرح می‌کند. در پایان، بخش هفتم به جمع‌بندی یافته‌ها می‌پردازد.



شکل (۱): روند رشد پژوهش‌ها در حوزه AQA طی دهه اخیر، تعداد انتشارات مرتبط (مقاله‌های کنفرانسی، ژورنالی و پایان‌نامه‌ها) نشان داده شده است.

۲. مفاهیم اصلی ارزیابی کیفیت فعالیت انسان مبتنی بر ویدئو

ارزیابی کیفیت فعالیت به توانایی محاسباتی برای اندازه‌گیری کمی کیفیت فعالیت‌ها یا سطح مهارت‌های انسانی اشاره دارد. هدف اصلی این حوزه، جایگزینی ارزیابی‌های ذهنی و ناهماهنگ انسانی با سامانه‌هایی خودکار، عینی و قابل‌اعتماد است. این ارزیابی‌ها معمولاً بر پایه داده‌های چندوجهی از جمله ویدئو، داده‌های اسکلتی و حسگرهای حرکتی انجام می‌شود تا تصویری جامع از نحوه اجرای فعالیت فراهم شود. در معماری‌های مبتنی بر یادگیری عمیق، فرآیند AQA عموماً شامل سه گام اساسی است: **استخراج ویژگی** از داده‌های ورودی، **یادگیری بازنمایی جامع** از این ویژگی‌ها، و **پیش‌بینی کیفیت** از طریق نگاشت بازنمایی به یک امتیاز عددی. این فرآیند را می‌توان به صورت زیر فرموله کرد:

$$\hat{y} = h\left(g\left(f_1(\mathbf{X}^{(1)}), f_2(\mathbf{X}^{(2)}), \dots, f_M(\mathbf{X}^{(M)})\right)\right) \quad (۱)$$

در این رابطه، $\mathbf{X}$ مجموعه داده‌های ورودی چندوجهی، f_m تابع استخراج ویژگی برای هر نوع داده، g لایه ادغام ویژگی‌ها و h تابع پیش‌بینی امتیاز نهایی است. در این حوزه، شناسایی عناصر کلیدی هر فعالیت اهمیت ویژه‌ای دارد، زیرا هر عنصر سهم متفاوتی در کیفیت نهایی اجرا دارد. به عنوان نمونه، در رشته شیرجه عواملی مانند نوع چرخش، زاویه ورود به آب و میزان خمیدگی بدن و درنتیجه اندازه پاشش آب، تعیین‌کننده کیفیت هستند؛ در ژیمناستیک، وضعیت بدن در حین چرخش، فشردگی و کشیدگی پاها و پرهیز از خطاهایی مانند باز بودن پاها یا پیچ‌خوردگی مچ‌ها نقش اساسی دارد؛ و در ورزش‌های زمستانی نیز نحوه فرود یا کنترل تخته، معیارهای کلیدی ارزیابی محسوب می‌شوند. این نمونه‌ها نشان می‌دهند که ارزیابی دقیق نیازمند مدل‌هایی است که بتوانند ویژگی‌های مکانی، زمانی و تکنیکی فعالیت‌ها را به‌طور هم‌زمان تحلیل کرده، شدت خطا را تعیین و جریمه متناسب اعمال کنند.

با وجود شباهت‌های ظاهری میان AQA و بازشناسی فعالیت انسان[۱] (HAR)، این دو وظیفه از نظر هدف و ماهیت کاملاً متمایز هستند. درحالی‌که HAR صرفاً به شناسایی نوع فعالیت انجام‌شده (مانند دویدن یا پریدن) می‌پردازد، AQA بر ارزیابی دقیق کیفیت و ظرافت‌های اجرای یک فعالیت خاص متمرکز است. به عنوان مثال، HAR تنها فعالیت «شیرجه‌زدن» را تشخیص می‌دهد، اما AQA عواملی همچون وضعیت بدنی، زاویه ورود به آب و میزان پاشش را نیز تحلیل می‌کند. برخلاف HAR که بر تمایز میان دسته‌های مختلف فعالیت متمرکز

[۱] Human Action Recognition (HAR)

است، AQA باید طیفی پیوسته از عملکرد را درون یک دسته خاص درک و ارزیابی کند [۵]. مدل‌ها در وظایف HAR فقط باید تفاوت‌های خارجی[۱] بین دسته‌های فعالیت مختلف را به تصویر بکشند، اما مدل‌ها در وظایف AQA بر تمایز تفاوت‌های داخلی[۲] در یک فعالیت خاص تمرکز دارند و باید تفاوت‌های ظریف بین حرکات پیوسته را بدست آورند. علاوه بر این، مجموعه‌داده‌های AQA در مقایسه با مجموعه‌داده‌های HAR معمولاً کوچک‌تر هستند، که این محدودیت، عامل مهمی در کارآیی سیستم‌های AQA محسوب می‌شود. این تمایزها، AQA را به مسئله‌ای چالش‌برانگیزتر و ظریف‌تر در حوزه یادگیری عمیق تبدیل کرده است.

۳. روش‌ها و الگوریتم‌های ارزیابی کیفیت فعالیت

مدل‌های AQA عموماً بر پایه شبکه‌های عصبی کانولوشنی دوبعدی و سه‌بعدی، همراه با شبکه‌های حافظه بلندمدت کوتاه‌مدت[۳] (LSTM)، طراحی می‌شوند تا ویژگی‌های ویدئویی را استخراج، ترکیب و تحلیل کرده و در نهایت فرآیند ارزیابی را تکمیل کنند. همانطور که در شکل (۲) نشان داده شده است، چارچوب‌های مبتنی بر یادگیری عمیق در AQA معمولاً از دو مرحله‌ی اصلی تشکیل می‌شوند. در مرحله نخست، **ویژگی‌های معنادار و مرتبط** از توالی‌های ویدئویی استخراج می‌شود و در مرحله دوم، **ماژول ارزیابی عملکرد** با استفاده از این ویژگی‌ها، میزان کیفیت اجرای فعالیت را کمی‌سازی می‌کند. در این بخش، ابتدا خلاصه‌ای از روش‌های یادگیری بازنمایی ویژگی‌های فعالیت ارائه شده است. سپس به بررسی جامع روش‌های ارزیابی کیفیت فعالیت موجود پرداخته می‌شود.

شکل (۲): چارچوب مدل ارزیابی کیفیت فعالیت مبتنی بر یادگیری عمیق.

۱،۳. استخراج ویژگی‌ها در یادگیری بازنمایی ویدئو

دقت ارزیابی کیفیت فعالیت، به‌طور مستقیم به توانایی مدل در استخراج بازنمایی‌های دقیق و متمایز از ویدئو وابسته است. در روش‌های اولیه، سیستم‌های AQA بر پایه ویژگی‌های سنتی و طراحی‌شده‌ی دستی نظیر تبدیل فوریه گسسته[۴] (DFT)، تبدیل کسینوسی گسسته[۵] (DCT) و مسیرهای چگال[۶] عمل می‌کردند. این روش‌ها قادر به درک تفاوت‌های ظریف بین فعالیت‌های یک کلاس نبودند [۷]. در برخی پژوهش‌ها نیز تلاش شد با استخراج ویژگی‌های حالت[۷] از هر فریم، بازخوردی قابل‌تفسیر فراهم شود. برای نمونه، در مطالعه [۸]، جهت‌های بهینه حرکت اندام‌ها به‌منظور افزایش امتیاز کلی پیشنهاد شد. اما خطا در برآورد وضعیت اسکلتی به‌طور مستقیم بر دقت ارزیابی، اثر منفی داشت. همچنین این بازنمایی‌ها قادر به مدل‌سازی اشیای درگیر در فعالیت‌ها (مانند توپ ورزشی) نیستند.

---

[۱] external differences
[۲] internal differences
[۳] Long Short-Term Memory (LSTM)
[۴] Discrete Fourier Transformation (DFT)
[۵] Discrete Cosine Transformation (DCT)
[۶] dense trajectories
[۷] pose

پیشرفت شبکه‌های عصبی عمیق موجب تحولی بنیادین در یادگیری بازنمایی ویدئویی شد. در شبکه‌های عصبی کانولوشنی (CNN)، بازنمایی‌های سلسله‌مراتبی از داده‌ها به‌صورت خودکار از طریق یادگیری مبتنی بر انتشار پس‌رو[۱] استخراج می‌شود. موفقیت این شبکه‌ها در طبقه‌بندی تصاویر، مسیر استفاده از آن‌ها برای تحلیل ویدئو را هموار کرد. در این رویکرد، ویدئو به‌عنوان مجموعه‌ای از فریم‌های متوالی در نظر گرفته می‌شود؛ ویژگی‌های هر فریم جداگانه استخراج و سپس تجمیع می‌شوند. CNNها در استخراج ویژگی‌های مکانی از فریم‌های منفرد ویدئو عملکرد بالایی دارند، در حالی که شبکه‌های عصبی بازگشتی RNNها (مانند LSTMها یا GRUها) قادر به مدل‌سازی وابستگی‌های زمانی میان فریم‌ها هستند. ترکیب این دو ساختار امکان تحلیل هم‌زمان ویژگی‌های مکانی و زمانی را فراهم می‌کند که برای ارزیابی دقیق فعالیت‌های انسانی حیاتی است. با وجود این، شبکه‌های کانولوشنی دوبعدی از لحاظ محاسباتی سبک‌ترند اما قادر به مدل‌سازی روابط زمانی طولانی‌مدت نیستند. در نتیجه، ساختار زمانی ویدئو نادیده گرفته می‌شود؛ برای مثال، مدل ممکن است قادر به تشخیص تفاوت میان باز و بسته شدن درب نباشد، یا در موقعیتی مشابه، نتواند تمایز میان برخاستن از صندلی و نشستن بر آن را تشخیص دهد. به‌منظور رفع این کاستی، بسیاری از پژوهش‌ها از ساختارهای دو جریانی (RGB و جریان نوری[۲]) یا CNNهای سه‌بعدی استفاده کرده‌اند. سیمونیان و زیسرمن [۹] شبکه‌ای دوجریانی معرفی کردند که یکی به استخراج ویژگی‌های مکانی و دیگری به استخراج اطلاعات حرکتی اختصاص داشت؛ هرچند محاسبه جریان نوری در این روش مستلزم پردازش پیش‌محاسبه‌ای است. شبکه‌های کانولوشنی سه‌بعدی با افزودن بعد زمانی به کانولوشن‌های دوبعدی، توانایی مدل‌سازی پویایی‌های ویدئو را به‌طور مستقیم فراهم کردند. تران و همکاران [۱۰] شبکه‌ای موسوم به C3D را برای یادگیری ویژگی‌های مکانی–زمانی در داده‌های ویدئویی بزرگ‌مقیاس پیشنهاد کرده و نشان دادند که این مدل‌ها می‌توانند ظاهر و حرکت را به‌طور هم‌زمان توصیف کنند. به دنبال آن، کریرا و زیسرمن [۱۱] مدل Inflated 3D ConvNet (I3D) را ارائه کردند که از داده‌های RGB و جریان نوری به‌طور هم‌زمان استفاده کرده و نتایج دو جریان را ادغام می‌کند. در ادامه، کیو و همکارانش [۱۲] شبکه‌ای شبه‌سه‌بعدی مبتنی بر ResNet معرفی کردند که در آن کانولوشن‌های سه‌بعدی به دو کانولوشن دوبعدی تفکیک می‌شود. این تفکیک ضمن کاهش تعداد پارامترها، کارایی محاسباتی را افزایش می‌دهد. در سال‌های اخیر، معماری‌های مبتنی بر ترنسفورمر به دلیل توانایی در مدل‌سازی وابستگی‌های بلندمدت زمانی–مکانی، به‌طور فزاینده‌ای در AQA به کار گرفته شده‌اند.

۲،۳. روش‌های AQA

پس از استخراج ویژگی، مرحله بعدی شامل پردازش آن‌ها توسط یک ماژول ارزیابی عملکرد است. روش‌های موجود برای ارزیابی کیفیت فعالیت عمدتاً بر اساس فرمول‌بندی مسئله به سه دسته اصلی، روش‌های مبتنی بر رگرسیون[۳]، مبتنی بر مقایسه زوجی[۴] و مبتنی بر درجه‌بندی[۵] تقسیم می‌شوند. این طبقه‌بندی، چارچوبی کلی برای درک رویکردهای مختلف موجود در AQA فراهم می‌سازد و به تحلیل نقاط قوت و ضعف هر گروه کمک می‌کند.

۱،۲،۳. امتیازدهی رگرسیون

---

[۱] Back-Propagation
[۲] optical flow
[۳] Regression-based
[۴] Pairwise-comparison based
[۵] Grading-based

امتیازدهی رگرسیونی یکی از متداول‌ترین رویکردها در AQA است که به‌ویژه در حوزه‌های ورزشی کاربرد گسترده‌ای دارد. در این رویکرد، هر اجرا از سوی داوران انسانی با یک امتیاز پیوسته ارزیابی می‌شود و هدف مدل، یادگیری تابعی است که بتواند این امتیاز عددی را از داده‌های ورودی پیش‌بینی کند. معمولاً برای این منظور از مدل‌های یادگیری ماشین مانند رگرسیون بردار پشتیبان[۱] (SVR) یا شبکه‌های کاملاً متصل (FCN) استفاده می‌شود تا رابطه‌ای میان ویژگی‌های استخراج‌شده از ویدئو و امتیاز واقعی برقرار گردد. در بسیاری از پژوهش‌های اولیه، بازنمایی اسکلت بدن انسان به‌عنوان ورودی اصلی مدل‌ها مورد استفاده قرار گرفته است. این داده‌ها از سامانه‌های ضبط حرکت یا الگوریتم‌های تخمین حالت بدن استخراج می‌شوند و شامل توالی‌هایی از مختصات دوبعدی یا سه‌بعدی مفاصل بدن هستند. چنین بازنمایی‌هایی نسبت به تغییرات ظاهری، شرایط نوری و شلوغی پس‌زمینه مقاوم بوده و امکان تمرکز بر ویژگی‌های سینماتیکی دقیق را فراهم می‌سازند. این ویژگی‌ها به‌ویژه در فعالیت‌هایی که دقت حرکات و هماهنگی بدن اهمیت بالایی دارد، نقش مؤثری در تحلیل کیفی ایفا می‌کنند. در یکی از نخستین پژوهش‌های این حوزه [۸]، مجموعه‌داده‌ای عمومی برای داوری المپیک معرفی شد. در این مطالعه، ابتدا حالت بدن ورزشکار در هر فریم تخمین زده شده و سپس ویژگی‌های حالت در سطح فریم با ویژگی‌های مبتنی بر گرادیان پیکسلی ترکیب گردید. پس از آن، داده‌ها با استفاده از DCT جهت کاهش ابعاد و حذف نویزهای فرکانس بالا پردازش شدند. در نهایت، مدل رگرسیون بردار پشتیبان خطی[۲] (L-SVR) برای پیش‌بینی امتیاز نهایی فعالیت به‌کار رفت. یک سال بعد، پژوهش دیگری نشان داد که استفاده از آنتروپی تقریبی[۳] در ویژگی‌های حالت می‌تواند اطلاعات حرکتی را بهتر از روش‌های DCT/DFT رمزگذاری کند [۱۳]. در ادامه، پاریسی و همکاران [۱۴] یک شبکه عصبی بازگشتی را پیشنهاد دادند که از ساختار خودسازماندهی در حال رشد برای یادگیری مؤثر توالی حرکت بدن استفاده می‌کند.

تخمین حالت بدن در محیط‌های واقعی ورزشی با چالش‌های متعددی همراه است. اتکای صرف بر داده‌های اسکلت، توانایی مدل را در درک عوامل غیرانسانی مؤثر در ارزیابی محدود می‌سازد. برای مثال، در رشته‌هایی مانند شیرجه، عناصری نظیر میزان پاشش آب در تعیین کیفیت اجرا تأثیرگذارند. پارمار و موریس [۵] برای اولین بار استفاده از روش‌های یادگیری عمیق مختلفی را برای امتیازدهی به رویدادهای المپیک بررسی کردند و مجموعه داده‌های UNLV-Dive و UNLV-Vault را بر اساس کار [۸] معرفی نمودند. آن‌ها سه چارچوب مبتنی بر شبکه کانولوشنی سه‌بعدی C3D [۱۰] پیشنهاد دادند که از ویژگی‌های مکانی-زمانی برای پیش‌بینی امتیاز استفاده می‌کنند. این چارچوب‌ها شامل رگرسیون با استفاده از SVR، LSTM، و ترکیبی از LSTM و SVR هستند. علاوه بر این، روشی کارآمد برای آموزش مدل‌ها در مجموعه‌داده‌های محدود ارائه شد که مبتنی بر قطعه‌بندی ویدئوها و تجمیع ویژگی‌ها در سطح کل ویدئو از طریق میانگین‌گیری و LSTM بود. در ادامه، لی و همکاران [۱۵] چارچوبی سرتاسری مبتنی بر C3D [۱۰] معرفی کردند که با استخراج ویژگی‌ها از قطعه‌های ویدئویی، از ترکیب تابع هزینه میانگین مربعات خطا و تابع هزینه رتبه‌بندی[۴] برای لحاظ‌کردن هم‌زمان ارزش امتیاز و ترتیب نسبی رتبه‌ها بهره می‌گیرد. همچنین در پژوهش [۱۶]، مدلی ارائه شد که با استخراج ویژگی‌های هر بخش، نمایش‌های مکانی-زمانی استخراج شده از شبکه کانولوشنی سه‌بعدی سنگین را به مدل‌های سبک‌تر دوبعدی منتقل می‌کند. برخلاف روش‌های مبتنی بر تقسیم ویدئو به کلیپ‌های هم‌طول و تجمیع ساده‌ی ویژگی‌ها [۵]، شیانگ و همکاران [۱۷] نخستین **رویکرد آگاه از بخش‌بندی معنایی زمانی** را ارائه کردند. این روش بر پایه‌ی چارچوب بخش‌بندی زمانی کاملاً نظارت شده ED-TCN[۵] [۱۸] و شبکه‌ی P3D [۱۲]

---

[۱] Support Vector Regression (SVR)

[۲] Linear Support Vector Regression (L-SVR)

[۳] approximate entropy

[۴] ranking loss

[۵] Encoder-Decoder TCN (ED-TCN)

عمل می‌کند. در آن، ویژگی‌های هر زیرفعالیت توسط P3D استخراج شده و سپس با مدل‌های رگرسیونی مانند SVR یا LR به امتیاز نهایی نگاشت می‌شوند. لی و همکاران [۱۹] نیز روشی برای ویدئوهای ورزشی طولانی‌مدت معرفی کردند که شامل دو مؤلفه بخش‌بندی قطعه‌های کلیدی[۱] (KFS) و پیش‌بینی امتیاز[۲] (SP) است. در KFS، شبکه کانولوشنی سه‌بعدی اطلاعات زمانی کوتاه‌مدت را پردازش کرده و شبکه Bi-LSTM اطلاعات زمانی بلندمدت را تحلیل می‌کند تا قطعه‌های کلیدی معنایی استخراج شده و قطعه‌های نامربوط فیلتر شوند. در SP، ویژگی‌های قطعه‌های کلیدی توسط شبکه کانولوشنی سه‌بعدی استخراج شده و با استفاده از تابع هزینه آنتروپی متقابل برای KFS و میانگین مربعات خطا برای SP، امتیاز نهایی پیش‌بینی می‌شود. در پژوهش [۲۰]، ویدئو با ED-TCN به پنج زیرفعالیت تقسیم و ویژگی‌های هر بخش توسط شبکه P3D [۱۲] استخراج شد. پنج شبکه کاملاً متصل به‌صورت مستقل امتیاز هر زیرفعالیت را پیش‌بینی کرده و مدل رگرسیون پنهان چندمرحله‌ای آن‌ها را به امتیاز کلی تبدیل می‌کند. برای بهبود دقت، شبکه رگرسیون تک‌مرحله‌ای ویژگی‌های سری‌شده زیرفعالیت‌ها را از طریق یک شبکه کاملاً متصل پردازش کرده و امتیاز نهایی را تولید می‌نماید. علاوه بر این، در [۲۱]، از برچسب امتیاز کلی برای تولید شبه‌زیرامتیازهای زیرفعالیت‌ها استفاده شد تا مدل AQA چند مرحله‌ای بتواند دقت پیش‌بینی امتیاز کلی را افزایش داده و بازخورد دقیق‌تری از عملکرد زیرفعالیت‌ها ارائه دهد.

شبکه‌های کانولوشنی و بازگشتی اگرچه در استخراج ویژگی‌های مکانی-زمانی کوتاه‌مدت مؤثرند، در مدل‌سازی وابستگی‌های بلندمدت محدودیت دارند. برای رفع این ضعف، **معماری‌های ترنسفورمر** به دلیل توانایی در مدل‌سازی وابستگی‌های بلندمدت مکانی-زمانی، کاربرد گسترده‌ای در AQA یافته‌اند. مکانیزم خودتوجهی ترنسفورمرها با تمرکز بر نواحی کلیدی ویدئو، دقت ارزیابی را بهبود می‌بخشد. در [۲۲]، شبکه I3D [۱۱] با رمزگشای ترانسفورمر[۳] ترکیب و از TimeSformer برای استخراج ویژگی و به دنبال آن MLP به عنوان جزء رمزگشا گسترش داده می‌شود. برای آموزش، تابع هزینه ترکیبی MSE-Spearman Correlation معرفی شده است. ژو و همکاران [۲۳] یک معماری عمیق شامل دو مؤلفه مکمل، یعنی شبکه LSTM خود توجه[۴] (S-LSTM) و LSTM کانولوشنی چند مقیاسی با پرش[۵] (M-LSTM) را معرفی کردند. S-LSTM از یک استراتژی ساده خودتوجهی برای انتخاب ویژگی‌های مهم کلیپ استفاده می‌کند که مستقیماً برای کارهای رگرسیون استفاده می‌شود. M-LSTM اطلاعات متوالی محلی و جهانی را در مقیاس چندگانه مدل می‌کند. در نهایت، رگرسیون خطی برای پیش‌بینی امتیازهای کیفیت فعالیت بر اساس ویژگی‌های ترکیب شده استفاده می‌شود. زنگ و همکاران [۲۴] با معرفی شبکه توجه آگاه از زمینه پویا-ایستا ترکیبی[۶] به ارزیابی فعالیت در ویدئوهای طولانی می‌پردازند. این رویکرد از یک معماری ترکیبی پویا-ایستا با یک جریان پویا برای استخراج اطلاعات حرکت و یک جریان ایستا برای کاوش حالت‌های ایستا ورزشکاران شناسایی‌شده در فریم‌های خاص استفاده می‌کند. ماژول توجه آگاه از زمینه، شامل شبکه کانولوشنی گراف زمانی (GCN) و واحد توجه، روابط بین بخش‌ها را مدل‌سازی کرده و ویژگی‌های غنی‌تری تولید می‌نماید. ویژگی‌های دو جریان ترکیب شده و از طریق ماژول رگرسیون، امتیاز نهایی ویدئو پیش‌بینی می‌شود. در [۲۵]، مکانیزم توجه زمانی برای وزن‌دهی پویا به کلیپ‌ها پیشنهاد شد تا قطعات مهم‌تر تأکید بیشتری یابند. وانگ و همکاران [۲۶] با بهره‌گیری از ماژول خودتوجهی لوله[۷] (TSA-Net) و ردیاب شی منفرد، اطلاعات مکانی-زمانی غنی مرتبط با حرکت ورزشکار را تولید می‌کنند. ویدئو به کلیپ‌هایی تقسیم شده، ویژگی‌های اولیه توسط I3D [۱۱]

---

[۱] Key Fragment Segmentation (KFS)

[۲] Score Prediction (SP)

[۳] Transformer decoder

[۴] Self-attentive LSTM (S-LSTM)

[۵] Multi-scale Convolutional Skip LSTM (M-LSTM)

[۶] hybrid dynAmic-static Context-aware attenTION NETwork (ACTION-NET)

[۷] Tube Self-Attention Network (TSA-Net)

استخراج می‌گردد و ماژول TSA ویژگی‌ها را تجمیع کرده و اطلاعات پس‌زمینه نویزی را حذف می‌نماید. در نهایت، شبکه MLP امتیازهای نهایی را پیش‌بینی می‌کند. ناگای و همکاران [۲۷] نیز روشی مبتنی بر یادگیری تقابلی[۱] برای ارزیابی کیفیت فعالیت پیشنهاد کردند که تأثیر زمینه صحنه را حذف می‌کند. این روش با دو تابع هزینه، شامل هزینه تقابلی[۲] صحنه برای نادیده گرفتن ویژگی‌های پس‌زمینه و هزینه رگرسیون ماسک‌شده انسانی برای تمرکز بر فعالیت‌های انسانی در صورت عدم مشاهده هدف، دقت ارزیابی را با تمرکز بر فعالیت هدف بهبود می‌بخشد. فارابی و همکاران [۲۸] تکنیک میانگین وزنی مبتنی بر یادگیری به نام Weight-Decider (WD) را برای تجمیع هوشمند ویژگی‌های استخراج‌شده توسط 3D ResNets و (2+1)D ResNets معرفی کردند. بطور مشابه، ژانگ و همکاران [۲۹] ویدئو را به کلیپ‌هایی تقسیم کرده و با مکانیزم توجه آگاه از زمان و تابع هزینه تقابلی، روابط و پویایی‌های زمانی بین کلیپ‌ها را مدل‌سازی می‌کنند. ویژگی‌های کلیپ‌ها با استفاده از چارچوب I3D [۱۱] استخراج شده، سپس با ضرایب وزنی محاسبه‌شده توسط ماژول توجه ترکیب می‌شوند تا بازنمایی نهایی ویدئو ایجاد گردد. این بازنمایی از طریق دو لایه کاملاً متصل به امتیاز نهایی نگاشت شده و با تابع هزینه ترکیبی فاصله L1 و L2 آموزش می‌بینند. در [۳۰] با الهام از مقیاس لیکرت[۳]، به‌جای رگرسیون مستقیم، درجات عملکرد مختلف در یک فعالیت به‌صورت صریح کمی‌سازی شدند و با استفاده از معماری ترنسفورمر رمزگشا و پرس‌وجوهای قابل یادگیری متنوع، ویژگی‌های خاص هر درجه را استخراج کرده و با ترکیب مقادیر کمی و پاسخ‌های تخمین‌زده‌شده از ویدئو، امتیاز نهایی محاسبه گردید. پژوهش [۳۱] نیز با ماژول توجه مکان‌یابی چندمقیاسی برای شناسایی ورزشکاران در فریم‌ها و LSTMها اطلاعات توالی‌های محلی و جهانی را ترکیب کرد. سون و همکاران [۳۲] چارچوبی نابینا مبتنی بر شبکه GRU چندسری[۴] و مکانیزم توجه پیشنهاد دادند. استفاده از (چند–سر) به این معنی است که مدل می‌تواند چندین دیدگاه مختلف از ویژگی‌های فعالیت را به طور همزمان پردازش و تجزیه و تحلیل کند که این امر غنای نمایش‌های یادگرفته شده را افزایش می‌دهد. این روش با استخراج ویژگی‌های مکانی–زمانی و پردازش موازی آن‌ها توسط سرهای GRU، وابستگی‌های زمانی را در سطوح مختلف مدل‌سازی می‌کند. مکانیزم توجه، تمرکز بر بخش‌های کلیدی حرکتی را تقویت کرده و تابع هزینه مبتنی بر آنتروپی اطلاعات، تطبیق مدل با ارزیابی‌های ذهنی را بهبود می‌بخشد.

