ExpertEdit: Learning Skill-Aware Motion Editing from Expert Videos

Arjun Somayazulu[1] and Kristen Grauman[1]

The University of Texas at Austin, USA

Abstract. Visual feedback is critical for motor skill acquisition in sports and rehabilitation, and psychological studies show that observing near-perfect versions of one's *own* performance accelerates learning more effectively than watching expert demonstrations alone. We propose to enable such personalized feedback by automatically editing a person's motion to reflect higher skill. Existing motion editing approaches are poorly suited for this setting because they assume paired input-output data—rare and expensive to curate for skill-driven tasks—and explicit edit guidance at inference. We introduce **ExpertEdit**, a framework for skill-driven motion editing trained exclusively on *unpaired* expert video demonstrations. ExpertEdit learns an expert motion prior with a masked language modeling objective that infills masked motion spans with expert-level refinements. At inference, novice motion is masked at skill-critical moments and projected into the learned expert manifold, producing localized skill improvements without paired supervision or manual edit guidance. Across eight diverse techniques and three sports from Ego-Exo4D [15] and Karate Kyokushin [52], ExpertEdit outperforms state-of-the-art supervised motion editing methods on multiple metrics of motion realism and expert quality. Project page and benchmark available at: https://vision.cs.utexas.edu/projects/expert_edit/

Keywords: Pose editing · Motion generation · Skilled activity

1 Introduction

Imagine watching a video of yourself performing a layup, but with the smooth form, cadence, and precision of an NBA-level player. The shot still rises from your hands and arcs toward the basket, yet your body moves with trained efficiency, timing and rhythm, capturing the expert nuances of a player who's practiced thousands of times. Just as Auto-Tune subtly corrects a singer's pitch by a semitone while preserving their vocal timbre and phrasing, this imagined edit auto-tunes *your motion*: infusing expert-like form into your exact execution while preserving your gross movement path and orientation in space.

Such personalized edited video has broad applications. Psychological research on video self-modeling shows that observing near-perfect versions of one's *own* performance improves motor skill acquisition, confidence, and retention [46,50,51]. Seeing one's own body perform with expertise engages the perceptual–motor

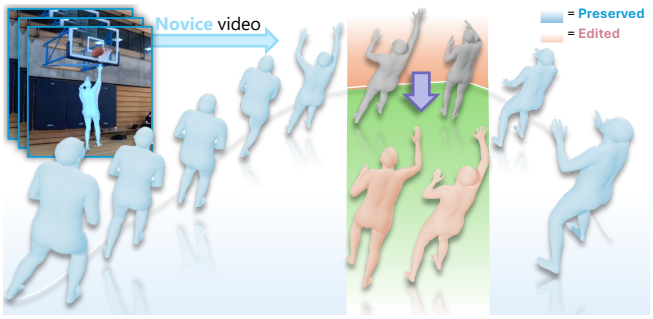

Fig. 1: Skill-driven motion editing. Given a 3D motion sequence extracted from a novice activity video, ExpertEdit produces personalized skill edits by refining poses within regions where skill differences are pronounced. Our model tweaks them to exhibit expert-like precision and form, while preserving the original execution's motion path and body orientation, as well as its source poses at all non-skill-critical moments. We learn to perform these edits solely from expert videos, without paired supervision or heavy edit guidance from text or reference motions—whether during inference or training. Preserved source poses are shown in blue; edited poses are shown in orange.

system more effectively than watching another person. This means personalized motion editing would be especially valuable for AI coaching, where individuals could receive feedback by viewing improved versions of their own movement within the original scene and viewpoint. Beyond coaching, such editing could help actors appear more skilled in stunts or choreography without body doubles, allow users to enhance their own performances (e.g., TikTok dances), and could even be used to refine the vast web of amateur activity footage into higher-quality human demonstrations for downstream robotics applications.

Despite this promise, skill-driven motion editing exposes a fundamental limitation of existing methods. Prior motion editing approaches assume either explicit edit conditioning—via text descriptions or reference clips [27, 30, 32, 55, 56, 73]—or access to large-scale motion-text-motion triplets [5, 8, 27]. While these assumptions are reasonable for coarse edits (e.g., walking to dancing), they break down for skill refinement, where the goal is to improve execution of the same action rather than alter the action itself. Creating such supervision for skilled activities requires expert knowledge and significant manual effort. For example, MotionFix [5] constructs candidate motion pairs through large-scale motion similarity search followed by human annotation of edit descriptions, yet over 55% of candidate pairs are rejected as 'too similar' for labeling. Extending such pipelines to skill refinement would additionally require domain experts capable of diagnosing and describing subtle execution errors, making annotation difficult to scale. Consequently, no large-scale paired novice–expert motion dataset exists, and skill annotations remain scarce. At inference time, adapting the existing editing methods would require a human expert-in-the-loop to provide detailed text corrections or precisely aligned reference clips—precisely the form of feedback that AI coaching seeks to automate.

We therefore introduce **ExpertEdit** for *skill-driven motion editing*: given a novice motion sequence, ExpertEdit transforms it into an expert-like performance (Fig. 1). Unlike traditional motion editing formulations, our goal is to improve execution of the same action—transforming a novice performance into an expert-like one—while preserving the actor's identity and overall motion structure. Crucially, this process must operate without explicit edit guidance or reference demonstrations. In particular, we design ExpertEdit so it can be trained solely on unpaired expert motion and operate fully conditioning-free at inference time. By learning directly from expert demonstrations, our approach eliminates the need for paired supervision, expensive data curation, or explicit 'correction' annotations.