توزیع نامتوازن امتیازهای کیفیت در مجموعه‌داده‌ها منجر به سوگیری مدل به سمت کلاس‌های پرتکرار و افت عملکرد در کلاس‌های کم‌نمونه می‌شود. برای رفع این مسئله، لیان و شائو [۳۳] با معرفی یک چارچوب استدلال زمانی بین‌مرحله‌ای و تنظیم وزن‌های مدل، توانایی تشخیص تفاوت‌های ظریف کیفیت را بهبود بخشیدند. آن‌ها با استفاده از برآورد چگالی هسته‌ای[۵] (KDE) و وزن‌دهی تابع هزینه بر اساس معکوس جذر چگالی برچسب‌ها، اثر سوگیری داده را کاهش دادند. همچنین، هوانگ و لی [۳۴] چارچوبی مبتنی بر رگرسیون عملکرد توالی معنایی همراه با وزن‌دهی متراکم نمونه‌ها را پیشنهاد کردند. این روش با استخراج ویژگی‌های ویدئویی توسط شبکه I3D [۱۱]، فعالیت را به بخش‌های نابرابر معنایی تقسیم کرده و با استفاده از کانولوشن‌های یک‌بعدی آبشاری، ویژگی‌ها را ادغام می‌کنند تا نمایش غنی‌تری ارائه دهند. استراتژی وزن‌دهی متراکم با تقویت نمونه‌های چالش‌برانگیز و اعمال تابع هزینه میانگین مربعات خطا با وزن‌دهی مثبت، ضمن مقابله با عدم‌تعادل برچسب‌ها، دقت رتبه‌بندی کیفیت را بهبود می‌دهد. در کنار این رویکردها، **یادگیری متضاد**[۶] از طریق طراحی توزیع‌های ویژگی حساس به امتیاز انجام می‌شود و به مدل امکان می‌دهد که ویژگی‌هایی مقاوم، دقیق و تعمیم‌پذیر را استخراج کند. لیو و همکاران [۳۵] یک چارچوب یادگیری متضاد هدایت‌شده با بازپخش ارائه کردند که با بهره‌گیری از بازپخش‌های

---

[۱] adversarial
[۲] adversarial loss
[۳] Likert scale
[۴] multi-headed GRU
[۵] Kernel Density Estimation (KDE)
[۶] contrastive

ویدئویی، اطلاعات زمانی و مکانی را برای بهبود دقت ارزیابی استخراج می‌کند. این روش از سه مسیر (ویدئوی ورودی، نمونه نمایه و بازپخش)، تفاوت‌های ظریف بین اجراها را مدل‌سازی کرده و با مقایسه ویدئوهای زنده با نمونه و بازپخش، ویژگی‌های مکانی-زمانی مرتبط را تحت شرایط متنوع استخراج می‌نماید. آن‌ها در پژوهشی دیگر [۳۶]، ترنسفورمر متضاد هدایت‌شده با بازپخش مبتنی بر آموزش مشترک سلسله‌مراتبی را پیشنهاد کردند. این مدل با ماژول تمرکز زمانی، بخش‌های کلیدی مانند خطاها یا لحظات برجسته را شناسایی کرده و از پردازش ویژگی‌های زائد جلوگیری می‌کند. همچنین، سه جریان موازی یادگیری متضاد ویژگی‌های مشترک را از دیدگاه‌های مختلف استخراج کرده و آموزش مشترک سلسله‌مراتبی با نظارت هم‌زمان بر لایه‌های سطحی و عمیق، پایداری و دقت بهینه‌سازی را افزایش می‌دهد. همچنین، که و همکاران [۳۷] روشی دومسیره با ترکیب رگرسیون مستقیم و یادگیری متضاد ارائه کردند که مسیر اول به پیش‌بینی امتیاز کلی می‌پردازد و مسیر دوم تفاوت‌های محلی زیرفعالیت‌ها را مدل‌سازی می‌کند.

شبکه‌های گرافی کانولوشنی (GCN) به‌طور فزاینده‌ای در AQA مبتنی بر اسکلت به کار گرفته شده‌اند، زیرا قادرند تعاملات مکانی و زمانی میان مفاصل بدن را با درنظرگرفتن اسکلت به عنوان یک گراف مدل‌سازی کنند. پان و همکاران [۳۸] شبکه کانولوشنی گرافی مبتنی بر روابط مفصلی پیشنهاد کردند که با استفاده از گراف‌های مکانی و زمانی و ماژول‌های اشتراک و تفاوت مفصلی، ویژگی‌های محلی و کلی فعالیت‌ها را استخراج و آن‌ها را از طریق ماژول رگرسیون به امتیاز نهایی نگاشت می‌کند. لی و همکاران [۳۹] نیز با بهره‌گیری از داده‌های اسکلتی استخراج‌شده از ویدئوهای RGB، ویژگی‌های دینامیک حرکتی را با شبکه گرافی زمانی-مکانی تحلیل کرده و تفاوت‌های ظریف حرکات را مدل‌سازی می‌نماید. این ویژگی‌ها سپس به شبکه رگرسیون وارد شده تا امتیازهای کیفیت پیش‌بینی شوند. برای رسیدگی به مشکل سردرگمی درون کلیپ و عدم انسجام بین کلیپ، ژائو و همکاران [۷] یک شبکه کانولوشنی گرافی سلسله مراتبی را معرفی کردند که با پالایش اطلاعات معنایی کلیپ‌ها، آن‌ها را به واحدهای دقیق‌تر (شات) تبدیل و از تجمیع شات‌ها، صحنه‌های معنادار و در نهایت نمایش ویدئویی کامل را ایجاد می‌کند. امتیاز نهایی نیز با مدل‌سازی توزیع گاوسی امتیازهای داوران به دست می‌آید. در ارزیابی فعالیت‌های گروهی، ژانگ و همکاران [۴۰] با معرفی ماژول توجه گروهی (GOAT) از GCN برای مدل‌سازی روابط مکانی بین بازیکنان و ترکیب آن با ویژگی‌های زمانی در سطح کلیپ استفاده کردند. لی و همکاران [۴۱] با معرفی شبکه کانولوشنی گرافی چنداسکلتی، الگوهای حرکتی مفاصل و بخش‌های بدن را مدل‌سازی می‌کنند. در این روش، سه نوع گراف اسکلتی (اتصال خودی، اتصال درون‌قسمتی و اتصال بین‌قسمتی) برای استخراج ویژگی‌های حالت ساخته می‌شوند. ساختار فیزیکی-بیولوژیکی (اتصال خودی) که روابط فیزیکی بین مفاصل را نشان می‌دهد، ساختار همسایگی نزدیک (اتصال درون‌قسمتی) که روابط بین مفاصل نزدیک به هم را در هر لحظه زمانی نشان می‌دهد، ساختار همسایگی دور (اتصال بین‌قسمتی) که روابط بین مفاصل دورتر را ثبت می‌کند که ممکن است در فعالیت‌های پیچیده تعامل داشته باشند. برای ادغام مؤثر اطلاعات از این ساختارهای چنداسکلتی، یک مکانیزم توجه تطبیقی به کار گرفته شده است. ژنگ و همکاران [۴۲] سیستمی برای ارزیابی تمرینات توان‌بخشی مبتنی بر داده‌های اسکلتی با خاصیت عدم حساسیت به چرخش معرفی کردند. این روش با استفاده از ماتریس ضرب داخلی اسکلت، اطلاعات مربوط به جهت‌گیری اسکلت را حذف کرده، در حالی که سایر اطلاعات مفید را حفظ می‌نماید. شبکه‌های کانولوشنی گرافی مکانی-زمانی با لایه‌های کاهش‌یافته برای استخراج ویژگی‌ها به‌کار رفته و نقشه‌برداری فعال‌سازی کلاس وزن‌دار با گرادیان، مفاصل مرتبط با حرکات نادرست را برای بازخورد بصری شناسایی می‌نماید. EGCN++ [۴۳] نیز یک راهبرد جامع برای یادگیری گروهی در زمینه ارزیابی تمرینات توان‌بخشی مبتنی بر داده‌های اسکلتی ارائه می‌دهد. این رویکرد با بهره‌گیری از شبکه‌های گرافی بهبودیافته، ویژگی‌های مکانی-زمانی را از داده‌های اسکلتی استخراج می‌کند و سپس از یک سازوکار ادغام تطبیقی برای تلفیق پیش‌بینی‌های حاصل از مدل‌های گوناگون استفاده می‌نماید. در [۴۴]

با افزودن مفاصل مجازی به اسکلت، فضای نمایش غنی‌تر شده و یادگیری متضاد برای افزایش تمایزپذیری بین ویدئوهای با کیفیت متفاوت به کار گرفته شد.

چن و همکاران [۴۵] نخستین روش ثبت هم‌زمان حرکات سه‌بعدی انسان و درک ریزدانه فعالیت‌ها در ویدئوهای ورزشی تک‌دوربینی را پیشنهاد دادند. این روش از یک ماژول نهان‌سازی حرکتی برای استخراج ویژگی‌های ضمنی و صریح حرکت سه‌بعدی استفاده کرده و با بهره‌گیری از شبکه‌های کانولوشنی گرافی مکانی-زمانی چندجریانی، روابط بین مفاصل بدن و ویژگی‌های معنایی دقیق فعالیت را مدل‌سازی می‌کند. این روش همچنین از یک بلوک نگاشت ویژگی معنایی برای ترکیب ویژگی‌های مرتبط با برچسب‌های سطح بالا بهره می‌برد. بطور مشابه، هوانگ و همکاران [۴۶] از بازسازی پارامتری سه‌بعدی بدن برای ارزیابی کیفیت فعالیت استفاده کردند. در [۴۷] نیز شبکه‌های کانولوشنی سه‌بعدی با ماژول‌های ارزیابی مختصات مفاصل و دینامیک ظاهری، ویژگی‌های چندمقیاسی مکانی-زمانی را استخراج کرده و با مکانیزم توجه، تمرکز مدل بر نواحی کلیدی را تقویت می‌کنند. در ادامه، هوانگ و همکاران [۴۸] با معرفی شبکه دومرجعی کمکی، ویژگی‌های درشت‌دانه را به بازنمایی‌های ریزدانه و کیفیت‌محور پالایش کرده و با بهره‌گیری از مرجع‌های معنایی در سطح گروهی[۱] و مرجع‌های فردی[۲] همراه با ماژول توجه هدایت‌شده توسط امتیاز، دقت پیش‌بینی را بهبود بخشیدند.

از آنجا که داوران انسانی ممکن است برای یک فعالیت واحد، امتیازهای متفاوتی ارائه دهند، برخی پژوهش‌ها به **مدل‌سازی عدم قطعیت در برچسب‌گذاری امتیازها** پرداخته‌اند. تانگ و همکاران [۴۹] چارچوب یادگیری توزیع امتیاز آگاه از عدم قطعیت[۳] (USDL) را معرفی کردند که با استفاده از توزیع گاوسی و کمینه‌سازی فاصله Kullback-Leibler میان توزیع هدف و پیش‌بینی‌شده، توزیع امتیاز را مدل‌سازی می‌کند. نسخه چندمسیری آن (MUSDL) نیز با ترکیب امتیازهای داوران مختلف، پیش‌بینی نهایی دقیق‌تری ارائه می‌دهد. بطور مشابه، ژو و هوانگ [۵۰] مدلی مبتنی بر عدم قطعیت برای تولید پیش‌بینی‌های متعدد پیشنهاد می‌کنند که شامل دو شاخه قطعی و نهان است؛ در شاخه قطعی، ویدئوها به کلیپ‌های هم‌پوشان تقسیم و ویژگی‌ها با I3D [۱۱] استخراج می‌شوند، در حالی که شاخه نهان با رمزگذار خودکار متغیر شرطی، ابهام در ویدئو را مدل کرده و از طریق نمونه‌برداری مکرر، پیش‌بینی‌های متعدد تولید می‌کند. سپس، ماژول توجه وزن، ویژگی‌های کلیپ را تجمیع کرده و با وزن‌دهی پویا در تابع هزینه رگرسیون، تأثیر نمونه‌های با اطمینان پایین را کاهش می‌دهد. جی و همکاران [۵۱] شبکه‌ای مبتنی بر تفکیک امتیاز عدم‌قطعیت با کمک محلی‌سازی طراحی کردند که شامل ماژول مکان‌یابی زیرفعالیت‌ها و ماژول تفکیک امتیاز با عدم‌قطعیت برای ادغام ویژگی‌های برش‌های مهم و ثبت روابط زمینه‌ای است. این مدل با پیش‌بینی جداگانه امتیاز کل عنصر[۴] (TES) و امتیاز کل اجزای برنامه[۵] (PCS) و بهره‌گیری از رگرسیون عدم‌قطعیت، قابلیت اطمینان امتیازها را افزایش می‌دهد. در [۵۲] نیز با استخراج ویژگی‌های بصری توسط ResNet [۱۲] و مدل‌سازی وابستگی‌های زمانی از طریق شبکه رمزگذار زمانی، تابع هزینه گاوسی برای یادگیری عدم قطعیت به کار رفته است. ژانگ و همکاران [۵۳] رگرسیون رمزگذار خودکار توزیع را معرفی کردند که با ترکیب مزایای رگرسیون و یادگیری توزیع برچسب، ویژگی‌های ویدئویی را به توزیع‌های امتیازی نگاشت می‌کند. این روش از پارامترسازی مجدد در رمزگذارهای خودکار متغیر برای نمونه‌برداری از امتیازها استفاده می‌کند. بنابراین نگاشت دقیق‌تری میان ویدئو و امتیاز برقرار می‌سازد. برای تسریع آموزش، یک تابع هزینه ترکیبی شامل هزینه بازسازی[۶] و پشتیبانی[۷] ساخته

---

[۱] semantic-level grade prototypes

[۲] individual-level reference

[۳] Uncertainty-aware Score Distribution Learning (USDL)

[۴] Total Element Score

[۵] Program Component Score

[۶] reconstruction loss

[۷] support loss

شده است. ماجیدی و همکاران [۵۴] روش ارزیابی کالیبره‌شده مبتنی بر روبریک را پیشنهاد کردند که مراحل فعالیت و معیارهای امتیازدهی را به‌صورت گراف غیرمدور جهت‌دار[۱] (DAG) مدل‌سازی می‌کند. در این روش، شبکه‌های گرافی عصبی (GNN) تعبیه‌های احتمالاتی مراحل را تولید و از طریق انتشار در گراف، امتیاز نهایی را با درنظرگرفتن عدم قطعیت محاسبه می‌کنند. از سوی دیگر، گائو و همکاران [۵۵] مدل AIM را به‌عنوان نخستین چارچوب ارزیابی کیفیت فعالیت مبتنی بر روابط نامتقارن ارائه کردند که تعاملات نامتقارن بین عامل‌ها را مدل‌سازی می‌کند. عامل‌های اولیه و ثانویه به‌صورت دستی تفکیک شده، تفاوت‌های آن‌ها محاسبه و با ویژگی‌های اولیه در شبکه LSTM تلفیق می‌گردد تا ویژگی‌های AIM تولید شود. سپس، ویژگی‌های کلی صحنه با I3D [۱۱] استخراج و از طریق مکانیزم توجه با AIM ادغام شده امتیازهای فعالیت پیش‌بینی می‌شوند. مدل AIL [۵۶] با گسترش AIM، دو ماژول کلیدی شامل ماژول تخصیص خودکار و ماژول جستجوی عملیات را معرفی کرده است. این ماژول‌ها قادرند عامل‌های اولیه و ثانویه را در فعالیت‌ها به‌طور خودکار طبقه‌بندی نموده و ساختار آن‌ها را متناسب با سناریوهای مختلف فعالیتی به‌صورت پویا تنظیم کنند.

پژوهش‌های اولیه در زمینه‌ی ارزیابی کیفیت فعالیت انسان با تکیه بر داده‌های بصری یا اسکلت، به نتایج قابل‌توجهی دست یافته‌اند؛ با این حال، وابستگی به داده‌های تک‌وجهی، به‌ویژه در محیط‌های پیچیده، مانع دستیابی به ارزیابی جامع شده است. ازاین‌رو، **رویکردهای چندوجهی** با هدف بهره‌گیری هم‌زمان از منابع متنوع داده‌ای نظیر ویدئوهای RGB، اسکلت، صدا، متن و سیگنال‌های حسگری توسعه یافته‌اند. این رویکردها با ترکیب اطلاعات مکمل از هر منبع، بازنمایی غنی‌تری از فعالیت انسانی فراهم آورده و موجب افزایش دقت و پایداری ارزیابی می‌شوند. برای نمونه، داده‌های اسکلت، حرکات و دینامیک بدنی را مدل‌سازی کرده، در حالی که داده‌های بصری و صوتی نشانه‌های ظریف و زمینه‌ای را تکمیل می‌نمایند و درنتیجه استحکام مدل را بهبود می‌بخشند. پارمار و همکاران [۵۷] یک روش چندوجهی برای ارزیابی مهارت‌های نوازندگی پیانو ارائه کردند که در آن، داده‌های ویدئویی توسط شبکه C3D [۱۰] و داده‌های صوتی توسط شبکه ResNet18-2D پردازش می‌شوند. ویژگی‌های استخراج‌شده از هر دو جریان ترکیب شده و امتیاز نهایی مهارت نوازنده پیش‌بینی می‌گردد. مدل پیشنهادی دو و همکاران [۵۸] از معماری معلم–شاگرد برای تلفیق اطلاعات بصری و معنایی بهره می‌برد. شبکه معلم دانش معنایی مرتبط با عناصر اسکیت (نظیر انواع پرش‌ها، چرخش‌ها، یا خطاهای خاص) را فراهم کرده و شبکه شاگرد با استفاده از پرس‌وجوهای اتمی قابل‌آموزش و مکانیزم توجه، این دانش را به فضای نمایش بصری تطبیق می‌دهد تا امتیازدهی دقیق‌تری حاصل شود. در [۵۹] چالش ارزیابی ویدئوهای طولانی اسکیت نمایشی بررسی شده است؛ جایی که هماهنگی حرکات فنی با موسیقی پس‌زمینه نقشی کلیدی ایفا می‌کند. در این روش، ویژگی‌های مکانی–زمانی از جریان ویدئویی و ویژگی‌های صوتی از جریان شنیداری استخراج و توسط رمزگذارها به نمایش‌های قابل پردازش تبدیل می‌شوند. سپس، مدل مبتنی بر MLP و واحدهای حافظه بازگشتی، با در نظر گرفتن همبستگی‌های درون‌وجهی و بین‌وجهی، بازنمایی جامعی را برای پیش‌بینی امتیاز نهایی ارائه می‌دهد. گدامو و همکاران [۶۰] شبکه تجزیه زمانی با هم‌ترازی بصری–معنایی را پیشنهاد کردند که با بهره‌گیری از ماژول تجزیه زمانی خودنظارتی، زیرفعالیت‌ها را استخراج کرده و معناشناسی سطح بالا و پویایی زمانی را به‌طور هم‌زمان مدل‌سازی می‌نماید. علاوه بر این، ماژول تعامل چندوجهی برای ثبت ارتباطات متقابل بین ویژگی‌های بصری و معنایی طراحی شده است تا ارزیابی دقیق‌تری از جزئیات فعالیت حاصل شود.

ژانگ و همکاران [۶۱] وظیفه جدید ارزیابی روایی فعالیت[۲] (NAE) را معرفی کردند که با استفاده از تعامل چندوجهی هدایت‌شده با پرامپت، رگرسیون امتیاز را به تطبیق ویدئو–متن تبدیل کرده و توضیحات حرفه‌ای و دقیق در قالب زبان طبیعی ارائه می‌دهد. در این

[۱] Directed Acyclic Graph (DAG)

[۲] Narrative Action Evaluation (NAE)

روش، ویژگی‌های ویدئویی توسط رمزگذار ویدئو استخراج و از طریق مکانیزم توجه متقاطع چندسری با پرامپت‌های قابل‌آموزش ترکیب می‌شوند. سپس، مولد چندوجهی متن با بهره‌گیری از ماسک توجه سه-توکنی[۱]، شرح‌های زبانی شامل جزئیات فعالیت، امتیازها و ارزیابی‌های کیفی تولید می‌کند. [۶۲] نیز شامل یک ماژول ارزیابی چندوجهی است که ویژگی‌های بصری و اسکلتی را از طریق یادگیری خودنظارتی با استفاده از نمایش‌های متضاد ادغام می‌کند. این ویژگی‌ها توسط یک ماژول تلفیق ترجیحی ترکیب شده و به یک شبکه رگرسیون وارد می‌شوند تا امتیازها پیش‌بینی گردد. [۶۳] با بهره‌گیری از دانش تخصصی درمانگران، چارچوبی مبتنی بر GCN پیشنهاد می‌دهد که قادر به استخراج ویژگی‌های مرتبط با بازتوانی از داده‌های اسکلتی است. این مدل از دو جریان موازی تشکیل شده است: جریان اول، ویژگی‌های عمومی مکانی-زمانی را با از شبکه‌های کانولوشنی گرافی چندمقیاسی و شبکه‌های کانولوشنی زمانی چندمقیاسی استخراج می‌نماید؛ و جریان دوم، با استفاده از GCN مبتنی بر دانش تخصصی، ویژگی‌های خاص بازتوانی را هدف قرار می‌دهد. در نهایت، ماژول تجمیع توجه دروازه‌ای[۲] ویژگی‌های دو جریان را ترکیب کرده و امتیاز نهایی تولید می‌شود. روش پیشنهادی در [۶۴] از ماژول توجه تفکیکی برای تبدیل فضای ویژگی‌های متراکم به نمایش‌های پراکنده استفاده می‌کند تا روابط پیچیده مکانی-زمانی مؤثرتر مدل‌سازی شوند. این ماژول با تحلیل همبستگی داده‌های بصری و اسکلتی، ویژگی‌های کلیدی را رمزگذاری کرده و ویژگی‌های کم‌اهمیت را حذف می‌کند. ویژگی‌های منتخب سپس به شبکه رگرسیون وارد شده تا امتیاز کیفیت فعالیت پیش‌بینی شود. زنگ و ژنگ [۶۵] با معرفی شبکه تلفیق چندوجهی تطبیقی پیش‌رونده، رویکردی نوین برای بهره‌گیری از داده‌های بصری (RGB و جریان نوری) و صوتی معرفی کرده‌اند. این چارچوب شامل سه شاخه مستقل برای استخراج ویژگی‌های خاص هر منبع داده و یک شاخه تلفیقی برای ادغام تطبیقی این ویژگی‌ها است. با استفاده از ماژول‌های رمزگشای ویژه‌وجه، تلفیقی تطبیقی، و رمزگشای ویژگی بین‌وجهی، اطلاعات به‌صورت سلسله‌مراتبی پردازش شده و سیاست‌های تلفیق برای بخش‌های مختلف فعالیت بهینه می‌گردد.

اغلب روش‌های AQA بر یک نوع فعالیت خاص متمرکزند که توسعه سیستم‌های یکپارچه و مقیاس‌پذیر برای **ارزیابی چندفعالیتی** را دشوار می‌کند. چالش اصلی در ایجاد مدلی واحد برای یادگیری تفاوت‌های ظریف معیارهای کیفیت در فعالیت‌های متنوع است. پارمار و موریس [۴] نخستین چارچوب AQA چندفعالیتی را معرفی کردند که با استفاده از ویژگی‌های زمانی-مکانی استخراج‌شده توسط C3D [۱۰] و مدل‌های SVR، LSTM و LSTM-SVR، امتیاز کیفیت فعالیت‌ها را پیش‌بینی می‌کند. در ادامه، [۶۶] نخستین شبکه تطبیقی را برای رفع ناسازگاری معماری‌های ثابت با انواع مختلف فعالیت‌ها پیشنهاد کرد. این چارچوب با بهره‌گیری از یادگیری تقویتی و شبکه کانولوشنی گرافی مبتنی بر مفاصل، روابط مفصلی خاص هر فعالیت را بر اساس ساختار اسکلت انسانی مدل‌سازی کرده و معماری ارزیابی را به‌صورت تطبیقی طراحی می‌کند. برای آموزش، از توابع هزینه میانگین مربعات خطا نرمال‌شده و پیرسون برای نرمال‌سازی خودکار امتیازها استفاده شده و معماری بهینه با مکانیزم جستجوی تمایزی و انتخاب‌گرهای عملیات تعیین می‌گردد. ژانگ و همکاران [۶۷] چارچوب انتقال مهارت ارزیابی آگاه از مرحله تطبیقی را معرفی کردند که با تقسیم فعالیت‌ها به مراحل مختلف و بهره‌گیری از طرح جستجوی ژنتیکی، فعالیت‌های مبدأ مرتبط را شناسایی کرده و دانش مهارتی آن‌ها را به فعالیت‌های هدف منتقل می‌نماید. این فرآیند با استفاده از یادگیری انتقالی، دقت ارزیابی در فعالیت‌های جدید را افزایش می‌دهد. ژو و همکاران [۶۸] با معرفی چارچوب هم‌راستاسازی دستورالعمل‌های درشت‌به‌ریز به مشکلات جابجایی دامنه و بیش‌برازش در شرایط داده محدود پرداختند. این روش با

[۱] tri-token attention mask

[۲] Gated Attention Pooling

تبدیل رگرسیون امتیاز به طبقه‌بندی دو مرحله‌ای، فرآیند داوری انسانی را شبیه‌سازی کرده و با هم‌راستاسازی ویژگی‌های مدل‌های پیش‌آموزش‌دیده طبقه‌بندی، تفاسیر قابل‌فهمی ارائه می‌دهد.

برخلاف مدل‌های سنتی که از توزیع‌های داده‌ای ایستا می‌آموزند، یادگیری مداوم شامل یادگیری از توزیع‌های داده‌ای پویا است. در حوزه **ارزیابی فعالیت مداوم**[۱]، که هدف آن ارزیابی کیفیت فعالیت در طول زمان و با مواجهه با داده‌های جدید است، چالش اصلی پدیده فراموشی فاجعه‌بار[۲] می‌باشد که به کاهش عملکرد مدل در وظایف قبلی هنگام یادگیری وظایف جدید منجر می‌شود. داداش‌زاده و همکاران [۶۹] چارچوب پیش‌آموزش مداوم با کارایی پارامتری بالا را پیشنهاد کردند. این روش با افزودن آداپتورهای سه‌بعدی به مدل پیش‌آموزش‌دیده (مانند I3D [۱۲] آموزش‌دیده بر Kinetics-400)، ویژگی‌های مکانی-زمانی خاص دامنه را از طریق یادگیری خودنظارتی استخراج می‌کند، در حالی که تنها پارامترهای آداپتورها به‌روزرسانی شده و وزن‌های اصلی مدل ثابت می‌مانند. این امر نیاز به تنظیم کامل مدل را حذف و شکاف دامنه را کاهش می‌دهد. لی و همکاران [۷۰] رویکردی برای ارزیابی مداوم کیفیت فعالیت معرفی کردند که بر یادگیری ویژگی‌های متمایز امتیاز با حفظ سازگاری وظیفه تمرکز دارد. برای مقابله با فراموشی، ابتدا ماژول بازپخش آگاه از همبستگی ویژگی-امتیاز با نمونه‌برداری گروهی در بازه‌های امتیاز و افزایش داده‌ها، پیوستگی توزیع ویژگی‌ها را حفظ می‌کند. سپس، ماژول گراف عمومی-اختصاصی فعالیت ویژگی‌های وظیفه جدید را با دانش وظایف پیشین تلفیق کرده و ویژگی‌های امتیازی سازگار با وظیفه استخراج می‌نماید. در [۷۱] نیز با منظم‌سازی گرافی و تابع هزینه مبتنی بر منیفلد، هم‌راستاسازی ویژگی‌های قدیمی و جدید و بازپخش آن‌ها، یادگیری مداوم ممکن شده است. پارمار و موریس [۷۲] **رویکردی چند وظیفه‌ای**[۳] طراحی کردند، که با انجام سه وظیفه موازی شامل بازشناسی فعالیت، تولید نظر و تفسیر عملکرد، و تخمین امتیاز کیفیت فعالیت، عملکرد را بهبود می‌بخشد. ویژگی‌های استخراج‌شده از C3D [۱۰] به شاخه‌های خاص‌وظیفه منتقل و بهینه‌سازی با توابع هزینه متناسب با هر وظیفه انجام می‌شود.

ماهیت «جعبه سیاه» مدل‌های یادگیری عمیق و عدم شفافیت در تصمیم‌گیری آن‌ها، چالش کلیدی بسیاری از مدل‌های شبکه عصبی سرتاسری است. در مقابل، یک الگوی نوظهور و بسیار امیدبخش در حوزه AQA، **رویکرد عصبی-نمادین** است. این رویکرد با هدف بهره‌گیری هم‌زمان از توان بالای شبکه‌های عصبی در شناسایی الگوهای پیچیده و قابلیت استدلال منطقی و تفسیرپذیری سامانه‌های نمادین طراحی شده است. این سیستم‌ها علاوه بر امتیازدهی دقیق، بازخوردهای قابل‌فهم و مبتنی بر شواهد بصری ارائه می‌کنند. اوکاموتو و پارمار [۷۳] رویکردی عصبی-نمادین برای AQA پیشنهاد کردند که با ترکیب قدرت شبکه‌های عصبی عمیق و استدلال مبتنی بر قواعد هوش مصنوعی، ارزیابی جامعی را با قابلیت تفسیرپذیری بالا ارائه می‌دهد. این رویکرد، ابتدا کل فعالیت (مانند شیرجه) و محیط آن را با مدل‌های عصبی عمیق به مؤلفه‌های معنایی قابل فهم (مانند سکو، حالت‌های بدن ورزشکار، موقعیت اعضای بدن، میزان پاشش آب و غیره) تجزیه می‌کند. سپس، سیستم قواعد سلسله‌مراتبی، مراحل اجرا، فازهای زمانی مختلف، نوع فعالیت و خطاهای جزئی را تحلیل کرده و با بهره‌گیری از میکروبرنامه‌ها، کیفیت هر مرحله را با امتیازدهی درصدی ارزیابی می‌نماید. این سیستم گزارش‌های تفصیلی و شواهد بصری تولید کرده و امکان شخصی‌سازی لحظه‌ای را برای تطبیق با نیازهای کاربران فراهم می‌آورد. ماتسویاما و لیم [۷۴] نیز رویکرد بخش‌بندی مبتنی بر روبریک قابل تفسیر[۴] را پیشنهاد دادند. روبریک‌ها؛ مجموعه‌ای از معیارهای ثابتی که توسط داوران ورزشی برای امتیاز دادن به عملکردها استفاده می‌شوند؛ در مدل هوش مصنوعی است. با استفاده از این روبریک‌ها، فرآیند تصمیم‌گیری قضات

[۱] Continual Action Assessment
[۲] Catastrophic Forgetting
[۳] Multi-Task Learning (MTL)
[۴] Interpretable Rubric-Informed Segmentation (IRIS)

انسانی و ارائه توضیحاتی برای قضاوت‌های خود را شبیه‌سازی می‌کنند. این روش با پیش‌بینی بخش‌های فعالیت، تفاوت امتیاز عناصر فنی هر بخش نسبت به امتیازهای پایه، امتیازهای مؤلفه‌های چندگانه برنامه و امتیاز نهایی، توضیحات شفافی ارائه می‌دهد. [۷۵] از یک شبکه رمزگشای ترنسفورمر مبتنی بر پرس‌وجو شامل استخراج‌کننده ویژگی، رمزگشا زمانی و سر رگرسیون وزن–امتیاز استفاده می‌کند. برای افزایش دقت و تفسیرپذیری، تابع هزینه توجه و روش مقداردهی اولیه پرس‌وجو جدیدی پیشنهاد گردید. تابع هزینه توجه با کاهش شباهت میان نقشه‌های خودتوجهی و توجه متقابل، مشکل پرش زمانی را در ترنسفورمرها رفع کرده و مقداردهی اولیه پرس‌وجوها با تنظیم واریانس توزیع گاوسی، همبستگی نقشه‌های خودتوجهی را تقویت می‌کند. ماژول رگرسیون وزن–امتیاز، با پیش‌بینی جداگانه وزن و امتیاز هر کلیپ و تجمیع آن‌ها، امتیاز نهایی را محاسبه می‌نماید. ونگ و همکاران [۷۶] رویکردی برای ارزیابی دقیق و قابل تفسیر مهارت جراحی ارائه می‌دهند. این روش متشکل از سه چارچوب همبسته است. چارچوب اول، MTL-VF، از یادگیری چند وظیفه‌ای با ویژگی‌های بصری معنایی برای پیش‌بینی امتیازهای مهارت نهایی کارآزمایی‌های جراحی استفاده می‌کند. چارچوب دوم، IMTL-AGF، با هدف ارائه بازخورد برای بهبود مهارت با شناسایی حرکات مشکل‌ساز است. همچنین سطوح مهارت حرکات فردی را به عنوان یک کار کمکی ارزیابی می‌کند. در حالی که چارچوب سوم، PG-GS، به بررسی توالی بهینه حرکات اشاره می‌کند.

در بسیاری از مطالعات، مدل‌ها به دلیل وابستگی به امتیازهای مغرضانه داوران انسانی، تمام عوامل را هنگام کمی‌سازی کیفیت یک فعالیت در نظر نمی‌گیرند. داوران انسانی معمولاً باید در اکثر موارد و حوزه‌ها، بدون امکان بازبینی ویدئو، به‌سرعت امتیازهایی ارائه دهند. در نتیجه، بخش‌های پایانی اجرا را پررنگ‌تر ارزیابی می‌کنند. به‌منظور کاهش این سوگیری، پژوهش‌هایی با **در نظر گرفتن عوامل چندگانه**، دقت و جامعیت ارزیابی را در سناریوهای پیچیده بهبود بخشیده‌اند. لیو و همکاران [۷۷] چارچوبی چندمسیره برای ارزیابی مهارت‌های جراحی معرفی کردند که جنبه‌هایی چون استفاده از ابزار جراحی، وضوح میدان جراحی و الگوی رویدادهای حین جراحی را در نظر می‌گیرد. هر مسیر ویژگی‌های مرتبط را استخراج کرده و به توالی‌های امتیاز مهارت تبدیل می‌کند. ماژول وابستگی مسیر روابط و اهمیت زمانی میان آن‌ها را مدل‌سازی می‌کند. در نهایت، این توالی‌های وزن‌دار ترکیب شده و امتیاز نهایی کیفیت و نیز لحظات خطا یا فازهای جراحی به‌صورت خودکار شناسایی می‌شوند. وانگ و همکاران [۷۸] مدلی برای پیش‌بینی وایرال‌شدن کلیپ‌های رقص ارائه کردند که از ترکیب ویژگی‌های اسکلتی، ظاهری، چهره، و عوامل محیطی استفاده می‌کند تا کیفیت و محبوبیت محتوا را پیش‌بینی کند.

با توجه به هزینه‌ی بالای برچسب‌گذاری داده‌ها، توجه پژوهش‌ها به **یادگیری بازنمایی خودنظارتی و نیمه‌نظارتی** افزایش یافته است. رودیتاکیس و همکاران [۷۹] روشی خودنظارتی ارائه دادند که بر هم‌ترازی خودنظارتی بین توالی‌های ویدئویی متکی است. این روش با استفاده از کدگذاری‌های سازگاری چرخه‌ای زمانی[۱] (TCC) به‌عنوان ویژگی‌های مکمل، ویژگی‌های ظاهری را با اطلاعات زمانی هم‌تراز می‌کند. پارمار و همکاران [۸۰] با استفاده از یادگیری متضاد حالت، حرکات نادرست را از حرکات ایده‌آل تفکیک کرده و دانش دامنه (مانند نظم تکرارها، دامنه حرکتی، یا فرم صحیح) را در مدل اعمال کردند. [۸۱] با هدف مدل‌سازی روابط میان فعالیت‌ها و اثرات آن‌ها در ویدئوهای پیچیده طراحی شده است. در این چارچوب، دو وظیفه‌ی اصلی شامل انتخاب فعالیت مؤثر و تعیین ارتباط فریم‌ها با فعالیت تعریف گردید. مدل پیشنهادی، با بهره‌گیری از شبکه عصبی عمیق و مکانیزم توجه، الگوهای بصری و زمانی مرتبط با کیفیت را شناسایی و تلفیق می‌کند. همچنین، یک تابع هزینه مبتنی بر رتبه‌بندی معرفی شده است که ویژگی‌های معنادار مکانی–زمانی مانند ردیابی اشیاء و بازنمایی حالت بدن را بدون نیاز به برچسب‌گذاری دستی یاد می‌گیرد. ژانگ و همکاران [۸۲] چارچوبی نیمه‌نظارتی با

[۱] Temporal Cycle Consistency (TCC)

سه ماژول بازیابی ویژگی بخش پوشانده‌شده برای یادگیری بازنمایی‌های ویدئوهای بدون برچسب، ارزیابی فعالیت برای یادگیری بازنمایی‌های ویدئوهای برچسب‌دار و هم‌ترازی توزیع نمایش برای تراز کردن توزیع ویژگی‌های داده‌های برچسب‌دار و بدون برچسب معرفی کردند که با یادگیری بازسازی خودنظارتی و مکانیزم آموزشی تقابلی با یک لایه معکوس گرادیان[۱] (GRL)، از داده‌های بدون برچسب برای بهبود تعمیم‌پذیری استفاده می‌کند. گدامو و همکاران [۸۳] شبکه‌ای نیمه‌نظارتی برای تجزیه زیرفعالیت‌ها را طراحی کردند که بر اساس ساختار شبکه معلم–دانش‌آموز، شبه‌برچسب‌ها و یادگیری متضاد گروهی، ویژگی‌های معنایی سازگار بین داده‌های برچسب‌دار و بدون برچسب ایجاد می‌کند. یون و همکاران [۸۴] نیز معماری نیمه‌نظارتی معلم–مرجع–دانش‌آموز را معرفی کردند. در این معماری، مدل معلم ویژگی‌های اولیه را از داده‌های بدون برچسب استخراج کرده و شبه‌برچسب‌هایی را برای آن‌ها تولید می‌کند. مدل مرجع با بهره‌گیری از اطلاعات جانبی مرتبط با فعالیت، برچسب‌های دقیق‌تری ایجاد کرده و اطلاعات زمینه‌ای باکیفیتی فراهم می‌آورد. مدل دانش‌آموز سپس با استفاده از این اطلاعات و داده‌های برچسب‌دار محدود، ارزیابی دقیقی از کیفیت فعالیت انجام می‌دهد. این روش از مکانیزم انتقال دانش برای بهره‌برداری از داده‌های بدون برچسب استفاده کرده و با طراحی یک حافظه اطمینان برای ذخیره دقیق‌ترین خروجی‌های شبکه معلم و مرجع، دقت شبه‌برچسب‌ها را بهبود می‌بخشد.

۲،۲،۳. مقایسه زوجی

در روش‌های مبتنی بر مقایسه زوجی، دو ویدئو از مجموعه‌داده به‌عنوان ورودی انتخاب می‌شوند تا کیفیت اجرای آن‌ها به‌صورت نسبی ارزیابی گردد. در این رویکرد، مدل به‌جای پیش‌بینی امتیاز مطلق، تفاوت یا فاصله کیفی میان دو نمونه را می‌آموزد تا رتبه‌بندی دقیق‌تری از کیفیت فعالیت‌ها ارائه دهد. این دسته از روش‌ها به‌ویژه زمانی مؤثرند که تعیین امتیاز عددی مطلق به دلیل ماهیت ذهنی یا پیچیدگی ارزیابی دشوار باشد. برتاسیوس و همکاران [۸۵] نخستین مدل یادگیری متضاد را برای ارزیابی مهارت در ویدئوهای دید اول‌شخص بسکتبال معرفی کردند. در این مدل، ویژگی‌های مکانی–زمانی غیرخطی با ترکیب شبکه‌های کانولوشنی، LSTM و مدل‌های مخلوط گاوسی استخراج شده و کیفیت فعالیت از طریق وزن‌دهی خطی به این ویژگی‌ها تخمین زده می‌شود. در ادامه، داوتی و همکاران [۸۶] مجموعه‌داده‌ای با نام EPIC-Skills و مدلی مبتنی بر رتبه‌بندی متضاد را معرفی کردند. این مدل با بهره‌گیری از شبکه‌های بخش‌بندی مکانی–زمانی و کانولوشنی، ویژگی‌های هر بخش را استخراج کرده و و از ترکیب تابع هزینه رتبه‌بندی و تابع هزینه شباهت برای یادگیری روابط بین جفت فعالیت‌ها استفاده می‌کند. سپس همان گروه در پژوهش‌های بعدی [۸۷] مدلی مبتنی بر توجه زمانی آگاه از رتبه‌بندی ارائه کردند که با استفاده از دو ماژول توجه مجزا، نواحی مفید و غیرمفید را شناسایی کرده و به کلیپ‌ها وزن‌های متفاوتی اختصاص می‌دهد تا تفاوت عملکردی میان جفت ویدئوها را برجسته سازد. آموزش مدل با ترکیبی از توابع هزینه‌ی رتبه‌بندی، اختلاف، آگاهی از رتبه و تنوع انجام می‌شود. بطور مشابه، لی و همکاران [۸۸] نیز یک مدل توجه مکانی مبتنی بر شبکه عصبی بازگشتی پیشنهاد کردند که در آن مکانیزم‌های ادغام توجه[۲] و تجمیع زمانی[۳] برای بهبود استخراج ویژگی‌ها به کار می‌روند.

سیستم پیشنهادی در [۸۹] شامل دو ماژول اصلی است. ابتدا، یک ماژول یادگیری عمیق متریک[۴] از یک شبکه سیامی برای یادگیری شباهت میان دو توالی فعالیت استفاده می‌کند. دوم، یک ماژول تخمین امتیاز از شبکه سیامی آموخته شده استفاده می‌کند و از رگرسیون برای تخمین امتیاز ویدئوها بر اساس شباهت آن با اجرای خبره همان فعالیت بهره می‌برد. یو و همکاران [۹۰] مسأله AQA را به صورت

[۱] Gradient Reverse Layer (GRL)
[۲] Attention pooling
[۳] Temporal aggregation
[۴] Deep Metric Learning (DML)

رگرسیون متضاد امتیازهای نسبی، با ارجاع به ویدئوی دیگری که ویژگی‌های مشابهی دارند، فرموله می‌کنند. در این روش، ویژگی‌های مکانی-زمانی با استفاده از I3D [۱۱] استخراج شده و با امتیاز مرجع ترکیب می‌شوند، سپس به یک درخت رگرسیون آگاه از گروه[۱] وارد می‌گردند. این درخت دامنه امتیازهای نسبی را به فواصل غیرهمپوشان تقسیم کرده و دسته‌بندی درشت به ریز[۲] انجام می‌دهد، سپس رگرسیون را در این فواصل زمانی کوچک انجام می‌دهد. بای و همکاران [۹۱] نیز روشی مبتنی بر رگرسیون متضاد پیشنهاد کردند که با بهره‌گیری از تنوع جزئی درون کلاسی، نمایش‌های سطح بخش با معانی معنایی قابل تفسیر استخراج می‌کند. این روش نمایش کلی ویدئو را به نمایش‌های زمانی جزئی در سطح بخش تجزیه کرده و از پرس‌وجوهای قابل یادگیری برای ثبت الگوهای زمانی اتمی مرتبط با دسته‌های خاص فعالیت استفاده می‌کند. این الگوها زیرفعالیت‌های کلیدی را در بر می‌گیرند. ترنسفورمر تجزیه زمانی برای رمزگشایی فریم‌های ویدئویی به مجموعه‌ای از نمایش‌های مرتب‌شده زمانی به کار گرفته شده و برای تخمین امتیاز کیفی فعالیت استفاده می‌شود. دو تابع هزینه رتبه‌بندی و پراکندگی[۳] برای آموزش مدل در پاسخ‌های توجه متقاطع پیشنهاد شده‌اند.

به منظور یادگیری تفاوت‌های بین ویدئوها برای کمک به کار امتیازدهی نهایی، لی و همکاران [۹۲] یک شبکه یادگیری متضاد زوجی[۴] پیشنهاد کردند که با شبکه رگرسیون پایه ترکیب شده و مدلی سرتاسری را تشکیل می‌دهد. در مرحله آموزش، جفت‌های ویدئویی به صورت تصادفی از مجموعه آموزشی انتخاب شده و ویژگی‌های آن‌ها با استفاده از شبکه ResNet استخراج می‌شود. این ویژگی‌ها به یک شبکه MLP وارد شده تا امتیاز کیفیت پیش‌بینی گردد. این مدل سه محدودیت بهینه‌سازی شامل خطای امتیاز مطلق، خطای امتیاز نسبی و سازگاری میان دو شبکه را هم‌زمان در نظر می‌گیرد. گدامو و همکاران [۹۳] نیز یک شبکه‌ی تجزیه‌کننده‌ی فعالیت مکانی-زمانی ریزدانه ارائه کردند که بر نواحی انسان‌محور تمرکز کرده و با استفاده از ترنسفورمر چندمقیاسی، وابستگی‌های بلندمدت میان زیرفعالیت‌ها را مدل‌سازی می‌کند. در حوزه‌های خاص، هیپینی و همکاران [۹۴] برای ارزیابی رقص‌های دونفره از تابع هزینه رتبه‌بندی جفتی استفاده کردند تا مقایسه‌ی مستقیم کیفیت اجراها را ممکن سازند. لیو و همکاران [۹۵] ویدئوی کامل فعالیت را به کلیپ‌های مجزا تفکیک کرده و با ترکیب امتیازهای هر بخش از طریق وزن‌دهی، امتیاز نهایی فعالیت را برآورد کردند. آن و همکاران [۹۶] نیز یک رویکرد رگرسیون متضاد چندمرحله‌ای را پیشنهاد دادند. این روش ویدئو را به مراحل پیوسته تفکیک می‌کند و سپس با بهره‌گیری از رگرسیون متضاد مرحله‌ای، تفاوت‌های بین ویدئوی هدف و ویدئوی نمونه با محتوای مشابه را محاسبه می‌کند تا امتیاز نسبی عملکرد استخراج شود.

ژو و همکاران [۹۷] با توسعه تجزیه‌کننده فعالیت مکانی-زمانی ریزدانه، با استفاده از ماسک‌های پیش‌زمینه انسانی، نواحی کلیدی را استخراج و ساختارهای زمانی زیرفعالیت‌ها را مدل کردند. در ادامه، رگرسیون متضاد ریزدانه برای ارزیابی تفاوت‌های کیفی و تولید امتیاز نهایی به‌کار رفت. لو و همکاران [۹۸] مدلی مبتنی بر ترنسفورمر آگاه از ریتم پیشنهاد کردند که با الهام از ریتم‌های موسیقی، الگوهای ریتمی مرتبط با مدت‌زمان وظایف را به‌صورت تطبیقی استخراج می‌کند. این مدل با تنظیم پویای تمرکز بر ویژگی‌های زمانی و توالی‌های ویدئویی، از یک ماژول هم‌توجهی برای برجسته‌سازی اطلاعات مرتبط با مدت‌زمان در مقایسه قطعه‌های ویدئویی با زمان‌های اجرای مشابه استفاده می‌کند. ترکیب اطلاعات مدت‌زمان با ویژگی‌های رویه‌ای، رتبه‌بندی کیفیت را در فعالیت‌هایی مانند مهارت‌های پزشکی و زندگی روزمره بهبود می‌بخشد. ژو و همکاران [۹۹] روشی با تمرکز بر رویدادهای شیرجه متنوع ارائه کردند که بر مجموعه داده‌ای با حاشیه‌نویسی‌های دقیق رویه‌های فعالیت استوار است. با توجه به جفت ویدئوها، ابتدا ویژگی‌های بصری مکانی-

---

[۱] Group-Aware Regression Tree (GART)

[۲] coarse-to-fine

[۳] sparsity loss

[۴] Pairwise Contrastive Learning Network (PCLN)

زمانی با استفاده از I3D [۱۱] استخراج می‌شود. سپس ماژول توجه بخش‌بندی زمانی[۱] برای یادگیری تعبیه‌های آگاه از رویه معرفی شده که با تجزیه جفت ویدئو ورودی و نمونه به مراحل متوالی با تناظرهای معنایی و زمانی، ارزیابی کیفیت را بهبود می‌بخشد. سپس توجه متقاطع آگاه از رویه[۲] برای یادگیری تعبیه‌های بین مراحل جفت ویدئو استفاده می‌شود و مطابقت‌های معنایی، مکانی و زمانی آن‌ها را به تصویر می‌کشد. این ویژگی‌ها برای رگرسیون متضاد ریز دانه[۳] استفاده می‌شوند تا مکانیزم امتیازدهی قابل اعتماد به دست آید. آن‌ها همچنین با توسعه مجموعه‌داده +FineDiving به تحلیل روابط معنایی و مکانی–زمانی بین مراحل مختلف یک فعالیت می‌پردازد [۱۰۰]. این روش، موسوم به توجه بخش‌بندی مکانی–زمانی، با بخش‌بندی رویه‌های حرکتی به مراحل متوالی، ویژگی‌های قدرتمندی را از طریق ماژول توجه حرکتی مکانی و توجه متقاطع آگاه از رویه استخراج می‌کند و به مدل امکان می‌دهد مناطق فعالیت پیش‌زمینه انسان‌محور را شناسایی کرده و تأثیر پس‌زمینه‌های نامربوط را کاهش دهد. در نهایت، با استفاده از رگرسیون متضاد ریزدانه، مکانیزم امتیازدهی قابل‌تفسیری را ارائه می‌شود. در [۱۰۱] مدلی سلسله‌مراتبی برای رتبه‌بندی کیفیت فعالیت‌های آموزشی معلمان ارائه شده است که با تابع هزینه‌ی رتبه‌بندی جفتی آموزش دیده و توانایی تمایز میان جفت فعالیت‌های با سطوح متفاوت را افزایش می‌دهد. به‌طور کلی، روش‌های مبتنی بر مقایسه‌ی زوجی در یادگیری تفاوت‌های نسبی عملکردی بسیار مؤثرند، اما نیاز به داده‌های حاشیه‌نویسی‌شده‌ی گسترده و مقایسه‌های زوجی معنادار دارند که یکی از مهم‌ترین موانع مقیاس‌پذیری آن‌ها به‌شمار می‌رود.

۳،۲،۳. درجه‌بندی

رویکرد درجه‌بندی زمانی به‌کار می‌رود که کیفیت فعالیت بتواند در قالب سطوح گسسته و از پیش تعریف‌شده طبقه‌بندی شود. این روش به‌ویژه در ارزیابی مهارت‌های پزشکی کاربرد دارد؛ به‌طور مثال، عملکردها در سه سطح «مبتدی»، «متوسط» و «حرفه‌ای» درجه‌بندی می‌شوند. یکی از نخستین پژوهش‌ها در این زمینه [۱۰۲]، روشی خودکار برای تحلیل ویدئوهای فعالیت انسانی معرفی کرد که با استخراج ویژگی‌های رفتاری از دنباله‌های تصویری و مقایسه‌ی آن‌ها با الگوهای از پیش تعریف‌شده، میزان مهارت را ارزیابی می‌کرد. تائو و همکاران [۱۰۳] از مدل‌های مخفی مارکوف پراکنده برای طبقه‌بندی حالت‌های جراحی رباتیک استفاده کردند. در این روش، واژه‌نامه‌ای از حرکات پایه استخراج شده و با گرامر مارکوفی ترکیب می‌شود تا مرزهای زمانی و انتقال بین حالت‌ها مدل‌سازی گردد. مشاهده‌ها به‌صورت ترکیب خطی پراکنده از اتم‌های واژه‌نامه بازنمایی می‌شوند که این رویکرد مقاومت بیشتری در برابر نویز و چالش‌های داده‌های با ابعاد بالا فراهم می‌کند. پارمار و موریس [۱۰۴] با تبدیل مسئله ارزیابی کیفیت به یک مسئله طبقه‌بندی دودویی (خوب یا بد)، چارچوبی برای ارزیابی تمرینات حرکتی در درمان فلج مغزی ارائه کردند. در این پژوهش، الگوریتم‌هایی نظیر SVM، شبکه‌های عصبی تک‌لایه و دولایه، درخت‌های تصمیم تقویت‌شده، و تطبیق زمانی پویا برای مقایسه‌ی بازنمایی‌های حوزه‌ی زمان و فرکانس به‌کار رفتند. ضیاء و همکاران [۱۰۵] روشی مبتنی بر داده‌های سینماتیکی ربات، ویژگی‌های استخراج‌شده مبتنی بر آنتروپی و ترکیب وزن‌دار ویژگی‌ها برای ارزیابی مهارت‌های جراحی در آموزش جراحی با کمک ربات معرفی کردند و نشان دادند که روش آن‌ها عملکرد بهتری نسبت به مدل‌های مخفی مارکوف دارد. در ادامه، آنان [۱۰۶] با بهره‌گیری از معیارهای آنتروپی استخراج‌شده از داده‌های ویدئویی و شتاب‌سنج، و با ترکیب فریم‌های ویدئویی و حرکت اسکلتی، بهبود قابل‌توجهی در ارزیابی خودکار مهارت‌های جراحی به‌دست آوردند. اوگاتا و همکاران [۱۰۷] روشی مبتنی بر ماتریس‌های فاصله زمانی برای طبقه‌بندی کیفیت فعالیت اسکوات ارائه کردند.

---

[۱] Temporal Segmentation Attention (TSA)

[۲] procedure-aware cross-attention

[۳] fine-grained contrastive regression

در این روش، ابتدا با استفاده از مدل تخمین حالت، مختصات مفاصل کلیدی بدن در هر فریم استخراج می‌شود. سپس، موقعیت‌های اسکلتی نرمال‌سازی شده و ماتریس‌های فاصله بین مفاصل محاسبه می‌گردند که نسبت به تغییرات فردی، جابه‌جایی کلی، و چرخش‌های جهانی مقاوم هستند. این ماتریس‌ها، روابط مکانی–زمانی بین جفت‌های مفاصل بدن را در طول زمان ثبت می‌کنند. برای طبقه‌بندی، از شبکه‌های کانولوشنی یک‌بعدی استفاده می‌شود تا ویژگی‌های زمانی از این ماتریس‌ها استخراج گردد. در پژوهشی دیگر [۱۰۸]، شاخص پراکندگی حالت انسان برای شناسایی فریم‌های کلیدی معرفی گردید و از شبکه‌های بخش‌بندی زمانی برای طبقه‌بندی فعالیت‌ها بهره گرفته شد. همچنین، لی و همکاران [۱۰۹] یک شبکه سیامی کانولوشنی گرافی برای ارزیابی تمرینات توان‌بخشی فیزیکی پیشنهاد کردند. این روش با استفاده از شبکه‌های گرافی کانولوشنی و ساختار سیامی، ویژگی‌های مکانی–زمانی را از داده‌های اسکلتی استخراج کرده و امکان مقایسه جفتی فعالیت‌های آزمایشی با فعالیت‌های استاندارد را فراهم می‌سازد. داده‌های اسکلتی ابتدا به ویژگی‌های فشرده تبدیل شده و سپس با استفاده از تابع هزینه سیامی، کیفیت نسبی حرکات ارزیابی می‌شود. در نهایت، جدول (۱) خلاصه‌ای از پژوهش‌های شاخص در سه دسته‌ی رگرسیون، مقایسه‌ی زوجی و درجه‌بندی را ارائه می‌کند. این جدول شامل جزئیات مربوط به سال انتشار، حوزه‌ی کاربرد، نوع داده ورودی، روش‌های مورد استفاده، مجموعه‌داده و معیارهای ارزیابی است.

جدول (۱): مروری بر پژوهش‌های کلیدی ارزیابی کیفیت فعالیت انسان.

مرجع	سال	کاربرد	ارزیابی	داده ورودی	معماری یا روش	مجموعه‌داده	معیار ارزیابی
[۱۰۳]	۲۰۱۲	پزشکی	درجه‌بندی	ویدئو	Sparse HMM + Dictionary Learning	California	Acc
[۸]	۲۰۱۴	ورزش	رگرسیون	اسکلت	Pose + DCT + L-SVR	MIT-Dive, MIT-Skate	SRC
[۱۳]	۲۰۱۵	ورزش	رگرسیون	ویدئو، اسکلت	Pose + Approximate Entropy + SVR / LR	MIT Olympics	SRC, mAP
[۱۴]	۲۰۱۶	تناسب اندام	رگرسیون	ویدئو، اسکلت	Self-organizing NN	Powerlifting exercises	Time Series Analysis
[۱۰۴]	۲۰۱۶	فیزیوتراپی	درجه‌بندی	ویدئو، اسکلت	AdaBoosted-Tree, SVM	ویدئوهای ضبط‌شده از فعالیت‌های ورزشی	Acc
[۵]	۲۰۱۷	ورزش	رگرسیون	ویدئو	C3D-SVR, C3D-LSTM, C3D-LSTM-SVR	UNLV-Dive, UNLV-Vault, UNLV-Skate, MIT-Dive, MIT-Skate	SRC
[۸۵]	۲۰۱۷	ورزش	مقایسه زوجی	ویدئو	Conv-LSTM + Gaussian Mixture	BPAD	Maximum F-score
[۱۵]	۲۰۱۸	ورزش	رگرسیون	ویدئو	C3D + FCN	UNLV-Dive, UNLV-Vault, MIT-Skate	SRC, MED
[۱۷]	۲۰۱۸	ورزش	رگرسیون	ویدئو	ED-TCN + P3D + FC/LR/SVR	UNLV-Dive	SRC
[۸۶]	۲۰۱۸	ورزش، پزشکی	مقایسه زوجی	ویدئو	Siamese architecture + 2S-CNN + Attention	EPIC-Skill, JIGSAWS	Acc
[۱۰۵]	۲۰۱۸	پزشکی	درجه‌بندی	ویدئو	SMT + DFT + DCT + Approximate Entropy	JIGSAWS	Acc
[۱۰۶]	۲۰۱۸	پزشکی	درجه‌بندی	ویدئو	Approximate Entropy + Cross-Approximate Entropy	JIGSAWS	Acc
[۱۹]	۲۰۱۹	ورزش	رگرسیون	ویدئو	C3D + FCN	MIT-Dive, UNLV-Dive, UNLV-Vault	SRC, MED
[۳۸]	۲۰۱۹	ورزش، پزشکی	رگرسیون	ویدئو، اسکلت	I3D + FCN	AQA-7, JIGSAWS	SRC
[۷۲]	۲۰۱۹	ورزش	رگرسیون	ویدئو	C3D + FCN	MTL-AQA	SRC
[۴]	۲۰۱۹	ورزش	رگرسیون	ویدئو	C3D-LSTM	AQA-7	SRC
[۲۳]	۲۰۱۹	ورزش	رگرسیون	ویدئو	C3D + M-LSTM + S-LSTM	MIT-Skate, Fis-V	SRC, MSE

[۸۸]	۲۰۱۹	روزمره، پزشکی	مقایسه زوجی	ویدئو	ResNet101 + Attention	Infant Grasp, EPIC-Skills, JIGSAWS	Acc
[۸۷]	۲۰۱۹	روزمره	مقایسه زوجی	ویدئو	I3D + Attention	EPIC-Skill, BEST	Acc
[۱۰۷]	۲۰۱۹	ورزش	درجه‌بندی	ویدئو، اسکلت	3D-Pose + Temporal Distance Matrix	Squat	Acc
[۲۴]	۲۰۲۰	ورزش	رگرسیون	ویدئو	I3D, ResNet, GCN, Attention	MIT-Skate, RG	SRC
[۴۹]	۲۰۲۰	ورزش، پزشکی	رگرسیون	ویدئو	I3D + Temporal Pooling	AQA-7, MTL-AQA, JIGSAWS	SRC
[۵۵]	۲۰۲۰	ورزش، پزشکی	رگرسیون	ویدئو	I3D + Attention	AQA-7, TASD-2, JIGSAWS	SRC
[۷۶]	۲۰۲۰	پزشکی	رگرسیون	ویدئو	C3D + LSTM + FCN	JIGSAWS	SRC
[۱۰۸]	۲۰۲۰	ورزش	درجه‌بندی	ویدئو	C3D + Two-Stream + TSN	FSD-10	Acc
[۲۰]	۲۰۲۱	ورزش	رگرسیون	ویدئو	ED-TCN + P3D	UNLV-Dive	SRC, MSE
[۲۵]	۲۰۲۱	ورزش	رگرسیون	ویدئو	I3D + Attention	UNLV-Dive, UNLV-Vault	SRC
[۲۶]	۲۰۲۱	ورزش	رگرسیون	ویدئو	I3D + Self-Attention	AQA-7, MTL-AQA	SRC
[۲۷]	۲۰۲۱	ورزش	رگرسیون	ویدئو	C3D	MTL-AQA	SRC
[۴۵]	۲۰۲۱	ورزش	رگرسیون	ویدئو	Multi-stream ST-GCN	MTL-AQA, SMART	SRC
[۴۷]	۲۰۲۱	ورزش	رگرسیون	ویدئو، اسکلت	I3D, HRNet, Attention, TCN, SCN	AQA-7	SRC
[۷۹]	۲۰۲۱	ورزش	رگرسیون	ویدئو	I3D + Temporal Cycle-Consistency (TCC)	MTL-AQA	SRC
[۷۷]	۲۰۲۱	پزشکی	رگرسیون	ویدئو	ResNet-101 + SSL	JIGSAWS	SRC
[۸۹]	۲۰۲۱	ورزش	مقایسه زوجی	ویدئو	C3D + Siamese-LSTM network	UNLV-Dive, UNLV-Vault, MTL-AQA	SRC, MSE
[۹۰]	۲۰۲۱	ورزش، پزشکی	مقایسه زوجی	ویدئو	I3D + GART	AQA-7, MTL-AQA, JIGSAWS	SRC, R-ℓ 2
[۵۷]	۲۰۲۱	موسیقی	درجه‌بندی	ویدئو، صوت	3D CNN, ResNet-18	PISA	Acc
[۶۶]	۲۰۲۱	ورزش، پزشکی	رگرسیون	ویدئو	I3D + Adaptive Net	AQA-7, JIGSAWS	SRC
[۲۲]	۲۰۲۲	ورزش	رگرسیون	ویدئو	TimeSformer + Transformer	MTL-AQA	SRC
[۲۸]	۲۰۲۲	ورزش	رگرسیون	ویدئو	3D and (2+1)D ResNets	MTL-AQA	SRC
[۲۹]	۲۰۲۲	ورزش	رگرسیون	ویدئو	I3D + Time- aware Attention	MTL-AQA	SRC
[۳۰]	۲۰۲۲	ورزش	رگرسیون	ویدئو	VST + Grade-aware Decoder	RG, Fis-V	SRC
[۳۹]	۲۰۲۲	ورزش	رگرسیون	اسکلت	ST-GCN + LSTM	MIT-Skate, FIS-V	SRC, MSE
[۵۰]	۲۰۲۲	ورزش، پزشکی	رگرسیون	ویدئو	I3D + Attention	MTL-AQA, FineDiving, JIGSAWS	SRC, R-ℓ 2
[۸۰]	۲۰۲۲	ورزش	رگرسیون	ویدئو	Pose Contrastive Learning + Motion Disentangling	MTL-AQA	SRC
[۸۲]	۲۰۲۲	ورزش، پزشکی	رگرسیون	ویدئو	I3D + FCN	MTL-AQA, Rhy.Gym., JIGSAWS	SRC
[۱۱۰]	۲۰۲۲	پزشکی	رگرسیون	ویدئو	TCN	JIGSAWS, NETS	SRC
[۷۸]	۲۰۲۲	ورزش	رگرسیون	ویدئو	PSGCN-RTCN	MIT-Skate, UNLV-Dive, UNLV-Vault	SRC
[۹۱]	۲۰۲۲	ورزش	مقایسه زوجی	ویدئو	I3D + Transformer	MTL-AQA, AQA-7, JIGSAWS	SRC, R-ℓ 2
[۹۲]	۲۰۲۲	ورزش	مقایسه زوجی	ویدئو	ResNet + Temporal Encoder	AQA-7, MTL-AQA	SRC

[۹۹]	۲۰۲۲	ورزش	مقایسه زوجی	ویدئو	I3D + TSA	FineDiving	SRC, R-ℓ 2
[۷]	۲۰۲۳	ورزش، پزشکی	رگرسیون	ویدئو	I3D + GCN	AQA-7, MTL-AQA, JIGSAWS	SRC, R-ℓ 2
[۲۱]	۲۰۲۳	ورزش	رگرسیون	ویدئو	ED-TCN + P3D	UNLV-Dive	SRC, MSE, MED
[۳۱]	۲۰۲۳	ورزش	رگرسیون	ویدئو	3D CNN + M-LSTM + S-LSTM	MIT-Skate, Fis-V	SRC, MSE
[۳۲]	۲۰۲۳	ورزش، پزشکی	رگرسیون	ویدئو	RNN + Attention	AQA-7, JIGSAWS	MAE
[۳۳]	۲۰۲۳	ورزش	رگرسیون	ویدئو	(2+1)D ResNet	FineDiving	SRC, R-ℓ 2
[۳۴]	۲۰۲۳	ورزش	رگرسیون	ویدئو	I3D, Transformer, Attention	RFSJ	SRC, R-ℓ 2
[۴۰]	۲۰۲۳	ورزش	رگرسیون	ویدئو	I3D + GCN + Attention	LOGO	SRC, R-ℓ 2
[۴۱]	۲۰۲۳	ورزش	رگرسیون	اسکلت	GCN	MIT-Skating, RG	SRC
[۴۳]	۲۰۲۳	تناسب اندام	رگرسیون	اسکلت	GCN	SQUATAI-HUB FITNESS	F1
[۵۱]	۲۰۲۳	ورزش	رگرسیون	ویدئو	VST	FineFS, Fis-V	SRC
[۵۲]	۲۰۲۳	ورزش	رگرسیون	ویدئو	ResNet + Temporal Encoder	AQA-7, MTL-AQA	SRC
[۵۶]	۲۰۲۳	ورزش، پزشکی، روزمره	رگرسیون	ویدئو، اسکلت	I3D + Attention	TASD-2, PaSk, EPIC-Skills, BEST, JIGSAWS	SRC, Acc
[۵۸]	۲۰۲۳	ورزش	رگرسیون	ویدئو	VST + Transformer	MTL-AQA, Fis-V, FS1000	SRC, MSE
[۵۹]	۲۰۲۳	ورزش	رگرسیون	ویدئو، صوت	MLP-based Multimodal	Fis-V	SRC
[۶۷]	۲۰۲۳	ورزش، پزشکی، روزمره	رگرسیون	ویدئو	I3D + Attention	AQA-7, JIGSAWS, EPIC-Skill, BEST	SRC, Acc
[۷۴]	۲۰۲۳	ورزش	رگرسیون	ویدئو	P3D + CNN + TCN	MIT-Skate	SRC, PCC
[۹۳]	۲۰۲۳	ورزش	مقایسه زوجی	ویدئو	I3D + Transformer	AQA-7, MTL-AQA, FineDiving	SRC, R-ℓ 2
[۹۴]	۲۰۲۳	رقص	مقایسه زوجی	اسکلت	2D Pose + Visual Codebook	CDRG	Precision
[۹۵]	۲۰۲۳	ورزش	مقایسه زوجی	ویدئو	I3D + FCN	FineDiving	SRC
[۱۰۹]	۲۰۲۳	توان‌بخشی	درجه‌بندی	اسکلت	GCN-Siamese	UI-PRMD, IRDS	Acc
[۳۴]	۲۰۲۴	ورزش	رگرسیون	ویدئو	I3D + TCN	UNLV-Dive, AQA-7	SRC
[۳۷]	۲۰۲۴	ورزش، پزشکی	رگرسیون	ویدئو	I3D + Transformer	AQA-7, MTL-AQA, FineDiving, JIGSAWS	SRC, R-ℓ 2
[۵۳]	۲۰۲۴	ورزش، پزشکی	رگرسیون	ویدئو	I3D	AQA-7, MTL-AQA, JIGSAWS	SRC
[۵۴]	۲۰۲۴	ورزش، پزشکی	رگرسیون	ویدئو	DAG + GNN	MTL-AQA, FineDiving, JIGSAWS	SRC
[۶۰]	۲۰۲۴	ورزش	رگرسیون	ویدئو	SSL Temporal Parsing + Visual-Semantic	MTL-AQA, Fis-V, RG, FineFS	SRC
[۶۱]	۲۰۲۴	ورزش	رگرسیون	ویدئو	Dual Transformers	MTL-AQA	SRC
[۶۲]	۲۰۲۴	ورزش	رگرسیون	ویدئو	I3D + SSL	UNLV-Diving, AQA-7	SRC
[۶۳]	۲۰۲۴	توان‌بخشی	رگرسیون	ویدئو	GCN + Attention	KIMORE	SRC
[۶۴]	۲۰۲۴	ورزش	رگرسیون	ویدئو، اسکلت	Discriminative Non-local Attention	AGF-Olympics	SRC
[۶۵]	۲۰۲۴	ورزش	رگرسیون	ویدئو	VST, I3D, AST, UNMT, MAST	Fis-V, RG	SRC

[۶۹]	۲۰۲۴	ورزش، پزشکی، توان‌بخشی	رگرسیون	ویدئو	I3D/R(2+1)D-18+3D-Adapters + Continual SSL Pretraining	MTL-AQA, FineDiving, JIGSAWS, PD4T	SRC
[۷۱]	۲۰۲۴	ورزش	رگرسیون	ویدئو	I3D + Manifold Projector	UNLV-Dive, MTL-AQA, FineDiving	SRC
[۷۳]	۲۰۲۴	ورزش	رگرسیون	ویدئو	Hierarchical NeuroSymbolic	FineDiving	متخصصان
[۸۱]	۲۰۲۴	ورزش	رگرسیون	ویدئو	Video Swin Transformer + SSL-CATE	MTL-AQA	SRC
[۸۳]	۲۰۲۴	ورزش	رگرسیون	ویدئو	Teacher-Student + Self-Supervised Sub-action Parsing	MTL-AQA, RG, FineDiving, FineFS	SRC
[۸۴]	۲۰۲۴	ورزش، پزشکی	رگرسیون	ویدئو	I3D + Teacher-Reference-Student + Cross-Attention	MTL-AQA, RG, JIGSAWS	SRC
[۱۱۱]	۲۰۲۴	ورزش	رگرسیون	ویدئو، اسکلت	Multi-task HAU	SkatingVerse	SRC
[۷۰]	۲۰۲۴	ورزش	رگرسیون	ویدئو	I3D + Distillation	AQA-7, MTL-AQA	SRC
[۷۵]	۲۰۲۴	ورزش	رگرسیون	ویدئو	I3D + Attention	Fis-V, RG	SRC
[۹۶]	۲۰۲۴	ورزش	مقایسه زوجی	ویدئو	RegNet-Y + Bi-GRU	FineDiving	SRC
[۹۷]	۲۰۲۴	ورزش	مقایسه زوجی	ویدئو	I3D + PSNet + ResNet + Transformer Cross-Attention	MTL-AQA, FineDiving	SRC
[۹۸]	۲۰۲۴	روزمره	مقایسه زوجی	ویدئو	Transformer + Co-Attention	EPIC-Skill	Acc
[۱۰۰]	۲۰۲۴	ورزش	م زوجی	ویدئو	I3D + Attention	FineDiving	SRC
[۱۰۱]	۲۰۲۴	آموزش	م زوجی	ویدئو	I3D + Siamese Network	TAQR	SRC
[۴۳]	۲۰۲۴	توان‌بخشی	درجه‌بندی	اسکلت	GCN	UI-PRMD, KIMORE, EHE	Acc
[۱۱۲]	۲۰۲۴	ورزش	درجه‌بندی	ویدئو	VST + Grade Prototypes + Sub-grade Prototypes	Fis-V, RG	SRC
[۴۸]	۲۰۲۵	ورزش	رگرسیون	ویدئو	Rating-guided Attention + Consistency Preserving Constraints	AQA-7, MTL-AQA	SRC
[۳۶]	۲۰۲۵	ورزش	رگرسیون	ویدئوهای سه‌جریانی	Hierarchical Joint Training + Replay-Guided Contrastive Transformer	RFSJ	SRC, R-ℓ 2

شکل (۳): نمونه‌هایی از فعالیت‌های موجود در مجموعه‌داده‌های مرتبط با حوزه‌های مختلف ارزیابی کیفیت فعالیت.

۴. مجموعه‌داده‌ها و معیارهای ارزیابی

در قسمت ابتدایی این بخش، مجموعه‌داده‌های مبتنی بر ویدئو برای AQA در حوزه‌های کاربردی مختلف به‌صورت نظام‌مند گردآوری شده‌اند. در ادامه، شرح مفصلی از معیارهای ارزیابی که معمولاً در وظایف AQA استفاده می‌شوند، جمع‌بندی می‌گردد.

۱,۴. مجموعه‌داده‌های ارزیابی کیفیت فعالیت

مجموعه‌داده‌های معیار در مقیاس بزرگ، نقشی محوری در پیشبرد پژوهش‌های علوم کامپیوتر ایفا می‌کنند. تنوع در حوزه‌ی کاربرد، نوع داده و نحوه‌ی برچسب‌گذاری، این مجموعه‌ها را به منابعی کلیدی برای آموزش و ارزیابی مدل‌ها تبدیل کرده است. پیشرفت‌های اخیر در حوزه AQA تا حد زیادی مرهون توسعه مجموعه‌داده‌های دقیق و باکیفیت است. این مجموعه‌داده‌ها دامنه‌ی گسترده‌ای از کاربردها را در بر می‌گیرند؛ از پزشکی و توان‌بخشی گرفته تا فعالیت‌های روزمره، تناسب اندام، موسیقی، تولید صنعتی، رقص، آموزش و به‌ویژه ورزش که بیشترین سهم را در میان این حوزه‌ها داشته و به‌عنوان بستر اصلی سناریوی پژوهش‌های AQA شناخته می‌شود (شکل (۳)). این سازماندهی حوزه‌محور به پژوهشگران کمک می‌کند تا مجموعه‌داده‌هایی متناسب با هدف و دامنه پژوهش خود انتخاب کنند. ویژگی‌های کلیدی این مجموعه‌داده‌ها، شامل حوزه کاربردی، سال انتشار، نوع داده‌های ارائه‌شده (مانند اطلاعات اسکلتی، ویدئویی یا صوتی)، تعداد نمونه‌ها، کلاس‌های فعالیت، میانگین مدت‌زمان هر نمونه و نوع برچسب‌گذاری در جدول (۲) خلاصه شده است.

جدول (۲): خلاصه‌ای از مجموعه‌داده‌های ارزیابی کیفیت فعالیت.

نام مجموعه‌داده	حوزه	سال	ورودی	نمونه‌ها	نوع	زمان	برچسب‌گذاری‌ها
MIT-Dive [۸]	ورزش	۲۰۱۴	ویدئو، اسکلت دوبعدی	۱۵۹	۱	۲,۵ ثانیه	امتیاز
MIT-Skate [۸]	ورزش	۲۰۱۴	ویدئو، اسکلت دوبعدی	۱۵۰	۱	۱۷۵ ثانیه	امتیاز
UNLV-Dive [۵]	ورزش	۲۰۱۷	ویدئو	۳۷۰	۱	۳,۸ ثانیه	امتیاز
UNLV-Vault [۵]	ورزش	۲۰۱۷	ویدئو	۱۷۶	۱	۲,۸ ثانیه	امتیاز

UNLV-Skate [۵]	ورزش	۲۰۱۷	ویدئو	۱۷۱	۱	۷۵ فریم	امتیاز
BPAD [۸۵]	ورزش	۲۰۱۷	ویدئو	۴۸	۱	۱۳ دقیقه	امتیاز
AQA-7 [۴]	ورزش	۲۰۱۹	ویدئو	۱۱۸۹	۷	۶,۷ ثانیه	امتیاز، فعالیت
MTL-AQA [۷۲]	ورزش	۲۰۱۹	ویدئو	۱۴۱۲	۱۶	۴,۱ ثانیه	امتیاز، فعالیت، توضیحات
Squat [۱۰۷]	ورزش	۲۰۱۹	ویدئو	۴۶۲۴	۱	۱۰ ثانیه	برچسب دودویی
Fis-V [۲۳]	ورزش	۲۰۲۰	ویدئو	۵۰۰	۱	۱۷۰ ثانیه	امتیاز (TES و PCS)
RG [۲۴]	ورزش	۲۰۲۰	ویدئو	۱۰۰۰	۴	۹۵ ثانیه	امتیاز، فعالیت
TASD-2 [۵۵]	ورزش	۲۰۲۰	ویدئو	۶۰۶	۲	۴,۱ ثانیه	امتیاز، فعالیت
FSD-10 [۱۰۸]	ورزش	۲۰۲۰	ویدئو	۱۴۸۴	۱۰	۳۰ فریم بر ثانیه	امتیاز، فعالیت
FineGym [۱۱۳]	ورزش	۲۰۲۰	ویدئو	۱۱۶۷	۴	۵۵ ثانیه	امتیاز، فعالیت، زیرفعالیت
FR-FS [۲۶]	ورزش	۲۰۲۱	ویدئو	۴۱۷	۱	۱۰۳ فریم	درجه، فعالیت
SMART [۴۵]	ورزش	۲۰۲۱	ویدئو	۵۰۰۰	۱۰	۴۲۰ فریم	امتیاز، فعالیت
FineDiving [۹۹]	ورزش	۲۰۲۲	ویدئو	۳۰۰۰	۵۲	۴,۲ ثانیه	امتیاز، فعالیت، زیرفعالیت
FineFS [۵۱]	ورزش	۲۰۲۳	ویدئو، اسکلت دوبعدی و سه‌بعدی	۱۱۶۷	۴	۲۱۵ ثانیه	امتیاز، فعالیت، زیرفعالیت
LOGO [۴۰]	ورزش	۲۰۲۳	ویدئو	۲۰۰	۱۲	۲۰۴,۲ ثانیه	امتیاز، فعالیت، آرایش تیمی
PaSk [۵۶]	ورزش	۲۰۲۳	ویدئو	۱۰۱۸	۱	۱۰,۷ ثانیه	امتیاز
RFSJ [۳۵]	ورزش	۲۰۲۳	ویدئو، بازپخش	۱۳۰۴	۲۳	۹۳ فریم	امتیاز، فعالیت
FS1000 [۵۹]	ورزش	۲۰۲۳	ویدئو، صوت	۱۶۰۴	۸	۲۰۰ ثانیه	امتیاز (TES و PCS)، فعالیت
FineDiving-HM [۹۷]	ورزش	۲۰۲۴	ویدئو	۳۰۰۰	۵۲	۴,۲ ثانیه	امتیاز، فعالیت، ماسک پیش‌زمینه انسانی
AGF-Olympics [۶۴]	ورزش	۲۰۲۴	ویدئو، اسکلت	۵۰۰	۱	۲۵۰۰ فریم	امتیاز، جنسیت
SkatingVerse [۱۱۱]	ورزش	۲۰۲۴	ویدئو	۱۶۸۷	۱۱	۳۹۷,۹ ثانیه	امتیاز، فعالیت
BASKET [۱۱۴]	ورزش	۲۰۲۵	ویدئو	۳۲۲۳۲	۲۰	۵۰۰ ثانیه	درجه، فعالیت
JIGSAWS [۱۱۵]	پزشکی	۲۰۱۴	ویدئو	۱۰۳	۳	۹۲ ثانیه	امتیاز، فعالیت
NETS [۱۱۰]	پزشکی	۲۰۲۲	ویدئو	۱۰۰	۱	۲۵ فریم بر ثانیه	امتیاز
Walking Gait [۱۱۶]	توان‌بخشی	۲۰۱۸	ویدئو، زوایا و موقعیت‌های مفاصل	۸۱	۹	۱۲۰۰ فریم	درجه
UI-PRMD [۱۱۷]	توان‌بخشی	۲۰۱۸	ویدئو، اسکلت سه‌بعدی، زوایا و موقعیت‌های مفاصل	۱۰۰	۱۰	نامشخص	برچسب دودویی، فعالیت
KIMORE [۱۱۸]	توان‌بخشی	۲۰۱۹	ویدئو، زوایا و موقعیت‌های مفاصل	۳۵۳	۵	۲۹,۹ ثانیه	امتیاز، فعالیت
EHE [۱۱۹]	توان‌بخشی	۲۰۲۱	ویدئو، اسکلت سه‌بعدی، زوایا و موقعیت‌های مفاصل	۸۶۹	۶	نامشخص	برچسب دودویی، فعالیت
MMASD [۱۲۰]	توان‌بخشی	۲۰۲۳	جریان نوری، اسکلت دوبعدی، اسکلت سه‌بعدی	۱۳۱۵	۱۱	۷ ثانیه	امتیاز، فعالیت
FineRehab [۱۲۱]	توان‌بخشی	۲۰۲۴	ویدئو، اسکلت سه‌بعدی، زوایا و موقعیت‌های مفاصل	۴۲۱۵	۱۶	نامشخص	امتیاز، فعالیت
EPIC-Skill [۸۶]	روزمره	۲۰۱۸	اسکلت سه‌بعدی زوایا و موقعیت‌های مفاصل	۲۱۶	۴	۸۶,۶ ثانیه	رتبه‌بندی، فعالیت
Best [۸۷]	روزمره	۲۰۱۹	ویدئو	۵۰۰	۵	۱۸۷,۶ ثانیه	رتبه‌بندی، فعالیت
Infant Grasp [۸۸]	روزمره	۲۰۱۹	ویدئو	۹۴	۱	۶۰ فریم بر ثانیه	رتبه‌بندی

EgoExo4D [۱۲۲]	روزمره	۲۰۲۴	ویدئو، صدا، اسکلت سه‌بعدی	۱۲۲۴	۴	۴۷ ثانیه	سطح مهارت، توضیحات
UMONS-TAICHI [۱۲۳]	تناسب اندام	۲۰۱۸	ویدئو، اسکلت	۲۲۰۰	۱۳	نامشخص	درجه، فعالیت
Fitness-AQA [۸۰]	تناسب اندام	۲۰۲۲	ویدئو	۲۱۲۸۴	۳	۴٫۱ ثانیه	برچسب دودویی، فعالیت
EgoExo-Fitness [۱۲۴]	تناسب اندام	۲۰۲۴	ویدئو	۶۱۳۱	۱۲	۱۸٫۸ ثانیه	امتیاز، فعالیت، توضیحات
PISA [۵۷]	موسیقی	۲۰۲۱	ویدئو	۹۹۲	۱	۱۶۰ فریم	درجه، سختی آهنگ
Assembly101 [۶]	تولید صنعتی	۲۰۲۲	ویدئو	۴۳۲۱	۱۰۱	۴۲۶ ثانیه	درجه، فعالیت
CDRG [۹۴]	رقص	۲۰۲۳	ویدئو	۲۴۰	۱۲	۱۴٫۷ ثانیه	رتبه، فعالیت
TAQR [۱۰۱]	آموزش	۲۰۲۴	ویدئو	۳۰۰	۴	۲۵ فریم بر ثانیه	رتبه نسبی
GAIA [۱۲۵]	چندحوزه‌ای	۲۰۲۴	ویدئو	۹۱۸۰	۵۱۰	۲٫۸ ثانیه	امتیاز، فعالیت

۱،۱،۴. مجموعه‌داده‌های ورزشی

مجموعه‌داده‌های ورزشی از منابع بنیادی و پراستفاده در پژوهش‌های مرتبط با AQA به‌شمار می‌روند. این مجموعه‌ها معمولاً شامل حرکات پیچیده ورزشی بوده و گستره‌ای از حدود ده نوع رشته ورزشی مختلف را در بر می‌گیرند. در ادامه، مروری جامع بر مهم‌ترین مجموعه‌داده‌های مورد استفاده در این حوزه ارائه می‌شود.

مجموعه‌داده‌های MIT-Dive و MIT-Skate [۸]: این مجموعه‌داده‌ها از نخستین منابع شاخص در حوزه‌ی AQA ورزشی هستند و شامل ویدئوهای شیرجه و اسکیت نمایشی در سطح رقابت‌های جهانی و المپیک همراه با امتیازهای رسمی داوران‌اند. مجموعه MIT-Dive شامل ۱۵۹ ویدئوی شیرجه با نرخ ۶۰ فریم بر ثانیه و میانگین مدت ۲٫۵ ثانیه (در مجموع ۲۵۰۰۰ فریم) است. هر ویدئو به‌طور متوسط ۱۵۰ فریم دارد و از دو زاویه‌ی جلو و کنار فیلم‌برداری شده است. دامنه‌ی امتیازها بین ۲۰ تا ۱۰۰ متغیر است. مجموعه MIT-Skate نیز شامل ۱۵۰ ویدئوی اسکیت نمایشی با نرخ ۲۴ فریم بر ثانیه و میانگین مدت ۱۷۵ ثانیه است. کل مجموعه‌داده حدود ۶۳۰۰۰۰ فریم دارد. هر ویدئو به طور متوسط شامل ۴۲۰۰ فریم است و دارای دامنه امتیاز ۰ تا ۱۰۰ و زوایه‌های دید متنوع و پویا می‌باشد.

مجموعه‌داده‌های UNLV-Dive، UNLV-Vault و UNLV-Skate [۵]: مجموعه‌داده UNLV-Dive نسخه گسترش‌یافته‌ی MIT-Dive است و شامل ۳۷۰ ویدئوی شیرجه‌ی انفرادی از رقابت‌های سکوی ۱۰ متری المپیک ۲۰۱۲ لندن می‌باشد. ویدئوها با وضوح ۳۲۰ × ۲۴۰ پیکسل، میانگین مدت ۳٫۸ ثانیه و امتیاز کلی (۲۱٫۶ تا ۱۰۲٫۶)، سطح دشواری (۲٫۷ تا ۴٫۱) و برچسب‌های بخش‌بندی زیرفعالیت‌ها ارائه شده‌اند. از میان آن‌ها، ۳۰۰ ویدئو برای آموزش و ۷۰ ویدئو برای ارزیابی مدل استفاده شده است. مجموعه UNLV-Vault شامل ۱۷۶ ویدئوی ژیمناستیک با میانگین مدت ۲٫۸ ثانیه (حدود ۷۵ فریم) است که دارای برچسب‌های امتیاز اجرا (۰ تا ۲۰)، امتیاز دشواری (۰ تا ۱۰) و امتیاز نهایی می‌باشد. مجموعه UNLV-Skate نیز شامل ۱۷۱ ویدئوی اسکیت نمایشی با میانگین طول ۴۵۰۰ فریم است.

مجموعه‌داده BPAD [۸۵]: BPAD مخفف Basketball Performance Assessment Dataset شامل ۴۸ ویدئوی دید اول‌شخص از بازیکنان بسکتبال است که به دو بخش تقسیم شده‌اند: ۲۴ ویدئو مربوط به روز نخست برای آموزش و ۲۴ ویدئو مربوط به روز دوم برای تست. این مجموعه با برچسب‌های مرتب‌سازی زوجی به‌منظور ارزیابی مهارت بازیکنان حاشیه‌نویسی شده و دیدگاهی منحصربه‌فرد برای تحلیل عملکرد ارائه می‌دهد. پس از آنکه بازیکنان حرفه‌ای بسکتبال ویدئوها را به صورت جفتی مرتب‌سازی کردند، در مجموع ۲۵۰ جفت آموزش و ۲۵۰ جفت تست تولید شد.

مجموعه‌داده AQA-7 [۴]: این مجموعه شامل نمونه‌هایی از هفت رشته ورزشی مختلف و در مجموع ۱۱۸۹ ویدئو استخراج‌شده از رویدادهای المپیک‌های تابستانی و زمستانی است. داده‌ها شامل ۳۷۰ ویدئوی شیرجه انفرادی از سکوی ۱۰ متری، ۱۷۶ ویدئوی پرش ژیمناستیک، ۱۷۵ ویدئوی اسکی، ۲۰۶ ویدئوی اسنوبرد، ۸۸ ویدئوی شیرجه هماهنگ از تخته ۳ متری، ۹۱ ویدئوی شیرجه هماهنگ از سکوی ۱۰ متری، و ۸۳ ویدئوی ترامپولین است. میانگین طول ویدئوها در میان این دسته‌ها متغیر بوده و از ۸۷ فریم در پرش ژیمناستیک تا ۶۳۴ فریم در ترامپولین تغییر می‌کند. میانگین مدت ویدئوها ۶٫۷ ثانیه است. ویدئوهای پرش ژیمناستیک، اسکی و اسنوبرد از زوایای دید متنوع‌تری نسبت به رشته‌های شیرجه و ترامپولین برخوردارند. ویژگی بارز این مجموعه‌داده، بهره‌گیری از نظام امتیازدهی واقعی هر رشته بر اساس قواعد داوری رسمی است؛ به‌طوری‌که امتیاز نهایی از حاصل‌ضرب امتیاز اجرا در امتیاز دشواری به‌دست می‌آید. دامنه امتیازها در رشته‌های مختلف عبارت‌اند از: شیرجه انفرادی از سکوی ۱۰ متری (۲۱٫۶ تا ۱۰۲٫۶)، پرش ژیمناستیک (۱۲٫۳ تا ۱۶٫۸۷)، اسکی بیگ ایر (۸ تا ۵۰)، اسنوبرد بیگ ایر (۸ تا ۵۰)، شیرجه هماهنگ از تخته ۳ متری (۴۶٫۲ تا ۱۰۴٫۸۸)، شیرجه هماهنگ از سکوی ۱۰ متری (۴۹٫۸ تا ۹۹٫۳۶) و ترامپولین (۶٫۷۲ تا ۶۲٫۹۹). در مجموع، ۸۰۳ ویدئو برای آموزش و ۳۰۳ ویدئو برای ارزیابی مدل مورد استفاده قرار گرفته است.

مجموعه‌داده MTL-AQA [۷۲]: این مجموعه به‌عنوان یکی از بزرگ‌ترین منابع و نخستین مجموعه‌داده چند وظیفه‌ای در حوزه AQA شناخته می‌شود. مجموعه MTL-AQA شامل ۱۶ رویداد بین‌المللی و ۱۴۱۲ ویدئوی شیرجه است که انواع مختلفی از شیرجه‌های انفرادی و هماهنگ، اجراهای مردان و زنان، سکوی پرش ۳ متری و سکوی ۱۰ متری، و زوایای دید متنوع با پس‌زمینه‌های شلوغ را در بر می‌گیرد. میانگین مدت هر ویدئو ۴٫۱ ثانیه است. برچسب‌گذاری‌ها شامل امتیاز دشواری، امتیاز واقعی اعطا شده توسط هفت داور، کلاس‌های فعالیت ریزدانه (شامل پنج زیرفعالیت) و تفسیرهای توصیفی از اجرای ورزشکاران است که اغلب توسط مفسران حرفه‌ای تلویزیونی ارائه شده‌اند. در مجموع، ۱۰۵۹ ویدئو برای آموزش و ۳۵۳ ویدئو برای تست استفاده می‌شود.

مجموعه‌داده Squat [۱۰۷]: این مجموعه‌داده با هدف ارزیابی کیفیت اجرای حرکت اسکوات طراحی شده و شامل چهار زیرمجموعه مجزا است: تک‌فردی، چندفردی، تغییر پس‌زمینه و یوتیوب. زیرمجموعه تک‌فردی با ۲۰۰۱ ویدئو از اجرای اسکوات یک فرد است؛ هر ویدئو حدود ۱۰ ثانیه (معادل ۳۰۰ فریم و ۳ تا ۵ تکرار اسکوات) است. زیرمجموعه چندفردی شامل ۵۹۹ ویدئو از اجرای اسکوات توسط هفت فرد مختلف، همراه با مجموعه‌ای از ویدئوهای گردآوری‌شده از یوتیوب که ممکن است یک یا چند خطای حرکتی داشته باشند. ویدئوهای بدون خطا در دسته «اسکوات صحیح» قرار گرفته‌اند. زیرمجموعه تغییر پس‌زمینه با ۲۰۰۱ ویدئو برای ارزیابی پایداری الگوریتم‌ها در برابر تغییرات محیطی طراحی شده است. زیرمجموعه یوتیوب شامل ۲۳ ویدئو از منابع عمومی بوده و هر ویدئو به یکی از هفت کلاس (یک فرم صحیح و شش فرم نادرست اسکوات) تعلق دارد. برچسب‌گذاری حرکات توسط متخصصان ورزشی انجام شده و بسته به نوع داده، به‌صورت تک‌برچسبی یا چندبرچسبی صورت گرفته است.

مجموعه‌داده Fis-V [۲۳]: مجموعه Fis-V مخفف Figure Skating Video شامل ۵۰۰ ویدئوی باکیفیت از برنامه‌های کوتاه اسکیت نمایشی تک‌نفره زنان است که از رقابت‌های بین‌المللی معتبر مانند جام NHK، تروفه اریک بومپار[۱] (TEB)، جام چین و مسابقات قهرمانی چهار قاره گردآوری شده‌اند. این مجموعه شامل اجراهای ۱۴۹ ورزشکار از ۲۰ کشور است و میانگین مدت هر ویدئو ۲ دقیقه

[۱] Trophee Eric Bompard (TEB)

و ۵۰ ثانیه (حدود ۴۳۰۰ فریم با نرخ فریم ۲۵ فریم بر ثانیه) می‌باشد. برچسب امتیاز هر ویدئو توسط ۹ داور حرفه‌ای تعیین شده و شامل دو بخش امتیاز TES و PCS است. در مجموع، ۴۰۰ ویدئو برای آموزش و ۱۰۰ ویدئو برای ارزیابی مدل استفاده شده است.

مجموعه‌داده RG [۲۴]: مجموعه‌داده RG مخفف Rhythmic Gymnastics منبعی تخصصی برای ارزیابی کیفیت اجرا در ژیمناستیک ریتمیک است. این مجموعه شامل ۱۰۰۰ ویدئو از سی‌وششمین و سی‌وهفتمین مسابقات بین‌المللی ژیمناستیک هنری بوده و چهار نوع فعالیت شامل توپ، چوب‌دستی، حلقه و روبان را پوشش می‌دهد. هر فعالیت دارای ۲۵۰ ویدئو با میانگین مدت ۹۵ ثانیه و نرخ تصویربرداری ۲۵ فریم بر ثانیه است. ویدئوها با امتیازهای داوران حرفه‌ای شامل امتیاز دشواری، امتیاز اجرا و امتیاز کل، برچسب‌گذاری شده‌اند. برای هر نوع فعالیت، ۲۰۰ ویدئو برای آموزش و ۵۰ ویدئو برای ارزیابی در نظر گرفته شده است.

مجموعه‌داده TASD-2 [۵۵]: این مجموعه شامل ۶۰۶ ویدئوی شیرجه هماهنگ از تخته ۳ متری و سکوی ۱۰ متری است که از زاویه دید جلویی و با وضوح ۳۲۰×۲۴۰ ضبط شده‌اند. هر ویدئو با میانگین مدت‌زمان ۴٫۱ ثانیه (حدود ۱۰۲ فریم) برای ارزیابی تعاملی کیفیت اجرای فعالیت طراحی شده است. برچسب‌گذاری شامل فریم‌های آغاز و پایان فعالیت، امتیاز دشواری، امتیاز اجرا، امتیاز هماهنگی و امتیاز نهایی اعطا شده توسط داوران حرفه‌ای حاشیه‌نویسی شده‌اند.

مجموعه‌داده FSD-10 [۱۰۸]: این مجموعه شامل ۱۴۸۴ کلیپ ویدئویی از مسابقات قهرمانی جهانی اسکیت نمایشی ۲۰۱۷ تا ۲۰۱۸ است و ۱۰ نوع فعالیت مختلف در برنامه‌های مردان و زنان را پوشش می‌دهد. هر کلیپ با نرخ ۳۰ فریم بر ثانیه و وضوح ۱۰۸۰ × ۷۲۰ پیکسل ضبط شده و توسط کارشناسان با اطلاعات نوع فعالیت، درجه اجرا و جزئیات اسکیت‌باز حاشیه‌نویسی شده است.

مجموعه‌داده FineGym [۱۱۳]: شامل ویدئوهای باکیفیت از رقابت‌های ژیمناستیک با حاشیه‌نویسی‌های سلسله‌مراتبی دقیق در دو سطح معنایی (رویداد، گروه زیرفعالیت و عنصر) و زمانی (فعالیت، زیرفعالیت) است. این مجموعه ۱۰ دسته رویداد ورزشی (۶ دسته برای مردان و ۴ دسته برای زنان)، با تمرکز بر چهار رویداد زنان برای برچسب‌گذاری دقیق‌تر را پوشش می‌دهد. تعداد نمونه‌ها در هر دسته عنصر بین ۱ تا ۱۶۴۸ متغیر است و از میان ۵۳۰ دسته عنصر تعریف‌شده، ۳۵۴ دسته دارای حداقل یک نمونه هستند. همچنین شامل ۳۰۳ رکورد مسابقه‌ای با مجموع حدود ۷۰۸ ساعت داده است.

مجموعه‌داده FR-FS [۲۶]: مجموعه FR-FS مخفف Fall Recognition in Figure Skating شامل ۴۱۷ ویدئو از مجموعه Fis-V و المپیک زمستانی پیونگ‌چانگ ۲۰۱۸ است. هر ویدئو شامل حرکات کلیدی اسکیت نمایشی نظیر برخاستن، چرخش، و فرود است. از میان این نمونه‌ها، ۲۷۶ ویدئو به فرودهای صحیح و ۱۴۱ ویدئو مربوط به سقوط هستند. برچسب‌گذاری دقیق در سطح فریم انجام شده و شامل برچسب‌های حرکات کلیدی و برچسب‌های دودویی (صحیح/نادرست) برای ارزیابی کیفیت فرود است. برای آزمایش کارآیی مدل، ۵۰ درصد از ویدئوها به‌صورت تصادفی از ویدئوهای سقوط و فرود به عنوان مجموعه آموزش و مجموعه تست انتخاب می‌شوند.

مجموعه‌داده SMART [۴۵]: این مجموعه شامل ۵۰۰۰ ویدئوی ورزشی از ۱۰ نوع فعالیت متنوع، شامل چوب موازنه[۱]، شیرجه، میله‌های ناهم‌سطح، پرش از خرک[۲]، دو با مانع، پرش با نیزه، پرش ارتفاع، بوکس، تمرینات آمادگی جسمانی و بدمینتون است. هر ویدئو

[۱] balance beam

[۲] vault

با میانگین ۴۲۰ فریم و نرخ تصویربرداری مناسب برای تحلیل دقیق حرکات ضبط شده است. داده‌ها دارای برچسب‌های زیرفعالیت‌ها و امتیازهای ارزیابی کیفیت اجرا هستند.

مجموعه‌داده FineDiving [۹۹]: این مجموعه شامل ۳۰۰۰ ویدئوی شیرجه از ۳۰ رقابت مختلف، از جمله المپیک، جام جهانی، قهرمانی جهان و قهرمانی ورزش‌های آبی اروپا، گردآوری‌شده از یوتیوب است. این مجموعه ۵۲ نوع فعالیت، ۲۹ نوع زیرفعالیت و ۲۳ سطح دشواری را پوشش می‌دهد. ویدئوها با میانگین مدت ۴٫۲ ثانیه، شامل اجرای کامل و بازپخش‌های آهسته از زوایای دید مختلف هستند. داده‌ها به‌صورت دقیق و سلسله‌مراتبی از نظر معنایی و زمانی حاشیه‌نویسی شده‌اند. هر ویدئوی دارای دو سطح برچسب‌گذاری است: سطح فعالیت و سطح مرحله. در ساختار معنایی، برچسب «فعالیت» نوع شیرجه را توصیف می‌کند و برچسب «مرحله» به زیرفعالیت‌های متوالی اجرای حرکت اشاره دارد. در ساختار زمانی، برچسب سطح فعالیت، زمان شروع و پایان اجرای کامل شیرجه را مشخص می‌کند، در حالی‌که برچسب سطح زیرفعالیت فریم‌های آغاز و پایان هر زیرفعالیت را تعیین می‌نماید.

مجموعه‌داده FineFS [۵۱]: شامل ۱۱۶۷ ویدئوی RGB اسکیت نمایشی همراه با توالی داده‌های اسکلتی از ۷۲ رویداد بین‌المللی سطح A تحت قوانین اتحادیه بین‌المللی اسکیت (ISU) است. این مجموعه در دو دسته برنامه کوتاه (۷۲۹ ویدئو، حدود ۱۶۰ ثانیه) و اسکیت آزاد (۴۳۸ ویدئو، حدود ۲۴۰ ثانیه) با میانگین کلی ۲۱۵ ثانیه و نرخ ۲۵ فریم بر ثانیه ارائه شده است. برچسب‌گذاری چندسطحی شامل امتیاز TES، PCS، ارزش پایه، درجه اجرا، مهارت‌های اسکیت‌سواری، گذارها، طراحی رقص، تفسیر موسیقی و امتیازهای داوران است.

مجموعه‌داده LOGO [۴۰]: مجموعه‌داده LOGO، مخفف LOng-form GrOup، نخستین منبع جامع برای ارزیابی کیفیت فعالیت گروهی در ویدئوهای بلندمدت ورزشی است. این مجموعه شامل ۲۰۰ ویدئو از ۲۶ رویداد شنای هنری بین‌المللی (۲۰۱۸ تا ۲۰۲۲) است. در هر فریم به‌طور میانگین ۸ ورزشکار حضور دارند و میانگین مدت هر ویدئو ۲۰۴٫۲ ثانیه است. حاشیه‌نویسی‌ها شامل برچسب‌های نوع فعالیت در سطح فریم، مرزهای زمانی فعالیت‌ها، و فرم آرایش‌های تیمی با استفاده از چندضلعی‌های محدب برای نمایش موقعیت ورزشکاران است. در مجموع، ۱۲ نوع برچسب برای فعالیت‌های فردی و ۱۷ نوع برچسب برای آرایش‌های گروهی تعریف شده که همگی از طریق تحلیل دقیق فریم‌به‌فریم تعیین شده‌اند.

مجموعه‌داده PaSk [۵۶]: این مجموعه‌داده برای تکمیل AQA در اسکیت نمایشی زوجی طراحی شده است. در این رشته، تعامل میان دو ورزشکار (نقش حمایتی مرد و نقش اصلی زن) عامل کلیدی در کیفیت اجرا به‌شمار می‌رود. مجموعه PaSk شامل ۱۰۱۸ ویدئوی رسمی از مسابقات تحت نظارت اتحادیه بین‌المللی اسکیت (ISU) است. مدت ویدئوها بین ۱۰۰ تا ۱۰۰۰ فریم متغیر بوده و میانگین آن ۱۰٫۷ ثانیه است. تمامی ویدئوها با کیفیت بالا گردآوری شده‌اند و برچسب‌گذاری آن‌ها شامل تعیین دقیق فریم‌های آغاز و پایان هر فعالیت مستقل و امتیازهای رسمی داوران برای هر اجرا است.

مجموعه‌داده RFSJ [۳۵]: این مجموعه مخفف Replay-based Figure Skating Jumping شامل ۸۸۶ ویدئو از پرش‌های اسکیت نمایشی است که از ۱۰ مسابقه رسمی المپیک و قهرمانی اروپا استخراج و از پلتفرم یوتیوب گردآوری شده‌اند. این مجموعه شامل ۷۶۸ توالی ویدئویی زنده و ۵۳۶ توالی بازپخش است. ویدئوهای بازپخش با زوایای دید و بزرگ‌نمایی متنوع، جزئیات فعالیت‌ها را از دیدگاه‌های مختلف نمایش می‌دهند و به‌صورت جفت‌های متناظر یک‌به‌یک با توالی‌های زنده هم‌تراز شده‌اند. هر ویدئو دارای سه زاویه

دید (یک نمای کلی و دو نمای بازپخش نزدیک) بوده و ۱۰ نوع فعالیت منفرد و ۱۳ نوع فعالیت ترکیبی را پوشش می‌دهد. حاشیه‌نویسی شامل شاخص فریم شروع و پایان برای توالی‌های زنده و بازپخش، ارزش پایه، درجه اجرا و امتیاز نهایی داوران حرفه‌ای است.

مجموعه‌داده FS1000 [۵۹]: مجموعه‌داده FS1000 بزرگ‌ترین منبع ویدئویی باکیفیت در حوزه امتیازدهی اسکیت نمایشی است که شامل ۱۶۰۴ ویدئو از مسابقات بین‌المللی سطح بالا نظیر قهرمانی جهانی ISU، گرند پری ISU و المپیک زمستانی پکن ۲۰۲۲ می‌باشد. این مجموعه هشت دسته شامل برنامه کوتاه تک‌نفره مردان، برنامه کوتاه تک‌نفره زنان، برنامه کوتاه زوجی، اسکیت آزاد تک‌نفره مردان، اسکیت آزاد تک‌نفره زنان، اسکیت آزاد زوجی، رقص ریتمیک، و رقص آزاد را پوشش می‌دهد. ویدئوها با میانگین مدت ۳ دقیقه و ۲۰ ثانیه (حدود ۵۰۰۰ فریم با نرخ ۲۵ فریم بر ثانیه) و بازه زمانی ۲,۵ تا ۴,۳ دقیقه ضبط شده‌اند. هر ویدئو با امتیازهای رسمی داوران، شناسه ورزشکار و دسته‌بندی فعالیت برچسب‌گذاری شده‌اند. در تقسیم‌بندی داده‌ها، ۱۲۴۷ ویدئو برای آموزش و اعتبارسنجی و ۳۵۷ ویدئو برای تست در نظر گرفته شده‌اند. امتیازدهی بر اساس دو مؤلفه اصلی، امتیاز TES و PCS، انجام گرفته است.

مجموعه داده FineDiving-HM [۹۷]: مجموعه‌داده FineDiving-HM نسخه‌ای گسترش‌یافته از مجموعه FineDiving است که شامل ۳۱۲۲۵۶ ماسک پیش‌زمینه فعالیت برای همان ۳۰۰۰ ویدئو است. این ماسک‌ها نواحی هدف فعالیت انسانی را از پس‌زمینه تفکیک می‌کنند. ۲۴۸۷۱۳ ماسک مربوط به شیرجه انفرادی و ۶۳۵۴۳ ماسک مربوط به شیرجه هم‌زمان است. کیفیت حاشیه‌نویسی‌ها توسط سه برچسب‌گذار متخصص در حوزه شیرجه بررسی و تأیید شده است.

مجموعه‌داده AGF-Olympics [۶۴]: مجموعه‌داده AGF-Olympics شامل ۸۳ ساعت ویدئوی ژیمناستیک هنری زمینی از رویدادهای المپیک است که با وضوح ۲۵۶×۲۵۶ پیکسل (نمونه‌برداری‌شده از حداکثر وضوح ۱۹۲۰×۱۰۸۰ پیکسل) و میانگین مدت ۱,۳ تا ۲ دقیقه (حدود ۲۵۰۰ فریم) ارائه شده است. این مجموعه شامل موقعیت‌های اسکلتی ورزشکاران در هر فریم و برچسب‌هایی نظیر امتیاز دشواری، امتیاز اجرا و امتیاز جریمه است. افزون بر این، اطلاعات جانبی همچون نوع رویداد، جنسیت ورزشکار و سال برگزاری نیز ارائه شده است. ویدئوها با زاویه‌های دید و پس‌زمینه‌های متنوع و در شرایطی شامل انسدادهای جزئی ثبت شده‌اند که به بازنمایی واقع‌گرایانه‌تری از سناریوهای دنیای واقعی منجر شده است.

مجموعه‌داده SkatingVerse [۱۱۱]: این مجموعه‌داده وسیع و جامع به‌طور خاص برای درک فعالیت انسان (HAU) در اسکیت نمایشی طراحی شده است. این منبع شامل ۱۶۸۷ ویدئوی رسمی از مسابقات بین‌المللی با مجموع ۱۸۴,۴ ساعت است. این مجموعه با تمرکز بر محیطی کنترل‌شده، سوگیری‌های موجود در مجموعه‌داده‌های پیشین را کاهش داده و امکان انجام همزمان دو وظیفه بازشناسی فعالیت و ارزیابی کیفیت فعالیت را فراهم می‌سازد.

مجموعه‌داده BASKET [۱۱۴]: یکی از بزرگ‌ترین منابع موجود در حوزه بسکتبال است که شامل ۴۴۷۷ ساعت ویدئو از ۳۲۲۳۲ بازیکن در ۲۱ لیگ از بیش از ۳۰ کشور در چهار قاره می‌باشد. این مجموعه، داده‌هایی از شش فصل رقابتی (۲۰۱۷ تا ۲۰۲۳) را در بر می‌گیرد و میانگین مدت هر ویدئو حدود ۵۰۰ ثانیه می‌باشد. از این میان، ۷۵۶۳ ویدئو به بازیکنان زن اختصاص دارد. این مجموعه‌داده با تنوع چشمگیر در جنسیت، سن، سطح مهارت، و موقعیت جغرافیایی، ۲۰ مهارت ریزدانه بسکتبال را پوشش می‌دهد.

۲,۱,۴. مجموعه‌داده‌های مراقبت‌های پزشکی

در مقایسه با حوزه ورزش، مجموعه‌داده‌های AQA مبتنی بر ویدئو در حوزه مراقبت‌های پزشکی بسیار محدودتر هستند. توسعه چنین داده‌هایی با چالش‌های منحصربه‌فردی همراه است، زیرا داده‌های پزشکی به‌شدت تخصصی‌اند و نیازمند تجهیزات پیشرفته و تخصص کارشناسان برای حاشیه‌نویسی دقیق می‌باشند. از آنجا که هر مرکز تحقیقاتی اغلب پلتفرم اختصاصی خود را برای جمع‌آوری داده‌ها ایجاد می‌کند، دسترسی عمومی به داده‌های منبع‌باز در این حوزه بسیار اندک است.

مجموعه‌داده JIGSAWS [۱۱۵]: مجموعه JIGSAWS، نخستین مجموعه‌داده‌ی متن‌باز برای مدل‌سازی مهارت‌های جراحی است که در دانشگاه جانز هاپکینز گردآوری شده و برای تحلیل و ارزیابی حرکات جراحی با استفاده از سیستم رباتیک داوینچی طراحی شده است. این مجموعه شامل ۱۰۳ ویدئو با میانگین ۹۲ ثانیه است که سه عمل جراحی پایه، بخیه‌زنی (۳۹ نمونه)، عبور سوزن (۲۸ نمونه) و گره‌زنی (۳۶ نمونه) توسط هشت جراح در سطوح مهارتی مختلف انجام شده‌اند. داده‌ها شامل ویدئوهای استریو (چپ و راست) و داده‌های حرکتی ۷۶ بعدی (موقعیت، جهت‌گیری، سرعت خطی و زاویه‌ای، و زاویه گیره ابزار) هستند. برچسب‌گذاری در سطح فریم با ۱۵ برچسب فعالیت و امتیاز مهارتی متشکل از شش مؤلفه (هرکدام با مقیاس ۱ تا ۵) انجام شده است.

مجموعه‌داده NETS [۱۱۰]: این مجموعه شامل ۱۰۰ ویدئوی کوتاه از تمرینات نورواندوسکوپی ضبط‌شده با یک شبیه‌ساز جراحی جعبه‌ای است. در این مجموعه، شش متخصص و شش کارآموز با استفاده از ابزار فورسپس بیوپسی و اندوسکوپ، وظیفه‌ی جابه‌جایی شش حلقه را داشتند. ویدئوها با نرخ ۲۵ فریم بر ثانیه ضبط شده و سطح مهارت هر شرکت‌کننده توسط جراح مغز و اعصاب خبره در مقیاس ۱ تا ۱۰ ارزیابی شده است.

۳،۱،۴. مجموعه‌داده‌های توان‌بخشی

بیشتر مجموعه‌داده‌های مرتبط با توان‌بخشی به دلیل دو عامل اصلی به‌صورت عمومی در دسترس نیستند. نخست، مسائل مربوط به حریم خصوصی بیماران یا افراد سالمند در حین انجام فعالیت‌های درمانی آن‌ها که محدودیت‌های اخلاقی و قانونی برای انتشار داده‌ها ایجاد می‌کند و دوم، محدودیت‌های حقوق مالکیت معنوی که اغلب در اختیار مراکز درمانی یا نهادهای سرمایه‌گذار قرار دارد.

مجموعه‌داده Walking Gait [۱۱۶]: این مجموعه شامل ۸۱ نمونه از الگوهای حرکتی راه‌رفتن ۹ شرکت‌کننده در ۹ سطح متفاوت است. برای ایجاد الگوهای راه‌رفتن نامتقارن، هر شرکت‌کننده یک‌بار بدون پد و هشت‌بار با پدهای مختلف زیر پا راه رفته است. هر نمونه شامل ۱۲۰۰ فریم ضبط‌شده توسط حسگر Kinect V2 است.

مجموعه‌داده UI-PRMD [۱۱۷]: این مجموعه شامل ۱۰۰ ویدئو از ۱۰ نوع فعالیت فیزیوتراپی انجام‌شده توسط ۱۰ فرد سالم، با ۱۰ تکرار برای هر حرکت است. داده‌ها هم‌زمان توسط دو سامانه Vicon و Kinect ضبط گردیده‌اند. هر نمونه شامل توالی‌های اسکلتی با اطلاعات موقعیت مکانی و زاویه‌ای مفاصل است. برچسب‌گذاری شامل دو حالت «درست» یا «نادرست» برای ارزیابی کیفیت اجراست.

مجموعه‌داده KIMORE [۱۱۸]: مجموعه‌داده KIMORE منبعی عمومی برای ارزیابی حرکات توان‌بخشی با تصاویر RGB تک‌نماست. این مجموعه شامل ۳۵۳ نمونه از پنج نوع فعالیت، اجرا شده توسط ۴۴ فرد سالم و ۳۴ بیمار، با میانگین مدت ۲۹,۹ ثانیه است.

حاشیه‌نویسی با نمرات کیفی تخصصی شامل شاخص‌های امتیاز جهت‌گیری وضعیت بدن[۱] (POS) و امتیاز عملکرد بالینی[۲] (CFS) در بازه ۰ تا ۵۰، بر اساس ارزیابی بالینی متخصصان درمانی، انجام شده است.

مجموعه‌داده EHE [۱۱۹]: مجموعه‌داده EHE شامل ۸۶۹ تکرار فعالیت است که توسط ۲۵ فرد سالمند اجرا شده است. یکی از اهداف اصلی این پایگاه، شناسایی ویژگی‌های اسکلت‌بندی مؤثرتر در تشخیص ناهنجاری‌های حرکتی بوده است.

مجموعه‌داده MMASD [۱۲۰]: یک پایگاه متن‌باز چندوجهی مبتنی بر جلسات درمانی برای کودکان مبتلا به اختلال طیف اوتیسم (ASD) شامل ۱۳۱۵ نمونه از ۳۲ کودک و بیش از ۱۰۰ ساعت ویدئو است. داده‌ها ۱۱ فعالیت در ۳ موضوع اصلی را پوشش می‌دهند و شامل چهار نوع داده چندوجهی شامل جریان نوری، اسکلت دوبعدی و سه‌بعدی و امتیازهای ارزیابی ASD توسط درمانگران متخصص است. میانگین طول هر ویدئو حدود ۷ ثانیه بوده و با نرخ بین ۲۵ تا ۳۰ فریم بر ثانیه ضبط شده است.

مجموعه‌داده FineRehab [۱۲۱]: مجموعه‌داده FineRehab یک پایگاه چندوجهی برای تحلیل حرکات توان‌بخشی است که با بهره‌گیری از فناوری حسگرها و هوش مصنوعی طراحی شده است. این مجموعه شامل ۴۲۱۵ نمونه از ۱۶ نوع فعالیت انجام‌شده توسط ۵۰ شرکت‌کننده (شامل بیماران مبتلا به اختلالات اسکلتی‌عضلانی و افراد سالم) است.

۴،۱،۴. مجموعه‌داده‌های فعالیت‌های روزمره

AQA در مجموعه‌داده‌هایی که بر فعالیت‌های روزمره و ارزیابی مهارت‌های عمومی تمرکز دارند نیز مورد توجه قرار گرفته است.

مجموعه‌داده EPIC-Skill [۸۶]: مجموعه‌داده EPIC-Skill شامل ۲۱۶ ویدئو با میانگین مدت ۸۶٫۶ ثانیه، در چهار زیرمجموعه جراحی، خمیررول‌کنی، نقاشی و استفاده از چوب چاپستیک تقسیم‌بندی شده است. زیرمجموعه جراحی همان مجموعه‌داده JIGSAWS است. زیرمجموعه خمیررول‌کنی شامل ۳۳ نمونه انتخاب‌شده از مجموعه‌داده مبتنی بر آشپزخانه CMU-MMAC است. در زیرمجموعه نقاشی، چهار داوطلب دو طرح «SONIC» (کارتون سونیک جوجه‌تیغی) و «HAND» (تصویر خاکستری دست) را هر یک پنج بار ترسیم کرده‌اند. زیرمجموعه استفاده از چاپستیک نیز شامل هشت داوطلب است که هر یک پنج بار دانه‌های لوبیا را از یک جعبه جابه‌جا کرده‌اند. کیفیت فعالیت بر اساس تعداد انتقال موفق در دقیقه ارزیابی شده است. از آنجا که مجموعه JIGSAWS دارای برچسب‌گذاری دقیق است، سه زیرمجموعه‌ی دیگر از طریق مقایسه‌ی زوجی درون‌کلاسی بر اساس عملکرد رتبه‌بندی شده‌اند.

مجموعه‌داده BEST [۸۷]: با توجه به محدودیت‌های مجموعه‌داده EPIC-Skills، از جمله زاویه دید تک و پس‌زمینه‌های ساده که باعث غیرواقعی شدن صحنه‌ها برای کاربردهای عملی می‌شود، مجموعه‌داده‌ی Bristol Everyday Skill Tasks برای ارزیابی مهارت در ویدئوهای بلندمدت و در پنج فعالیت مختلف روزمره در سال ۲۰۱۹ معرفی شد. این مجموعه شامل ۵۰۰ ویدئو از پنج فعالیت مهارتی روزمره (هم‌زدن تخم‌مرغ، بافتن مو، بستن کراوات، ساخت اوریگامی درنا[۳] و کشیدن خط چشم) با میانگین مدت ۱۸۷٫۶ ثانیه است. ویدئوها از یوتیوب گردآوری شده و اجراهای واقعی با زوایای دید و پس‌زمینه‌های متنوع را دربرمی‌گیرند. برچسب‌گذاری داده‌ها شامل زمان شروع و پایان فعالیت‌ها، درجه‌بندی مهارت و رتبه‌بندی دوبه‌دو برای ۴۰ درصد از جفت‌های ممکن است.

---

[۱] Postural Orientation Score (POS)

[۲] Clinical Functional Score

[۳] an origami crane

مجموعه‌داده Infant Grasp [۸۸]: این مجموعه شامل ۹۴ ویدئو از نوزادان در حال گرفتن اشیاء، با طولی بین ۸۰ تا ۵۰۰ فریم است که از پژوهش‌های روان‌شناسی استخراج شده‌اند. پنج روان‌شناس به‌صورت جفتی ویدئوها را ارزیابی کردند و در نتیجه ۴۳۷۱ جفت داده ایجاد شد که ۳۳۱۸ جفت دارای تفاوت‌های دوبه‌دو بودند. این مجموعه‌داده دیدگاهی نوین برای کاربرد AQA در روانشناسی رشد و مداخلات زودهنگام فراهم می‌کند.

مجموعه‌داده EgoExo4D [۱۲۲]: این مجموعه‌داده شامل ۱۲۲۴ ویدئو با مجموع ۱۶ ساعت محتوای تصویری است که به‌صورت هم‌زمان از دیدگاه‌های اول‌شخص (Ego) و سوم‌شخص (Exo) ضبط شده‌اند. فعالیت‌ها شامل حوزه‌های متنوعی مانند ورزش، موسیقی، رقص و تعمیر دوچرخه بوده و در ۱۲۳ محیط طبیعی مختلف انجام گرفته‌اند. این مجموعه‌داده با داده‌های چندوجهی شامل صدا، نگاه چشم، ابرهای نقاط سه‌بعدی، موقعیت دوربین‌ها و توضیحات متنی، از جمله تفاسیر تخصصی توسط مربیان، غنی شده است که امکان تحلیل دقیق و چند وظیفه‌ای را فراهم می‌سازد.

۵،۱،۴. مجموعه‌داده‌های تناسب اندام

در حوزه‌ی AQA در فعالیت‌های مرتبط با تناسب اندام، چندین مجموعه‌داده شاخص معرفی شده‌اند که از نظر مقیاس، ساختار داده و نوع حاشیه‌نویسی، نقش مهمی در پیشرفت این حوزه ایفا کرده‌اند.

مجموعه‌ی داده UMONS-TAICHI [۱۲۳]: یک مجموعه بزرگ شامل داده‌های ضبط حرکت سه‌بعدی از حرکات هنر رزمی تای‌چی‌چوان است و ۲۲۰۰ نمونه از ۱۳ کلاس حرکتی، اجرا شده توسط ۱۲ شرکت‌کننده با سطوح مهارتی متفاوت را در بر می‌گیرد. مهارت‌ها توسط سه متخصص با مقیاس ۰ تا ۱۰ ارزیابی شده‌اند. داده‌ها با دو سیستم ضبط حرکت هم‌زمان جمع‌آوری شده‌اند: سیستم Qualisys با ۱۱ دوربین Oqus و ۶۸ نشانگر بازتابنده (با نرخ ۱۷۹ هرتز) و حسگر Kinect V2 با ردیابی ۲۵ نقطه اسکلتی (با نرخ ۳۰ هرتز). داده‌های Qualisys به‌صورت دستی اصلاح و برای تکمیل اطلاعات گم‌شده پردازش شده‌اند. حاشیه‌نویسی شامل بخش‌بندی دستی حرکات است و نسخه‌های بخش‌بندی‌شده و نشده در دسترس قرار دارند.

مجموعه‌داده Fitness-AQA [۸۰]: این مجموعه شامل ۲۱۲۸۴ ویدئو از منابع عمومی، مانند وب‌سایت‌های اشتراک‌گذاری ویدئو، است و سه حرکت متداول تمرین با وزنه شامل اسکوات با هالتر، پارویی با هالتر و پرس سرشانه را با میانگین مدت ۴٫۱ ثانیه پوشش می‌دهد. برچسب‌گذاری داده‌ها به‌صورت دودویی (درست یا نادرست) توسط دو مربی حرفه‌ای انجام شده است.

مجموعه‌داده EgoExo-Fitness [۱۲۴]: گامی فراتر در ثبت و تحلیل حرکات تناسب اندام برداشت. این مجموعه شامل ۶۱۳۱ نمونه از ۸۶ توالی حرکتی، با ۳ تا ۶ حرکت متفاوت از میان ۱۲ نوع فعالیت ورزشی رایج است. داده‌ها به‌صورت هم‌زمان با سه دوربین اول‌شخص و سه دوربین سوم‌شخص ضبط شده‌اند و ساختاری چنددیدگاهی ارائه می‌دهند. میانگین مدت هر نمونه ۱۸٫۸ ثانیه است. حاشیه‌نویسی شامل نقاط کلیدی فنی برای صحت اجرا، توضیحات متنی به زبان طبیعی درباره عملکرد، و امتیاز کیفی عددی (۱ تا ۵) است.

۶،۱،۴. مجموعه‌داده موسیقی

مجموعه‌داده PISA [۵۷]: مجموعه‌داده Piano-Skills یا PISA، نخستین مجموعه برای ارزیابی مهارت نواختن پیانو است که از ویدئوهای یوتیوب گردآوری شده و داده‌های شنیداری را به عنوان مؤلفه‌ای کلیدی در کنار داده‌های تصویری به‌کار می‌گیرد. این مجموعه

شامل ۹۹۲ نمونه منحصربه‌فرد از ۶۱ اجرای مختلف پیانو، با میانگین مدت ۱۶۰ فریم است. برچسب‌گذاری توسط پیانیست حرفه‌ای انجام شده و شامل سطوح مهارتی نوازنده، درجه دشواری قطعه، نام قطعه و کادر محدوده‌ای اطراف دست‌های نوازنده است.

۷،۱،۴. مجموعه‌داده تولید صنعتی

مجموعه‌داده Assembly101 [۶]: منبعی جامع در زمینه‌ی ارزیابی کیفیت عملکرد فعالیت‌های مونتاژ صنعتی است که با استفاده از ۸ دوربین ثابت و ۴ دوربین دید اول‌شخص به‌صورت هم‌زمان ثبت شده‌اند. این مجموعه شامل ۴۳۲۱ ویدئو از فعالیت‌های مونتاژ و جداسازی ۱۰۱ مدل ماشین اسباب‌بازی توسط ۳۶۲ داوطلب، با میانگین مدت ۴۲۶ ثانیه، است. حاشیه‌نویسی شامل بیش از ۱ میلیون قطعه فعالیت با ۱۳۸۰ کلاس حرکتی ریزدانه و ۲۰۲ کلاس کلی، همراه با امتیازهای کیفی مرتبط با سطح اجرای فعالیت است.

۸،۱،۴. مجموعه‌داده رقص

مجموعه‌داده CDRG [۹۴]: این مجموعه‌داده، از نخستین تلاش‌ها برای ساخت مجموعه‌ای در زمینه‌ی ارزیابی عملکرد رقص در بستر پلتفرم‌هایی مانند تیک‌تاک شناخته می‌شود. این مجموعه شامل ۲۴۰ ویدئو از ۸ چالش مختلف رقص است که توسط ۲۰ داوطلب با میانگین مدت ۱۴٫۷ ثانیه اجرا شده‌اند. ویدئوها توسط شرکت‌کنندگان با دستگاه‌های شخصی ضبط شده‌اند که منجر به تنوع در کیفیت تصویر و ایجاد چالش‌های واقعی برای تحلیل عملکرد در محیط‌های غیرکنترل‌شده می‌شود. حاشیه‌نویسی مبتنی بر قضاوت نسبی است؛ به‌طوری که ۱۰۰ ارزیاب انسانی برای هر جفت اجرا، برنده را بر اساس کیفیت کلی انتخاب کرده‌اند.

۹،۱،۴. مجموعه‌داده آموزش

مجموعه‌داده TAQR [۱۰۱]: به عنوان نخستین منبع اختصاصی برای رتبه‌بندی کیفیت فعالیت‌های معلمان طراحی شده است. این مجموعه‌داده شامل ۱۲۰۰ ویدئوی آموزشی از چهار نوع فعالیت رایج در کلاس‌های درس سنتی دبیرستان در نه رشته‌ی درسی مختلف، گردآوری‌شده از شش مسابقه ملی سخنرانی معلمان، با پس‌زمینه یکسان و تغییرات جزئی در زاویه دید است. ویدئوها با وضوح ۱۹۲۰×۱۰۸۰ پیکسل و نرخ ۲۵ فریم بر ثانیه ثانیه ضبط شده‌اند. فرآیند حاشیه‌نویسی توسط دو گروه شامل چهار متخصص آموزشی (دو دبیر و دو استاد دانشگاه) انجام شد. جفت‌های ویدئویی به‌صورت هم‌زمان ارزیابی و بر اساس کیفیت رتبه‌بندی شدند، و تنها در صورت اجماع کامل، رتبه‌بندی به‌عنوان واقعیت پایه ثبت شد. برای هر فعالیت، ۵۰ درصد جفت‌های ممکن برچسب‌گذاری شدند. پس از بازبینی نهایی، جفت‌های دارای امتیاز یکسان حذف و جفت‌های با رتبه‌های نزدیک حفظ شدند. داده‌ها با نسبت ۷۵ درصد (۲۲۵ نمونه) برای آموزش و ۲۵ درصد (۷۵ نمونه) برای تست تقسیم شدند.

۱۰،۱،۴. مجموعه‌داده‌های تولیدشده توسط هوش مصنوعی[۱] (AIGV)

مجموعه‌داده GAIA [۱۲۵]: نخستین مجموعه‌داده AQA برای ویدئوهای تولیدشده توسط مدل‌های متن‌به‌ویدئو (T2V) برای ارزیابی کیفیت تولید ویدئو ساخته شده است. این مجموعه شامل ۹۱۸۰ ویدئویی تولیدشده توسط ۱۸ آزمایشگاه تحقیقاتی و پلتفرم تجاری مختلف است و دامنه‌ای گسترده از ۵۱۰ نوع فعالیت در حوزه‌های حرکات کل بدن، دست و حالات چهره را پوشش می‌دهد. میانگین مدت‌زمان هر ویدئو ۲٫۸ ثانیه است که با هدف ارزیابی حرکات کوتاه و مستقل طراحی شده است. حاشیه‌نویسی‌ها از سه منظر کلیدی

[۱] AI Generated Videos (AIGV)

کیفیت ظاهر سوژه، کامل‌بودن اجرای فعالیت و تعامل میان فعالیت و صحنه انجام شده است. هر یک از این ابعاد توسط ارزیابان انسانی متخصص با مقیاس عددی ۰ تا ۱۰۰ امتیازدهی شده‌اند تا ارزیابی چندبعدی و دقیق‌تری از کیفیت تولید ویدئو ارائه گردد.

۲,۴. معیارهای ارزیابی عملکرد

ارزیابی دقت و کارایی مدل‌های AQA نقش اساسی در سنجش اثربخشی آن‌ها در کاربردهای مختلف دارد. از آنجا که ماهیت وظایف در این حوزه متنوع است، از امتیازدهی رگرسیون پیوسته تا رتبه‌بندی مقایسه‌ای و درجه‌بندی طبقه‌بندی‌شده، معیار واحدی برای ارزیابی وجود ندارد. بنابراین، انتخاب معیار باید متناسب با نوع مسئله و هدف ارزیابی انجام گیرد. در وظایف رگرسیونی که خروجی مدل به‌صورت عددی پیوسته است، عملکرد با استفاده از شاخص‌های همبستگی مانند ضریب همبستگی رتبه‌ای اسپیرمن[۱] (SRC)، ضریب همبستگی پیرسون[۲] (PCC)، ضریب همبستگی کندال[۳] (KC)، ضریب همبستگی درون‌گروهی[۴] (ICC) و ضریب همبستگی رتبه‌ای میانگین[۵] (MRC) سنجیده می‌شود. این معیارها میزان تطابق میان مقادیر پیش‌بینی‌شده و مقادیر واقعی را نشان می‌دهند. در کنار آن، معیارهای خطا مانند میانگین مربعات خطا[۶] (MSE)، ریشه میانگین مربعات خطا[۷] (RMSE)، میانگین فاصله اقلیدسی[۸] (MED)، میانگین خطای مطلق[۹] (MAE)، فاصله اقلیدسی[۱۰] (ED) و فاصله نسبی[۱۱] (R-ℓ 2) برای سنجش میزان انحراف پیش‌بینی از مقدار واقعی استفاده می‌شوند. در برخی پژوهش‌ها، کیفیت فعالیت‌ها را به سطوح گسسته و طبقه‌بندی‌شده مختلف برچسب‌گذاری می‌کنند که می‌تواند به‌عنوان یک مسئله درجه‌بندی فرموله شود. در این وظایف، معیارهایی مانند دقت طبقه‌بندی[۱۲] (Acc)، امتیاز F1 و دقت پیش‌بینی[۱۳]، میانگین دقت متوسط[۱۴] (mAP) و میانگین دقت متوسط مدل چندوجهی تصویر–متن[۱۵] (mmIT mAP) کاربرد دارند.

ضریب همبستگی رتبه‌ای اسپیرمن: ضریب همبستگی رتبه اسپیرمن (ρ) یکی از پرکاربردترین معیارهای حوزه AQA است که نخستین‌بار توسط پیرسیاوش و همکاران [۸] معرفی شد. این ضریب میزان انطباق رتبه‌بندی پیش‌بینی‌شده توسط مدل را با رتبه‌بندی واقعی نشان می‌دهد. مقدار ρ بین ۱− و ۱ متغیر است و هرچه مقدار آن به ۱ نزدیک‌تر باشد، انطباق بین ترتیب پیش‌بینی‌شده مدل و رتبه‌بندی داوران خبره بیشتر است. رابطه‌ی ریاضی آن به‌صورت زیر تعریف می‌شود:

$$\rho = \frac{cov(p,q)}{\sigma_p \sigma_q} = \frac{\sum_i (p_i - \bar{p})(q_i - \bar{q})}{\sqrt{\sum_i (p_i - \bar{p})^2 \sum_i (p_i - \bar{q})^2}} \qquad (۲)$$

که در آن p و q به‌ترتیب دنباله‌های رتبه‌بندی امتیازهای پیش‌بینی شده و واقعی است. cov، کوواریانس این دنباله‌ها و σ_p و σ_q انحرافات معیار استاندارد p و q هستند. $\bar{p}$ و $\bar{q}$ میانگین رتبه‌ها می‌باشند.

---

[۱] Spearman's Rank Correlation (SRC)
[۲] Pearson Correlation Coefficient (PCC)
[۳] Kendall Correlation (KC)
[۴] Intra-group Correlation Coefficient (ICC)
[۵] Mean Rank Correlation (MRC)
[۶] Mean Squared Error (MSE)
[۷] Root Mean Squared Error (RMSE)
[۸] Mean Euclidean Distance (MED)
[۹] Mean Absolute Error (MAE)
[۱۰] Euclidean Distance (ED)
[۱۱] Relative ℓ 2-distance (R-ℓ 2)
[۱۲] Accuracy (Acc)
[۱۳] Precision
[۱۴] mean Average Precision (mAP)
[۱۵] mean Average Precision of the multi-modal image-text (mAP mmIT)

میانگین مربعات خطا و میانگین فاصله اقلیدسی: در ارزیابی‌های عددی پیوسته، معیارهای MSE و MED برای سنجش اختلاف مستقیم میان امتیازهای واقعی و پیش‌بینی‌شده به‌کار می‌روند (معادله ۳ و ۴). MSE میانگین مربعات اختلاف میان امتیازهای پیش‌بینی‌شده و واقعی را محاسبه می‌کند و نسبت به خطاهای بزرگ حساس‌تر است، درحالی‌که MED میانگین اختلاف مطلق بین امتیازهای پیش‌بینی‌شده و واقعی را محاسبه کرده و دید دقیق‌تری از میزان انحراف کلی ارائه می‌دهد.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(S_i - \hat{S}_i\right)^2 \quad (۳)$$

$$MED = \frac{1}{N}\sum_{i=1}^{N}\left|S_i - \hat{S}_i\right| \quad (۴)$$

در این روابط، S_i و $\hat{S}_i$ به ترتیب نشان‌دهنده مقدار امتیاز واقعی و پیش‌بینی‌شده برای نمونه i–ام، و N تعداد نمونه‌ها است. هر چه هر یک از این مقادیر معیار کوچکتر باشد، فاصله بین امتیاز پیش‌بینی‌شده و واقعی کمتر است؛ بنابراین عملکرد بهتری خواهد داشت.

فاصله نسبی R-ℓ 2: برای رفع محدودیت‌های ضریب همبستگی رتبه‌ای اسپیرمن، یو و همکاران [۹۰] معیار فاصله نسبی (R-l_2) را پیشنهاد کردند. این معیار با نرمال‌سازی بازه امتیازها، امکان مقایسه‌ی مدل‌ها در فعالیت‌های مختلف را فراهم می‌کند. در بسیاری از مجموعه‌داده‌های AQA، دامنه امتیازها در دسته‌های مختلف عملکردی متفاوت است؛ برای مثال، امتیازها در یک فعالیت ورزشی ممکن است از ۰ تا ۱۰ باشد، در حالی که در فعالیتی دیگر از ۰ تا ۱۰۰ تعریف شده باشد که مقایسه مستقیم خطاها را دشوار می‌کند. برخلاف فاصله سنتی ℓ 2، معیار R-ℓ 2 بازه‌های امتیازی بین دسته‌های مختلف فعالیت را در نظر می‌گیرد و امکان آموزش مدل بین دسته‌های مختلف را تسهیل می‌کند. دامنه‌ی این معیار بین ۰ تا ۱ است، به‌طوری‌که مقدار نزدیک‌تر به صفر نشان‌دهنده‌ی دقت بالاتر مدل است. با توجه به بالاترین و کمترین امتیازها برای یک فعالیت، S_{min} و S_{max}، این معیار بطور رسمی به صورت معادله ۵ تعریف می‌شود.

$$R-\ell_2 = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{|S_i-\hat{S}_i|}{S_{max}-S_{min}}\right)^2 \times 100 \quad (۵)$$

دقت طبقه‌بندی: این معیار در بسیاری از پژوهش‌های حوزه‌ی AQA، به‌ویژه در وظایف درجه‌بندی و مقایسه زوجی، که در آن فعالیت‌ها به سطوح کیفیت گسسته‌ای دسته‌بندی می‌شوند و مقایسه‌ی زوجی رتبه‌ها، به‌عنوان شاخص اصلی ارزیابی عملکرد مدل مورد استفاده قرار گرفته است. معیار دقت طبقه‌بندی، توانایی مدل در تفکیک سطوح مهارتی را نشان می‌دهد و به‌صورت زیر محاسبه می‌شود:

$$Accuracy = \frac{\sum_{i=1}^{N} 1(\hat{y}_i = y_i)}{N} \times 100\% \quad (۶)$$

در این رابطه، تابع شاخص 1 در صورت برابری مقدار پیش‌بینی و مقدار واقعی برابر ۱ و در غیر این صورت صفر است. مقادیر بالاتر دقت بیانگر عملکرد بهتر مدل در تفکیک صحیح نمونه‌ها است.

برای ارزیابی جامع‌تر، معمولاً ترکیبی از معیارهای مبتنی بر همبستگی و خطا (مانند SRC و rMSE) و معیارهای دقت (مانند Acc) استفاده می‌شود تا توانایی مدل در پیش‌بینی امتیازها و رتبه‌ها به‌طور کامل سنجیده شود. در برخی کاربردهای خاص، معیارهای اختصاصی نیز مورد استفاده قرار می‌گیرند؛ به عنوان نمونه، در روش‌هایی که شامل بخش‌بندی معنایی زمانی فعالیت هستند، از معیار تقاطع بر اتحاد[۱] (IoU) برای سنجش دقت نواحی زمانی زیرفعالیت‌ها استفاده می‌شود [۹۹] و [۹۷] و [۱۰۰]. از آنجا که ضریب همبستگی رتبه‌ای اسپیرمن متداول‌ترین معیار ارزیابی در این حوزه است، عملکرد چندین روش پیشرفته بر روی هفت مجموعه‌داده‌ی رایج در جدول (۳)

[۱] Intersection over Union (IoU)

ارائه شده است. در این جدول، برای مجموعه‌داده‌های AQA-7، Fis-V، RG و JIGSAW، ارزش z فیشر[۱] به‌منظور محاسبه‌ی میانگین SRC میان فعالیت‌ها مورد استفاده قرار گرفته است [۴].

جدول (۳): نتایج روش‌های معیار ارزیابی کیفیت فعالیت بر اساس ضریب همبستگی رتبه‌ای اسپیرمن. بهترین نتایج با قلم پررنگ و دومین بهترین با قلم زیرخط‌دار مشخص شده‌اند. این نتایج بر اساس کدهای رسمی منتشرشده هر روش گزارش شده است.

مرجع	انتشارات	AQA-7میانگین	MTL-AQA	Fis-Vمیانگین	RGمیانگین	FineDiving	LOGO	JIGSAWمیانگین
[۴۹]	CVPR 2020	۰,۸۲۹۱	۰,۹۳۵۰	۰,۰۰۰۰	۰,۴۱۸۲	۰,۸۸۹۱	**۰,۷۰۴۴**	۰,۷۰۰۰
[۹۰]	ICCV 2021	۰,۸۴۱۰	۰,۹۵۱۲	۰,۷۰۶۸	۰,۷۰۳۸	**۰,۹۴۰۶**	۰,۵۹۶۸	۰,۸۵۰۰
[۳۰]	CVPR 2022	۰,۸۱۶۴	۰,۹۳۹۵	۰,۷۵۵۰	**۰,۷۴۸۶**	۰,۹۳۵۱	۰,۶۶۰۸	۰,۸۹۰۰
[۷]	TCSVT 2023	۰,۸۵۰۱	۰,۹۵۳۶	۰,۷۲۶۵	۰,۷۳۲۹	۰,۹۳۸۱	۰,۶۷۰۹	۰,۹۰۰۰
[۵۳]	NCAA 2024	۰,۷۹۰۷	۰,۹۴۹۷	۰,۷۴۲۲	۰,۷۳۹۰	۰,۹۳۵۶	۰,۶۷۰۱	۰,۷۶۰۰
[۳۷]	INFS 2024	**۰,۸۷۲۶**	**۰,۹۶۳۸**	**۰,۷۶۰۶**	۰,۶۵۸۱	۰,۹۳۸۲	۰,۶۰۷۴	**۰,۹۱۰۰**

۵. چالش‌ها و محدودیت‌ها

با وجود پیشرفت‌های چشمگیر در حوزه AQA، این زمینه همچنان با چالش‌ها و محدودیت‌های متعددی روبه‌رو است. این موانع هم ریشه در ماهیت ذاتی ارزیابی کیفیت فعالیت انسانی دارند و هم از چالش‌های کلان‌تر موجود در حوزه بینایی کامپیوتر و یادگیری ماشین تأثیر می‌پذیرند و پیشرفت بیشتر در این حوزه را محدود می‌کنند. در این بخش، مهم‌ترین چالش‌ها و محدودیت‌ها بررسی می‌شوند.

۱,۵. چالش‌های مرتبط با مجموعه‌داده‌ها

کیفیت و گستره‌ی مجموعه‌داده‌ها یکی از مهم‌ترین گلوگاه‌های توسعه سیستم‌های AQA به شمار می‌روند. در مقایسه با مجموعه‌داده‌های وسیع حوزه‌ی بازشناسی فعالیت، نظیر Kinetics-600 [۱۱]، داده‌های مورد استفاده در AQA از نظر مقیاس، تنوع و دقت برچسب‌گذاری محدودترند. هرچند در سال‌های اخیر پیشرفت‌هایی در بهبود کیفیت و تنوع این مجموعه‌داده‌ها حاصل شده است، اما محدودیت‌هایی مانند حجم اندک داده‌ها، کمبود فعالیت‌های چندوجهی و خطاهای برچسب‌گذاری همچنان مانع تعمیم‌پذیری مطلوب مدل‌ها می‌شوند.

کمبود داده‌ها، هزینه‌ی بالای برچسب‌گذاری و تنوع محدود فعالیت‌ها: یکی از چالش‌های بنیادین در AQA، محدودیت مقیاس و کمبود غنای معنایی مجموعه‌داده‌هاست. برخلاف بازشناسی فعالیت که به برچسب‌گذاری کلی متکی است، AQA نیازمند داده‌های ریزدانه و و حاصل از جمع‌آوری در حوزه‌های خاص است تا تفاوت‌های ظریف فعالیت‌ها را ثبت کند. این امر مستلزم حضور داوران متخصص و فرآیندهای پرهزینه و زمان‌بر برچسب‌گذاری است. در نتیجه، مجموعه‌داده‌های موجود اغلب کوچک، محدود به دامنه‌های خاص و فاقد تنوع کافی‌اند؛ وضعیتی که توسعه مدل‌های پیشرفته و تعمیم‌پذیر را با مانع جدی روبه‌رو می‌کند. هرچند تقسیم فعالیت‌های پیچیده به زیرفعالیت‌ها برای افزایش تنوع پیشنهاد شده است، اما تنوع واقعی فعالیت‌ها را بهبود چندانی نمی‌بخشد [۹۹].

سوگیری انسانی و داده‌های نامتوازن: وابستگی به ارزیابی انسانی برای تعیین برچسب‌های درست، زمینه‌ساز بروز سوگیری‌های ذهنی و خطاهای داوری می‌شود. تفاوت در قضاوت داوران، خستگی، یا تجربه‌ی متفاوت می‌تواند موجب نوسان در امتیازدهی و انتقال این

[۱] Fisher's z-value

خطاها به مدل‌های یادگیری شود. در بسیاری از مجموعه‌داده‌های اولیه، تنها از امتیاز نهایی داور به‌عنوان برچسب استفاده می‌شد؛ اما چنین امتیازهایی قادر به انتقال جزئیات عملکرد نیستند. بهره‌گیری از ارزیابی چند داور، همراه با ثبت توصیف‌های تفسیری، می‌تواند به بهبود دقت و انصاف در داده‌ها کمک کند. همچنین، نامتوازن بودن داده‌ها، به‌ویژه در فعالیت‌هایی با توزیع نابرابر سطوح مهارت، باعث گرایش مدل‌ها به کلاس‌های پرتکرار و افت عملکرد کلی می‌شود.

۲,۵. چالش‌های فنی و مرتبط با مدل

پیچیدگی ذاتی فعالیت‌های انسانی و محدودیت‌های تکنیکی مدل‌ها، موانع فنی مهمی در مسیر توسعه‌ی سیستم‌های AQA ایجاد کرده‌اند.

پیچیدگی زمانی و پیوستگی فعالیت‌ها: فعالیت‌های انسانی به‌ویژه در زمینه‌های ورزشی یا پزشکی، دارای ساختار زمانی پیچیده و تغییرات ظریف بصری هستند. این فعالیت‌ها نه‌تنها متشکل از حالت‌های ایستا نیستند، بلکه دنباله‌هایی پویا و به‌هم‌پیوسته از زیرفعالیت‌های مرتبط‌اند که ترتیب، ریتم و انتقال میان آن‌ها در کیفیت نهایی نقش اساسی دارند. بسیاری از مدل‌های فعلی با تقسیم ویدئو به قطعه‌های ثابت، پیوستگی طبیعی فعالیت‌ها را از بین می‌برند و در پیش‌بینی توالی‌های بلند عملکرد ضعیفی دارند.

انسداد دید و تنوع محیطی: در شرایط واقعی، عواملی چون حرکات پیچیده، انسداد اعضای بدن، پس‌زمینه‌های پویا، تغییر نور و حرکت دوربین، باعث کاهش دقت در تخمین وضعیت بدن و استخراج ویژگی‌های پایدار می‌شوند. این خطاها در نهایت به ماژول ارزیابی منتقل شده و موجب افت عملکرد مدل‌های AQA می‌گردند.

وابستگی به دامنه و ضعف در تعمیم‌پذیری: بیشتر مدل‌های فعلی برای حوزه‌های خاصی مانند اسکیت نمایشی یا ژیمناستیک طراحی و آموزش داده می‌شوند و در کاربردهای دیگر (مانند جراحی یا توان‌بخشی) عملکرد ضعیفی دارند. این وابستگی به دامنه، نیاز به تنظیم مجدد یا آموزش اختصاصی مدل برای هر کاربرد جدید را ایجاد می‌کند و هزینه‌ی توسعه را افزایش می‌دهد.

محدودیت در کارایی محاسباتی و تأخیر زمانی: مدل‌های پیچیده نیازمند منابع محاسباتی گسترده برای آموزش و استنتاج هستند. در کاربردهای بلادرنگ، مانند مربیگری ورزشی زنده یا هدایت عمل جراحی، تأخیر پردازش می‌تواند کارایی سیستم را کاهش دهد. دستیابی به توازن میان دقت، سرعت و پیچیدگی مدل از چالش‌های کلیدی برای پیاده‌سازی سیستم‌های AQA در محیط‌های عملی است.

عدم تفسیرپذیری مدل‌ها: اکثر مدل‌های یادگیری عمیق در AQA ماهیت «جعبه‌سیاه» دارند و توانایی توضیح علت تصمیم‌گیری خود را ندارند. در حوزه‌هایی مانند ورزش یا پزشکی، که شفافیت و اعتماد اهمیت بالایی دارد، این ویژگی می‌تواند مانعی جدی در پذیرش سیستم باشد. نبود قابلیت تفسیر، امکان ارائه‌ی بازخورد روشن به کاربران و تحلیل خطاها را محدود می‌کند.

پایداری پایین در شرایط چندوجهی: روش‌های AQA اغلب از داده‌های چندوجهی برای بهبود دقت ارزیابی بهره می‌برند. پژوهشگران اغلب بر چگونگی ترکیب مؤثر داده‌ها از وجوه مختلف تمرکز می‌کنند، اما عملکرد در شرایط فقدان برخی وجوه هنوز چالش‌برانگیز است و نیازمند توسعه روش‌هایی برای افزایش انعطاف‌پذیری مدل‌ها است.

پیچیدگی ذاتی فعالیت‌ها: فعالیت‌های انسانی حتی در یک دسته واحد، مانند «دویدن»، می‌توانند از نظر سبک اجرا، سرعت، و حالت بدنی تفاوت زیادی داشته باشند. این تنوع درون‌دسته‌ای و بین‌دسته‌ای، طراحی مدل‌هایی را ضروری می‌کند که بتوانند تفاوت‌های ظریف را در سناریوهای مختلف به‌دقت درک و تحلیل کنند.

۶. جهت‌گیری‌های آینده و پرسش‌های پژوهشی باز

با وجود پیشرفت‌های قابل‌توجه در سال‌های اخیر، حوزه AQA همچنان با چالش‌های بنیادی مواجه است. شناسایی و تحلیل این چالش‌ها، زمینه‌ساز ترسیم مسیرهای آینده و گسترش کاربردهای واقعی در این حوزه است.

۱،۶. پیشرفت‌های سطح مجموعه‌داده

بخش قابل‌توجهی از محدودیت‌های فعلی در AQA ناشی از کمیت و کیفیت پایین مجموعه‌داده‌هاست. بنابراین، ارتقای اساسی در طراحی، برچسب‌گذاری و تنوع داده‌ها، نقش کلیدی در رشد این حوزه دارد.

توسعه مجموعه داده‌های بزرگ‌تر و از نظر معنایی غنی‌تر: یکی از نیازهای اساسی، ایجاد مجموعه‌داده‌هایی در مقیاس بزرگ است که بتوانند تنوع فعالیت‌های انسانی، تفاوت‌های کیفی در اجرا، شرایط محیطی متنوع و کاهش سوگیری داده‌ها را پوشش دهند. مجموعه‌داده‌های آینده باید با الهام از حوزه‌هایی مانند تشخیص زیرفعالیت‌ها و بازشناسی فعالیت، بر ثبت تفاوت‌های ظریف در اجرا و زوایای دید واقعی تمرکز داشته باشند. علاوه‌بر امتیازدهی عددی، استفاده از برچسب‌های توصیفی، سلسله‌مراتبی و چندوجهی می‌تواند به انعکاس دقیق‌تر کیفیت واقعی اجرا کمک کند.

معیارهای استاندارد و یکپارچه: با وجود نقش مؤثر مجموعه‌داده‌های معیار مانند SkatingVerse [۱۱۱] در استانداردسازی حوزه‌های خاص، نیاز به مجموعه‌داده‌های جامع و یکپارچه که دامنه‌های مختلف را در قالبی مشترک گردآوری کنند همچنان برجسته است. چنین منابعی امکان مقایسه دقیق مدل‌های AQA در شرایط گوناگون را فراهم آورده و توسعه سیستم‌هایی با تعمیم‌پذیری بالا را تسهیل می‌کنند.

مجموعه داده‌های AIGC چندفعالیتی در مقیاس بزرگ و باکیفیت: پیشرفت‌های اخیر در مدل‌های تبدیل متن به ویدئو (T2V)، زمینه‌ساز تولید ویدئوهای مصنوعی برای آموزش سیستم‌های AQA شده‌اند. مجموعه‌داده‌های مصنوعی چندفعالیتی می‌توانند هزینه و زمان برچسب‌گذاری سنتی را کاهش داده و تنوع و تعمیم‌پذیری مدل‌ها را ارتقا دهند. برای مثال، مجموعه‌داده GAIA [۱۲۵] نخستین مجموعه‌داده غیرواقعی در این زمینه است، هرچند هنوز با چالش‌هایی نظیر بازنمایی ناقص بدن انسان و ناهماهنگی در توالی فعالیت‌ها روبه‌روست. استفاده از توصیف‌های دقیق ویدئویی در قالب «پرامپت‌های ویدئویی» می‌تواند به‌عنوان برچسب‌گذاری سطح‌بالا به کار رفته و دشواری و زمان‌بر بودن فرآیند حاشیه‌نویسی را به‌طور چشم‌گیری کاهش دهد. بنابراین، توسعه مجموعه‌داده‌های AIGC چندفعالیتی در مقیاس بزرگ و باکیفیت، با پوشش متنوع فعالیت‌ها، می‌تواند بهره‌برداری از معماری‌های پیشرفته یادگیری عمیق را در AQA تقویت کرده و تعمیم‌پذیری مدل‌ها را بهبود بخشد.

۲،۶. پیشرفت‌های سطح مدل

نوآوری در طراحی مدل‌های یادگیری ماشین، کلید غلبه بر چالش‌های فنی و ارتقای قابلیت‌های سیستم‌های AQA است.

طراحی مدل‌های دقیق و کارآمد: برای کاربردهای بلادرنگ مانند مربیگری ورزشی یا توان‌بخشی، توسعه‌ی مدل‌هایی با دقت بالا و کارایی محاسباتی مطلوب ضروری است. این هدف نیازمند طراحی شبکه‌های عصبی سبک‌وزن، بهینه‌سازی سرعت استنتاج و استفاده‌ی هوشمند از سخت‌افزارهایی مانند GPU است. همچنین، پیاده‌سازی مدل‌های سبک AQA بر روی دستگاه‌های نزدیک به منبع داده، مانند

گوشی‌های هوشمند یا حسگرهای پوشیدنی، می‌تواند تأخیر پردازش را کاهش داده، مصرف پهنای باند را کم کرده و با پردازش محلی داده‌ها، حریم خصوصی کاربران را حفظ کند.

بهبود مدل‌سازی اطلاعات زمانی: درک صحیح کیفیت اجرای فعالیت نیازمند تحلیل دقیق ساختارهای زمانی و توالی زیرفعالیت‌ها است. تحقیقات آینده باید بر توسعه مدل‌هایی تمرکز کنند که بتوانند توالی‌های زمانی پیچیده را تحلیل کرده، ساختار سلسله‌مراتبی فعالیت‌ها را شناسایی و پیوستگی زمانی را حفظ کنند.

کاهش وابستگی به برچسب‌گذاری: با توجه به هزینه و زمان‌بر بودن فرآیند برچسب‌گذاری تخصصی، توسعه‌ی روش‌های نیمه‌نظارتی و خودنظارتی برای آموزش مدل‌ها ضروری است. استفاده از داده‌های بدون برچسب همراه با یادگیری بازنمایی خودنظارتی و برچسب‌گذاری شبه‌خودکار می‌تواند بدون نیاز به دخالت مستقیم انسان، ویژگی‌های معنادار را استخراج کند. در این چارچوب، آموزش اولیه روی حجم بزرگی از داده‌های بدون برچسب و تنظیم دقیق روی مجموعه‌های محدودتر برچسب‌دار، موجب بهبود دقت و تعمیم‌پذیری مدل‌های AQA می‌شود.

هوش مصنوعی قابل توضیح در AQA: به دلیل ماهیت ذهنی ارزیابی انسانی، توسعه‌ی مدل‌های قابل‌توضیح که بتوانند علاوه بر پیش‌بینی امتیاز، دلایل تصمیم‌گیری خود را بیان کنند، اهمیت ویژه‌ای دارد. در ادامه‌ی پیشرفت‌های مبتنی بر روش‌های عصبی–نمادین، بهره‌گیری از نمادهای قابل‌فهم برای انسان، تفسیرپذیری را بهبود بخشیده و نتایج امیدوارکننده‌ای، به‌ویژه در رویکردهای سلسله‌مراتبی [۷۳]، ارائه داده است. جهت‌گیری‌های آتی باید بر استفاده از محاسبات عصبی–نمادین برای شفاف‌سازی فرآیند تصمیم‌گیری و ارائه گزارش‌های عینی و مبتنی بر شواهد بصری تمرکز کنند تا اعتماد و کارایی سیستم‌های AQA را ارتقا دهند.

پایداری در شرایط داده ناقص: ترکیب جریان‌های داده‌های متنوع، نظیر ویدئو، واحدهای اندازه‌گیری اینرسی (IMU)، حسگرهای عمق و صوت، می‌تواند درک دقیق‌تری از کیفیت فعالیت فراهم کند. این رویکرد در سناریوهایی با انسداد دید، حرکات مبهم، شرایط نوری نامناسب یا مواردی که نشانه‌های غیربصری (مانند نیرو، تعادل یا صدای ضربه) تعیین‌کننده کیفیت هستند، بسیار مؤثر است. با این حال، بسیاری از روش‌های چندوجهی کنونی به پایداری در شرایط فقدان برخی وجوه داده‌ای توجه کافی ندارند. پژوهش‌هایی مانند [۷۲] و [۱۲۱] با استفاده از یادگیری چند وظیفه‌ای، تأثیر فقدان وجوه را کاهش داده‌اند، اما همچنان دستیابی به یکپارچگی کامل داده‌ها چالشی باز است. جهت‌گیری‌های آینده باید بر یکپارچگی مؤثر داده‌های چندوجهی و تقویت پیش‌بینی در شرایط ناقص تمرکز کنند. بهره‌گیری از مدالیته‌های موجود برای جایگزینی یا نگاشت مدالیته‌های مفقود می‌تواند عملکرد سیستم را در سناریوهای واقعی بهبود دهد.

تحلیل تعامل انسان–شیء: در بسیاری از وظایف AQA، از جمله جراحی، مونتاژ صنعتی و فعالیت‌های روزمره نظیر آشپزی، تعامل انسان با اشیاء نقش تعیین‌کننده‌ای در کیفیت عملکرد دارد. در تحقیقات آینده، مدل‌ها باید به‌گونه‌ای طراحی شوند که رابطه بین فعالیت‌های انسان و حالت‌های شیء و چگونگی تأثیر تعامل آن‌ها بر کیفیت کلی فعالیت را درک کنند.

۳,۶. دامنه‌های کاربردی گسترده‌تر

افق‌های آینده‌ی AQA فراتر از حوزه‌های ورزش و پزشکی است. از جمله می‌توان به کاربرد در آموزش (ارزیابی مهارت‌های عملی دانش‌آموزان)، رباتیک (همکاری انسان و ربات یا آموزش ربات از طریق مشاهده‌ی فعالیت انسان)، و ارزیابی کیفیت اجرای هنری اشاره

کرد. کاربرد AQA در فعالیت‌های روزمره مانند رقص، فوتبال و بسکتبال همچنان محدود مانده و نیازمند بررسی‌های عمیق‌تر است. کاوش‌های اضافی در این زمینه‌ها می‌تواند کارایی AQA را در موقعیت‌های واقعی گوناگون تقویت کند. علاوه بر این، ترکیب AQA با فناوری‌های واقعیت مجازی[۱] (VR) و واقعیت افزوده[۲] (AR) می‌تواند محیط‌هایی تعاملی‌تر و ارزیابی‌هایی دقیق‌تر ایجاد کند.

۷. جمع‌بندی و نتیجه‌گیری

ارزیابی کیفیت فعالیت انسانی مبتنی بر ویدئو به‌عنوان یکی از حوزه‌های کلیدی و نوظهور در بینایی کامپیوتر و یادگیری ماشین، جایگاهی پایدار و رو به گسترش یافته است. هدف اصلی این حوزه، ارائه ارزیابی‌های کمی، عینی و دقیق از کیفیت اجرای فعالیت‌های انسانی است؛ امری که زمینه‌ساز کاربردهای گسترده در تحلیل عملکرد ورزشی، توان‌بخشی پزشکی، آموزش مهارت‌های تخصصی و نظارت صنعتی شده است. چنین رویکردهایی امکان ارائه راه‌حل‌های خودکار، مقیاس‌پذیر و قابل‌اعتماد را برای تحلیل و بهبود عملکرد انسانی مهیا می‌کنند. این مرور نظام‌مند با بررسی و تحلیل محتوای صد مقاله علمی منتشرشده، تصویری جامع از وضعیت کنونی پژوهش‌ها در زمینه AQA ارائه داده است. یافته‌ها نشان می‌دهد که با وجود پیشرفت‌های چشمگیر در طراحی مدل‌ها، روش‌های یادگیری و گسترش مجموعه‌داده‌های متنوع، هنوز چالش‌های بنیادینی در این حوزه پابرجاست. کمبود داده‌های متوازن و باکیفیت، محدودیت در تعمیم‌پذیری مدل‌ها و نیاز به تبیین‌پذیری تصمیمات سیستم‌های هوشمند از جمله مهم‌ترین این چالش‌ها به شمار می‌روند. بنابراین، پژوهش‌های آتی باید بر توسعه مدل‌های چندوجهی، افزایش تفسیرپذیری خروجی‌ها، و طراحی روش‌های ارزیابی قابل اعتماد در شرایط واقعی متمرکز شوند. دستیابی به این اهداف می‌تواند مسیر را برای ساخت سیستم‌های AQA پایدار، دقیق و کارآمد هموار کرده و زمینه‌ساز به‌کارگیری گسترده آن‌ها در محیط‌های عملی و روزمره شود.

مراجع

[1] M. Norouzi, H. Hassanpour, and A. Ghanbari, "Investigating object recognition models based on deep learning," *Soft Computing Journal (SCJ),* 2024, doi: 10.22052/scj.2024.252945.1149.

[2] F. Hoseini, E. Tabibzade Lamar, and S. M. Mirkazemi Niarag, "Face recognition with incomplete data by deep convolutional neural network," *Soft Computing Journal (SCJ),* vol. 13, pp. 158–171, 2024, Art no. 1, doi: 10.22052/scj.2024.252789.1143.

[3] M. Aboonajmi and Z. Mostafaei, "Overview of fruit and vegetables quality assessment surveys using soft computing," *Soft Computing Journal (SCJ),* 2024, doi: 10.22052/scj.2024.248418.1103.

[4] P. Parmar and B. T. Morris, "Action quality assessment across multiple actions," in *2019 IEEE winter conference on applications of computer vision (WACV)*, Waikoloa Village, HI, USA, 2019: IEEE, pp. 1468–1476, doi: 10.1109/WACV.2019.00161.

[5] P. Parmar and B. T. Morris, "Learning to score olympic events," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Honolulu, HI, USA, Jul. 2017, pp. 76-84, doi: 10.1109/CVPRW.2017.16.

[6] F. Sener *et al.*, "Assembly101: A large-scale multi-view video dataset for understanding procedural activities," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA, 2022, pp. 21096–21106, doi: 10.1109/CVPR52688.2022.02042.

[7] K. Zhou, Y. Ma, H. P. Shum, and X. Liang, "Hierarchical graph convolutional networks for action quality assessment," *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 33, no. 12, pp. 7749–7763, 2023, doi: 10.1109/TCSVT.2023.3281413.

[8] H. Pirsiavash, C. Vondrick, and A. Torralba, "Assessing the quality of actions," in *13th European Conference on Computer Vision (ECCV)*, Zurich, Switzerland, 2014, vol. 8694 LNCS: Springer, pp. 556–571, doi: 10.1007/978-3-319-10599-4_36.

[۱] Virtual Reality (VR)

[۲] Augmented Reality (AR)

[9] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," *Advances in Neural Information Processing Systems (NIPS)* ,vol. 27, 2014, doi: 10.48550/arXiv.1406.2199.

[10] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV),* Santiago, Chile, Dec. 2015, pp. 4489–4497, doi: 10.1109/ICCV.2015.510.

[11] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017, pp. 4724-4733, doi: 10.1109/CVPR.2017.502.

[12] Z. Qiu, T. Yao, and T. Mei, "Learning spatio-temporal representation with pseudo-3d residual networks," in *proceedings of the IEEE International Conference on Computer Vision, (ICCV)*, Venice, Italy, 2017, pp. 5533–5541, doi: 10.1109/ICCV.2017.590.

[13] V. Venkataraman, I. Vlachos, and P. K. Turaga, "Dynamical Regularity for Action Analysis," in *British Machine Vision Conference (BMVC)*, Swansea, UK, 2015, vol. 67, pp. 67.1-67.12, doi: 10.5244/c.29.67.

[14] G. I. Parisi, S. Magg, and S. Wermter, "Human motion assessment in real time using recurrent self-organization," in *2016 25th IEEE international symposium on robot and human interactive communication (RO-MAN)*, New York ,NY, USA, Aug. 2016: IEEE, pp. 71–76, doi: 10.1109/ROMAN.2016.7745093.

[15] Y. Li, X. Chai, and X. Chen, "End-To-End Learning for Action Quality Assessment," in *Advances in Multimedia Information Processing – PCM 2018*, Munich, Germany, 2018, vol. 11165: Springer, Cham, pp. 125–134, doi: 10.1007/978-3-030-00767-6_12.

[16] P. Parmar and B. Morris, "HalluciNet-ing spatiotemporal representations using a 2D-CNN," *Signals,* vol. 2, no. 3, pp. 604–618, 2021, doi: 10.3390/signals2030037.

[17] X. Xiang, Y. Tian, A .Reiter, G. D. Hager, and T. D. Tran, "S3d: Stacking segmental p3d for action quality assessment," in *2018 25th IEEE International conference on image processing (ICIP)*, Athens, Greece, 2018: IEEE, pp. 928–932, doi: 10.1109/ICIP.2018.8451364.

[18] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks for action segmentation and detection," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 1003-1012, doi: 10.1109/CVPR.2017.113.

[19] Y. Li, X. Chai, and X. Chen, "Scoringnet: Learning key fragment for action quality assessment with ranking loss in skilled sports," in *Asian Conference on Computer Vision (ACCV)*, Trento, Italy, 2018: Springer, pp. 149–164, doi: 10.1007/978-3-030-20876-9_10.

[20] L.-J. Dong, H.-B. Zhang, Q. Shi, Q. Lei, J.-X. Du, and S. Gao, "Learning and fusing multiple hidden substages for action quality assessment," *Knowledge-Based Systems,* vol. 229, p. 107388, 2021, doi: 10.1016/j.knosys.2021.107388.

[21] H.-B. Zhang, L.-J. Dong, Q. Lei, L.-J. Yang, and J.-X. Du, "Label-reconstruction-based pseudo-subscore learning for action quality assessment in sporting events," *Applied Intelligence,* vol. 53, no. 9, pp. 10053–10067, May 2023, doi: 10.1007/s10489-022-03984-5.

[22] A. Iyer, M. Alali, H. Bodala, and S. Vaidya, "Action quality assessment using transformers," *arXiv preprint arXiv:2207.12318,* Jul. 2022, doi: 10.48550/arXiv.2207.12318.

[23] C. Xu, Y. Fu, B. Zhang, Z. Chen, Y.-G. Jiang ,and X. Xue, "Learning to score figure skating sport videos," *IEEE transactions on circuits and systems for video technology,* vol. 30, no. 12, pp. 4578–4590, 2019, doi: 10.1109/TCSVT.2019.2927118.

[24] L.-A. Zeng *et al.*, "Hybrid dynamic-static context-aware attention network for action assessment in long videos," in *Proceedings of the 28th ACM international conference on multimedia*, New York, NY, USA, Oct. 2020, pp. 2526–2534, doi: 10.1145/3394171.3413560.

[25] Q. Lei, H. Zhang, and J. Du, "Temporal attention learning for action quality assessment in sports video," *Signal, Image and Video Processing,* vol. 15, no. 7, pp. 1575–1583, Oct. 2021, doi: 10.1007/s11760-021-01890-w.

[26] S. Wang, D. Yang, P. Zhai, C. Chen, and L. Zhang, "Tsa-net: Tube self-attention network for action quality assessment," in *Proceedings of the 29th ACM international conference on multimedia*, New York, NY, USA, 2021, pp. 4902–4910, doi: 10.1145/3474085.3475438.

[27] T. Nagai, S. Takeda, M. Matsumura, S. Shimizu, and S. Yamamoto, "Action quality assessment with ignoring scene context," in *2021 IEEE International Conference on Image Processing (ICIP)*, Anchorage, AK, USA, 2021: IEEE, pp. 1189–1193, doi: 10.1109/ICIP42928.2021.9506257.

[28] S. Farabi, H. Himel, F. Gazzali, M. B. Hasan, M. H. Kabir, and M. Farazi, "Improving action quality assessment using weighted aggregation," in *Iberian Conference on Pattern Recognition and Image Analysis*, Aveiro, Portugal, 2022, vol. 13256, pp. 576–587, doi: 10.1007/978-3-031-04881-4_46.

[29] Y .Zhang, W. Xiong, and S. Mi, "Learning time-aware features for action quality assessment," *Pattern Recognition Letters,* vol. 158, pp. 104–110, Jun. 2022, doi: 10.1016/j.patrec.2022.04.015.

[30] A. Xu, L.-A. Zeng, and W.-S. Zheng, "Likert scoring with grade decoupling for long-term action assessment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 3222–3231, doi: 10.1109/CVPR52688.2022.00323.

[31] C. Han, F. Shen, L. Chen ,X. Lian, H. Gou, and H. Gao, "Mla-lstm: A local and global location attention lstm learning model for scoring figure skating," *Systems,* vol. 11, no. 1, p. 21, Jan. 2023, doi: 10.3390/systems11010021.

[32] W. Sun, Y. Hu, B. Zhang, X. Chen, C. Hao, and Y .Gao, "A novel blind action quality assessment based on multi-headed GRU network and attention mechanism," in *3rd International Conference on Artificial Intelligence, Automation, and High-Performance Computing (AIAHPC 2023)*, Guangzhou, China, 2023, vol. 12, pp. 835–843, doi: 10.1117/12.2685368.

[33] P.-X. Lian and Z.-G. Shao, "Improving action quality assessment with across-staged temporal reasoning on imbalanced data," *Applied Intelligence,* vol. 53, no. 24, pp. 30443–30454, 2023, doi: 10.1007/s10489-023-05166-3.

[34] F. Huang and J. Li, "Assessing action quality with semantic-sequence performance regression and densely distributed sample weighting," *Applied Intelligence,* vol. 54, no. 4, pp. 3245–3259, 2024, doi: 10.1007/s10489-024-05349-6.

[35] Y. Liu, X. Cheng, and T. Ikenaga, "A figure skating jumping dataset for replay-guided action quality assessment," in *Proceedings of the 31st ACM International Conference on Multimedia*, Ottawa, ON, Canada, 2023, pp. 2437–2445, doi: 10.1145/3581783.3613774.

[36] Y. Liu, X. Cheng, and T. Ikenaga, "A hierarchical joint training based replay-guided contrastive transformer for action quality assessment of figure skating," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences* ,vol. E108-A, no. 3, pp. 332–341, Mar. 2025, doi: 10.1587/transfun.2024SMP0003.

[37] X. Ke, H. Xu, X. Lin, and W. Guo, "Two-path target-aware contrastive regression for action quality assessment," *Information Sciences (Ny),* vol. 664, p. 120347, Apr. 2024, doi: 10.1016/j.ins.2024.120347.

[38] J.-H. Pan, J. Gao, and W.-S. Zheng, "Action assessment by joint relation graphs," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea (South), Oct. 2019, pp. 6330-6339, doi: 10.1109/ICCV.2019.00643.

[39] H. Li, Q. Lei, H. Zhang, J. Du, and S. Gao, "Skeleton-based deep pose feature learning for action quality assessment on figure skating videos," *Journal of Visual Communication and Image Representation,* vol. 89, p. 103625, Nov. 2022, doi: 10.1016/j.jvcir.2022.103625.

[40] S. Zhang *et al.*, "Logo: A long-form video dataset for group action quality assessment," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, Vancouver, BC, Canada, Jun. 2023, pp. 2405–2414, doi: 10.1109/CVPR52729.2023.00238.

[41] Q. Lei, H. Li, H. Zhang, J. Du, and S. Gao, "Multi-skeleton structures graph convolutional network for action quality assessment in long videos," *Applied Intelligence,* vol. 53, no. 19, pp. 21692–21705, 2023, doi: 10.1007/s10489-023-04613-5.

[42] K. Zheng, J. Wu, J. Zhang, and C. Guo, "A skeleton-based rehabilitation exercise assessment system with rotation invariance," *IEEE Transactions on Neural Systems and Rehabilitation Engineering,* vol. 31, no. 8, pp. 2612–2621, 2023, doi: 10.1109/TNSRE.2023.3282675.

[43] X. Bruce, Y. Liu, K. C. Chan, and C. W. Chen, "EGCN++: A new fusion strategy for ensemble learning in skeleton-based rehabilitation exercise assessment," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 46, no. 9, pp. 6471–6485, Sep. 2024, doi: 10.1109/TPAMI.2024.3378753.

[44] C.-I. Joung, S. Byun, and S. Baek, "Contrastive learning for action assessment using graph convolutional networks with augmented virtual joints," *IEEE Access,* vol. 11, pp. 88895–88907, 2023, doi: 10.1109/ACCESS.2023.3305372.

[45] X. Chen, A. Pang, W. Yang, Y. Ma, L. Xu, and J. Yu, "Sportscap: Monocular 3d human motion capture and fine-grained understanding in challenging sports videos," *International Journal of Computer Vision (IJCV),* vol. 129, no. 10, pp. 2846–2864, 2021, doi: 10.1007/s11263-021-01486-4.

[46] Y. Huang *et al.*, "Full-reference motion quality assessment based on efficient monocular parametric 3d human body reconstruction," in *2024 IEEE International Conference on Multimedia and Expo (ICME)*, Niagara Falls, ON, Canada, 2024: IEEE, pp. 1–6, doi: 10.1109/ICME57554.2024.10687714.

[47] M. Nekoui, F. O. T. Cruz, and L. Cheng, "EAGLE-EYE: Extreme-pose action grader using detail bird's-eye view," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision (WACV)*, Waikoloa, HI, USA, 2021, pp. 394–402, doi: 10.1109/WACV48630.2021.00044.

[48] K. Huang, Y. Tian, C. Yu, and Y. Huang, "Dual-referenced assistive network for action quality assessment," *Neurocomputing,* vol. 614, p. 128786, Jan. 2025, doi: 10.1016/j.neucom.2024.128786.

[49] Y. Tang *et al.*, "Uncertainty-aware score distribution learning for action quality assessment," in *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 2020, pp. 9839–9848, doi: 10.1109/CVPR42600.2020.00986.

[50] C. Zhou, Y. Huang, and H. Ling, "Uncertainty-driven action quality assessment," *arXiv preprint arXiv:2207.14513,* vol. 14, no. 8, pp. 1–10, Jul. 2022, doi: 10.48550/arXiv.2207.14513.

[51] Y. Ji, L. Ye, H. Huang, L. Mao, Y. Zhou, and L. Gao, "Localization-assisted uncertainty score disentanglement network for action quality assessment," in *Proceedings of the 31st ACM International Conference on Multimedia*, New York, NY, USA, Oct. 2023, pp. 8590–8597, doi: 10.1145/3581783.3613795.

[52] M.-Z. Li, H.-B. Zhang, L.-J. Dong, Q. Lei, and J.-X. Du, "Gaussian guided frame sequence encoder network for action quality assessment," *Complex & Intelligent Systems,* vol. 9, no. 2, pp. 1963–1974, Apr. 2023, doi: 10.1007/s40747-022-00892-6.

[53] B. Zhang, J. Chen, Y. Xu, H. Zhang, X. Yang, and X. Geng, "Auto-encoding score distribution regression for action quality assessment," *Neural Computing and Applications,* vol. 36, no. 2, pp. 929–942, Jan. 2024, doi: 10.1007/s00521-023-09068-w.

[54] A. Majeedi, V. R. Gajjala, S. S. S. N. GNVV, and Y. Li, "Ricaˆ 2: Rubric-informed, calibrated assessment of actions," in *European Conference on Computer Vision (ECCV)*, Munich, Germany, 2024, vol. 13697: Springer, pp. 143–161, doi: 10.1007/978-3-031-73036-8_9.

[55] J. Gao *et al.*, "An asymmetric modeling for action assessment," in *16th European Conference on Computer Vision (ECCV)*, Glasgow, UK (Virtual), 2020: Springer, pp. 222–238, doi: 10.1007/978-3-030-58577-8_14.

[56] J. Gao, J.-H. Pan, S.-J. Zhang, and W.-S. Zheng, "Automatic modelling for interactive action assessment," *International Journal of Computer Vision,* vol. 131, no. 3, pp. 659–679, 2023, doi: 10.1007/s11263-022-01695-5.

[57] P. Parmar, J. Reddy, and B. Morris, "Piano skills assessment," in *2021 IEEE 23rd international workshop on multimedia signal processing (MMSP)*, Chengdu, China, Oct. 2021: IEEE, pp. 1–5, doi: 10.1109/MMSP53017.2021.9733638.

[58] Z. Du, D. He, X. Wang, and Q. Wang, "Learning semantics-guided representations for scoring figure skating," *IEEE Transactions on Multimedia,* vol. 26, no. 10, pp. 4987–4997, 2023, doi: 10.1109/TMM.2023.3328180.

[59] J. Xia *et al.*, "Skating-mixer: Long-term sport audio-visual modeling with mlps," in *Proceedings of the AAAI Conference on Artificial Intelligence*, Jun. 2023, vol. 37, no. 3, pp. 2901–2909, doi: 10.1609/aaai.v37i3.25392 .

[60] K. Gedamu, Y. Ji, Y. Yang, J. Shao, and H. T. Shen, "Visual-semantic alignment temporal parsing for action quality assessment," *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 35, no. 11, pp. 7984–7998, 2024, doi: 10.1109/TCSVT.2024.3487242.

[61] S. Zhang *et al.*, "Narrative action evaluation with prompt-guided multimodal interaction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 2024, pp. 18430–18439, doi: 10.1109/CVPR52733.2024.01744.

[62] Y. Ding *et al.*, "2m-af: A strong multi-modality framework for human action quality assessment with self-supervised representation learning," in *Proceedings of the 32nd ACM International Conference on Multimedia*, New York, NY, USA, Oct. 2024, pp. 1564–1572, doi: 10.1145/3664647.3681084.

[63] T. He ,Y. Chen, L. Wang, and H. Cheng, "An expert-knowledge-based graph convolutional network for skeleton-based physical rehabilitation exercises assessment," *IEEE Transactions on Neural Systems and Rehabilitation Engineering,* vol. 32, pp. 1916–1925, 2024, doi: 10.1109/TNSRE.2024.3400790.

[64] S. Zahan, G. M. Hassan, and A. Mian, "Learning sparse temporal video mapping for action quality assessment in floor gymnastics," *IEEE Transactions on Instrumentation and Measurement,* vol. 73, no. 12, pp. 1–11, 2024, doi: 10.1109/TIM.2024.3398072.

[65] L.-A. Zeng and W.-S. Zheng, "Multimodal action quality assessment," *IEEE Transactions on Image Processing,* vol. 33, pp. 1600–1613, 2024, doi: 10.1109/TIP.2024.3362135.

[66] J.-H. Pan, J. Gao, and W.-S. Zheng, "Adaptive action assessment," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 44, no. 12, pp. 8779–8795, 2021, doi: 10.1109/TPAMI.2021.3126534.

[67] S.-J. Zhang, J.-H. Pan, J. Gao, and W.-S. Zheng, "Adaptive stage-aware assessment skill transfer for skill determination," *IEEE Transactions on Multimedia,* vol. 26, no. 10, pp. 4061–4072, 2023, doi: 10.1109/TMM.2023.3294800.

[68] J. Peng, M. Li, and H. Wang, "Stabilizing holistic semantics in diffusion bridge for image inpainting," in *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI)*, California, 2025, pp. 1756–1764, doi: 10.24963/ijcai.2025/196.

[69] A. Dadashzadeh, S. Duan, A. Whone, and M. Mirmehdi, "Pecop: Parameter efficient continual pretraining for action quality assessment," in *Proceedings of the IEEE/CVF Winter Conference on applications of computer vision (WACV)*, Waikoloa, HI, USA, 2024, pp. 42–52, doi: 10.1109/WACV57701.2024.00012.

[70] Y.-M. Li, L.-A. Zeng, J.-K. Meng, and W.-S. Zheng, "Continual action assessment via task-consistent score-discriminative feature distribution modeling," *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 34, no. 10, pp. 9112–9124, 2024, doi: 10.1109/TCSVT.2024.3396692.

[71] K. Zhou *et al.*, "Magr: Manifold-aligned graph regularization for continual action quality assessment," in *European Conference on Computer Vision (ECCV)*, London, UK, 2025: Springer, pp. 375–392, doi: 10.1007/978-3-031-73247-8_22.

[72] P. Parmar and B. T. Morris, "What and how well you performed? a multitask learning approach to action quality assessment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, 2019, pp. 304–313, doi: 10.1109/CVPR.2019.00039.

[73] L .Okamoto and P. Parmar, "Hierarchical neurosymbolic approach for comprehensive and explainable action quality assessment," in *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition (CVPR)*, Seattle, WA, USA, 2024, pp. 3204–3213, doi: 10.1109/CVPRW63382.2024.00326.

[74] H. Matsuyama, N. Kawaguchi, and B. Y. Lim, "Iris: Interpretable rubric-informed segmentation for action quality assessment," in *Proceedings of the 28th International Conference on Intelligent User Interfaces*, New York, NY, USA, Mar. 2023, pp. 368–378, doi: 10.1145/3581641.3584048.

[75] X. Dong, X. Liu, W. Li, A. Adeyemi-Ejeye, and A. Gilbert, "Interpretable long-term action quality assessment," *arXiv preprint arXiv:2408.11687,* 2024, doi: 10.48550/arXiv.2408.11687.

[76] T. Wang, M. Jin, and M. Li, "Towards accurate and interpretable surgical skill assessment: a video-based method for skill score prediction and guiding feedback generation," *International Journal of Computer Assisted Radiology and Surgery,* vol. 16, no. 9, pp. 1595–1605, 2021, doi: 10.1007/s11548-021-02448-4.

[77] D. Liu *et al.*, "Towards unified surgical skill assessment," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, Virtual (Nashville, TN, USA), 2021 ,pp. 9522–9531, doi: 10.1109/CVPR46437.2021.00940.

[78] J. Wang *et al.*, "Will you ever become popular? Learning to predict virality of dance clips," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM),* vol. 18, no. 2, pp. 1–24, 2022, doi: 10.1145/3477533.

[79] K. Roditakis, A. Makris, and A. Argyros, "Towards improved and interpretable action quality assessment with self-supervised alignment," in *Proceedings of the 14th PErvasive Technologies Related to Assistive Environments Conference (PETRA)*, Corfu, Greece, June 2021, pp. 507–513, doi: 10.1145/3453892.3461624.

[80] P. Parmar, A. Gharat, and H. Rhodin, "Domain knowledge-informed self-supervised representations for workout form assessment," in *European Conference on Computer Vision (ECCV)*, Tel Aviv, Israel, 2022: Springer, pp. 105–123, doi: 10.1007/978-3-031-19839-7_7.

[81] P. Parmar, E. Peh, and B. Fernando, "Learning to visually connect actions and their effects," in *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, Tucson, AZ, USA, Feb. 2025: IEEE, pp. 1477–1487, doi: 10.1109/WACV61041.2025.00151.

[82] S.-J. Zhang, J.-H. Pan, J. Gao, and W.-S. Zheng, "Semi-supervised action quality assessment with self-supervised segment feature recovery," *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 32, no. 9, pp. 6017–6028, Sep. 2022, doi: 10.1109/TCSVT.2022.3143549.

[83] K. Gedamu, Y. Ji, Y. Yang, J. Shao, and H. T. Shen, "Self-supervised subaction parsing network for semi-supervised action quality assessment," *IEEE Transactions on Image Processing,* vol. 33, no. 11, pp. 6057–6070, 2024, doi: 10.1109/TIP.2024.3468870.

[84] W. Yun, M. Qi, F. Peng, and H. Ma, "Semi-supervised teacher-reference-student architecture for action quality assessment," in *European Conference on Computer Vision (ECCV)*, Milano, Italy, 2024: Springer, pp. 161–178, doi: 10.1007/978-3-031-72904-1_10.

[85] G. Bertasius, H. Soo Park, S. X. Yu, and J. Shi, "Am I a baller? Basketball performance assessment from first-person videos," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, Oct. 2017, pp. 2196-2204, doi: 10.1109/ICCV.2017.239.

[86] H. Doughty, D. Damen, and W. Mayol-Cuevas, "Who's better? who's best? pairwise deep ranking for skill determination," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, Salt Lake City, UT, USA, Jun. 2018, pp. 6057–6066, doi: 10.1109/CVPR.2018.00634.

[87] H. Doughty, W. Mayol-Cuevas, and D. Damen, "The pros and cons: Rank-aware temporal attention for skill determination in long videos," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 7862–7871, doi: 10.1109/CVPR.2019.00805.

[88] Z. Li ,Y. Huang, M. Cai, and Y. Sato, "Manipulation-skill assessment from videos with spatial attention network," in *Proceedings of the IEEE/CVF international conference on computer vision workshops (ICCVW)*, Seoul, South Korea, Oct. 2019, pp. 4385–4395, doi: 10.1109/ICCVW.2019.00539.

[89] H. Jain, G. Harit, and A. Sharma, "Action quality assessment using siamese network-based deep metric learning," *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 31, no. 6, pp. 2260–2273, 2021, doi: 10.1109/TCSVT.2020.3017727.

[90] X. Yu, Y. Rao, W. Zhao, J. Lu, and J. Zhou, "Group-aware contrastive regression for action quality assessment," in *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, Montreal, QC, Canada, Oct. 2021 ,pp. 7899-7908, doi: 10.1109/ICCV48922.2021.00782.

[91] Y. Bai *et al.*, "Action quality assessment with temporal parsing transformer," in *European conference on computer vision*, 2022, vol. 13664: Springer, pp. 422–438, doi: 10.1007/978-3-031-19772-7_25.

[92] M. Li, H.-B. Zhang, Q. Lei, Z. Fan, J. Liu, and J.-X. Du, "Pairwise contrastive learning network for action quality assessment," in *European Conference on Computer Vision*, 2022, vol. 13664: Springer, pp. 457–473, doi: 10.1007/978-3-031-19772-7_27.

[93] K. Gedamu, Y. Ji, Y. Yang, J. Shao, and H. T. Shen, "Fine-grained spatio-temporal parsing network for action quality assessment," *IEEE Transactions on Image Processing,* vol. 32, pp. 6386–6400, 2023, doi: 10.1109/TIP.2023.3331212.

[94] I. Hipiny, H. Ujir, A. A. Alias, M. Shanat, and M. K. Ishak, "Who danced better? ranked tiktok dance video dataset and pairwise action quality assessment method," *International Journal of Advances in Intelligent Informatics,* vol. 9, no. 1, pp. 96–107, Mar. 2023, doi: 10.26555/ijain.v9i1.919.

[95] L. Liu, P. Zhai, D. Zheng, and Y. Fang, "Multi-stage action quality assessment method," in *Proceedings of the 2023 4th International Conference on Control, Robotics and Intelligent System*, New York, NY, USA, 2023, pp. 116–122, doi: 10.1145/3622896.3622916.

[96] Q. An, M. Qi, and H. Ma, "Multi-stage contrastive regression for action quality assessment," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seoul, South Korea, Apr. 2024, pp. 4110–4114. doi: 10.1109/ICASSP48485.2024.10447069.

[97] J. Xu, S. Yin, G. Zhao, Z. Wang, and Y. 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 (CVPR)*, Seattle, WA, USA, 2024, pp. 14628–14637, doi: 10.1109/CVPR52733.2024.01386.

[98] Z. Luo, Y. Xiao, F. Yang, J. T. Zhou, and Z. Fang, "Rhythmer: Ranking-based skill assessment with rhythm-aware transformer," *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 35, no. 1, pp. 259–272, 2025, doi: 10.1109/TCSVT.2024.3459938.

[99] J. Xu, Y. Rao, X. Yu, G. Chen, J. Zhou, and J. 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 (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 2939-2948, doi: 10.1109/CVPR52688.2022.00296.

[100] J. Xu, Y. Rao, J. Zhou, and J. Lu, "Procedure-aware action quality assessment: Datasets and performance evaluation," *International Journal of Computer Vision,* vol. 132, no. 12, pp. 6069–6090, Dec. 2024, doi: 10.1007/s11263-024-02146-z.

[101] M. Fang, X. Du, Q. Liu, Y. Zhou, Q. Liang, and S. Liu" ,Which is the better teacher action? a new ranking model and dataset," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seoul, South Korea, 2024: IEEE, pp. 7695–7699, doi: 10.1109/ICASSP48485.2024.10448158.

[102] A. S. Gordon, "Automated video assessment of human performance," in *Proceedings of AI-ED*, Washington, D.C., USA, 1995, vol. 2, p. 10.

[103] L. Tao, E. Elhamifar, S. Khudanpur, G. D. Hager, and R. Vidal, "Sparse hidden markov models for surgical gesture classification and skill evaluation," in *International conference on information processing in computer-assisted interventions (IPCAI)*, Pisa, Italy, 2012: Springer, pp. 167–177, doi: 10.1007/978-3-642-30618-1_17.

[104] P. Parmar and B. T. Morris, "Measuring the quality of exercises," in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Orlando, FL, USA, 2016: IEEE, pp. 2241–2244, doi: 10.1109/EMBC.2016.7591175.

[105] A. Zia and I. Essa, "Automated surgical skill assessment in RMIS training," *International journal of computer assisted radiology and surgery,* vol. 13, no. 5, pp. 731–739, May 2018, doi: 10.1007/s11548-018-1735-5.

[106] A. Zia, Y. Sharma, V. Bettadapura, E. L. Sarin, and I. Essa, "Video and accelerometer-based motion analysis for automated surgical skills assessment," *International journal of computer assisted radiology and surgery,* vol. 13, no. 3, pp. 443–455, Mar. 2018, doi: 10.1007/s11548-018-1704-z.

[107] R. Ogata, E. Simo-Serra, S. Iizuka, and H. Ishikawa, "Temporal distance matrices for squat classification," in *Proceedings of the ieee/cvf conference on computer vision and pattern recognition workshops (CVPRW)*, Long Beach, CA, USA, 2019, pp. 1083–1091, doi: 10.1109/CVPRW.2019.00309.

[108] S. Liu *et al.*, "FSD-10: A fine-grained classification dataset for figure skating," *Neurocomputing,* vol. 413, no. 1, pp. 360–367, Nov. 2020, doi: 10.1016/j.neucom.2020.06.108.

[109] C. Li, X. Ling, and S. Xia, "A graph convolutional siamese network for the assessment and recognition of physical rehabilitation exercises," in *International conference on artificial neural networks (ICANN)*, Heraklion, Crete, Greece, 2023: Springer, pp. 229–240, doi: 10.1007/978-3-031-44216-2_19.

[110] B. Baby, M. Chasmai, T .Banerjee, A. Suri, S. Banerjee, and C. Arora, "Representation learning using rank loss for robust neurosurgical skills evaluation," in *2022 IEEE International Conference on Image Processing (ICIP)*, Bordeaux, France, 2022: IEEE, pp. 4048–4052, doi: 10.1109/ICIP46576.2022.9897932.

[111] Z. Gan *et al.*, "SkatingVerse: A large-scale benchmark for comprehensive evaluation on human action understanding," *IET Computer Vision,* vol. 18, no. 7, pp. 888–906, 2024, doi: 10.1049/cvi2.12287.

[112] K. Zhou, J. Li, R. Cai, L. Wang, X. Zhang, and X. Liang, "Cofinal: Enhancing action quality assessment with coarse-to-fine instruction alignment," *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence,* pp. 1771–1779, 2024, doi: 10.24963/ijcai.2024/196.

[113] D. Shao, Y. Zhao, B. Dai, and D. Lin, "Finegym: A hierarchical video dataset for fine-grained action understanding," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, Seattle, WA, USA, 2020 :IEEE, pp. 2613–2622, doi: 10.1109/CVPR42600.2020.00269.

[114] Y. Pan, C. Zhang, and G. Bertasius, "BASKET: A Large-Scale Video Dataset for Fine-Grained Skill Estimation," in *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, Phoenix, AZ, USA, 2025, pp. 28952–28962, doi: 10.1109/CVPR52734.2025.02696.

[115] Y. Gao *et al.*, "JHU-ISI Gesture and Skill Assessment Working Set (JIGSAWS) : A Surgical Activity Dataset for Human Motion Modeling," *Model. Monit. Comput. Assist. Interv. – MICCAI Work.*, vol. 3, no. 3, pp. 1–10, 2014.

[116] T.-N. Nguyen, H.-H. Huynh, and J. Meunier, "3D reconstruction with time-of-flight depth camera and multiple mirrors," *IEEE Access,* vol. 6, pp. 38106–38114, 2018 ,doi: 10.1109/ACCESS.2018.2854262.

[117] A. Vakanski, H.-p. Jun, D. Paul, and R. Baker, "A data set of human body movements for physical rehabilitation exercises," *Data,* vol. 3, no. 1, p. 2, 2018, doi: 10.3390/data3010002.

[118] M. Capecci *et al.*, "The kimore dataset: Kinematic assessment of movement and clinical scores for remote monitoring of physical rehabilitation," *IEEE Transactions on neural systems and rehabilitation engineering,* vol. 27, no. 7, pp. 1436–1448, 2019, doi: 10.1109/TNSRE.2019.2923060.

[119] X. Bruce, Y. Liu, K. C. Chan, Q. Yang, and X. Wang, "Skeleton-based human action evaluation using graph convolutional network for monitoring Alzheimer's progression," *Pattern Recognition,* vol. 119, p. 108095, 2021, doi: 10.1016/j.patcog.2021.108095.

[120] J. Li *et al.*, "Mmasd: A multimodal dataset for autism intervention analysis," in *Proceedings of the 25th International Conference on Multimodal Interaction (ICMI)*, Paris, France, 2023, pp. 397–405, doi: 10.1145/3577190.3614117.

[121] J. Li *et al.*, "Finerehab: A multi-modality and multi-task dataset for rehabilitation analysis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Seattle, WA, USA, 2024: IEEE, pp. 3184–3193, doi: 10.1109/CVPRW63382.2024.00324.

[122] K. Grauman *et al.*, "Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives," *International Journal of Computer Vision,* vol. 133, pp. 8356–8435, 2025, doi: 10.1007/s11263-025-02557-6.

[123] M. Tits, S .Laraba, E. Caulier, J. Tilmanne, and T. Dutoit, "UMONS-TAICHI: A multimodal motion capture dataset of expertise in Taijiquan gestures," *Data in brief,* vol. 19, pp. 1214–1221, 2018, doi: 10.1016/j.dib.2018.05.088.

[124] Y.-M. Li, W.-J. Huang, A.-L. Wang ,L.-A. Zeng, J.-K. Meng, and W.-S. Zheng, "Egoexo-fitness: Towards egocentric and exocentric full-body action understanding," in *European Conference on Computer Vision (ECCV)*, Milano, Italy, 2024, vol. 15078: Springer, Cham, pp. 363–382, doi: 10.1007/978-3-031-72661-3_21.

[125] Z. Chen *et al.*, "Gaia: Rethinking action quality assessment for ai-generated videos," *Advances in Neural Information Processing Systems (NeurIPS)*, Vancouver, BC, Canada, vol. 37, pp. 40111–40144, 2024, doi: 10.52202/079017-1267.