Our key idea is to cast skill refinement as *contextual motion infilling*. During training, ExpertEdit learns to reconstruct spans of motion using a masked language modeling (MLM) objective. Rather than masking arbitrarily, spans are centered around kinematic peaks—identified via simple motion statistics such as velocity or acceleration—where execution quality is most pronounced (e.g., takeoff in a layup). At test time, the same criteria select segments of a novice execution for infilling with expert-like refinements, effectively projecting key phases of novice motion onto the learned expert motion manifold.

Evaluated on eight diverse techniques spanning basketball, soccer, and karate across two datasets, ExpertEdit consistently outperforms state-of-the-art motion editing methods [5, 30, 32] across multiple metrics of expert motion quality and realism, despite requiring no edit guidance at inference. These results establish skill-driven motion editing as a new capability for motion editing.

2 Related Work

Skill assessment from video. Skill understanding in video has been a longstanding challenge, with early benchmarks and methods for action quality assessment (AQA) in domains such as diving [43, 61], strength training [41, 70], and figure skating [34]. Recent works expand this focus to multi-view, multi-agent sports including soccer [15, 39, 47, 68] and basketball [4, 15, 40, 59]. Early AQA works [42] treated skill as a regression task, while later efforts adopted ranking- and graph-based approaches for greater robustness [11]. More recent systems incorporate sub-action labels for procedure-aware evaluation [62]. Self-supervised approaches [9, 41] learn skill-sensitive motion representations from unlabeled data using priors like periodicity and cycle-consistency. Recent work leverages narrations or assessment rubrics [38, 60, 74]. However, these works focus on assessing or ranking skill rather than modifying motion to improve its execution quality. We address a generative problem: transforming novice motion into expert-like performance without paired supervision or explicit edit guidance.

AI coaching from video. Recent work in AI coaching seeks to move beyond skill assessment by generating actionable feedback from video in the form of text commentary or reference clip retrieval. Several approaches generate corrective text suggestions based on pose deviations or visual analysis [2, 3, 67, 69]. PoseTutor [10],

for example, highlights incorrect joint locations in static exercise images, relying on classifiers trained on labeled pose categories. Other works compute angular joint deviations between a source performance and an aligned expert reference to generate corrective text [13] or pose edits [7, 33], requiring frame-level alignment with expert demonstrations. Retrieval-based systems such as Vid2Coach [24] provide step-level feedback for procedural tasks by leveraging large collections of visual and textual how-to resources. More recently, ExpertAF [3] learns skill-corrective pose edits from paired novice–expert motion sequences annotated with expert commentary. In contrast to these approaches, which depend on text guidance, expert reference pairs, or supervision, ExpertEdit learns to refine motion directly from unpaired expert demonstrations. Our method generates skill-improved motion without requiring expert-provided commentary, frame-level alignment, or paired novice–expert samples during training.

Motion style transfer. Motion style transfer re-synthesizes a source motion sequence with stylistic characteristics derived from a reference motion. Classic approaches formulate this as a motion-style disentanglement problem using GANs [12], autoencoders [1, 21, 25, 53], or transformers [29]. Other work transfers motion across characters with different skeletons using kinematic constraints [57], and diffusion-based approaches perform motion-to-motion translation driven by reference style signals [22]. While effective for global attributes such as identity or affect, these approaches rely on synchronized reference motions at inference time. In contrast, ExpertEdit projects novice motion into the learned expert-style motion manifold at inference, refining novice execution without reference clips or explicit style signals.

Instruction-driven motion editing. Text-driven motion editing has recently emerged as a popular paradigm. Recent methods apply transformer-based diffusion models to SMPL motion sequences, editing motions using low-level kinematic text prescriptions [8, 32] (e.g., “raise the right arm by 15 degrees") or discretized, code-level motion editing operators [14]. Other works treat motion as a language modeling task, generating sequences autoregressively [27, 37, 71], with masked token modeling [16], or jointly with text [26, 58]. Other motion editing approaches animate actor images or video clips conditioned on motion from pose sequences [6, 23, 35, 63, 76, 77] or paired video clips [55, 56, 66]. While effective in human-in-the-loop settings, these approaches rely on privileged text or reference clips to guide the desired modifications. In contrast, ExpertEdit learns skill-driven motion refinements by observing expert demonstrations, and automatically determines where and how a novice execution should be improved.

3 ExpertEdit

We explore *skill-driven motion editing*: the automatic transformation of a novice 3D motion sequence into an expert-like execution of the same action and duration, without text prompts, reference clips, or user-specified pose corrections.

Our goal is to synthesize body motion that exhibits expert-level precision and control while preserving the actor’s *global translation*, *root orientation*, and

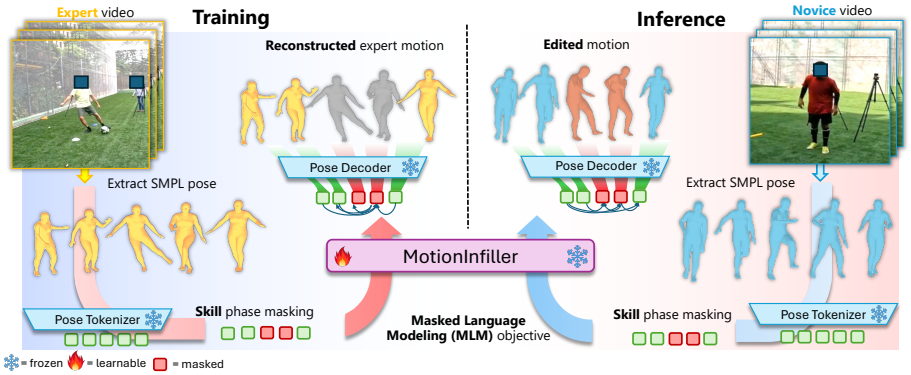

Fig. 2: ExpertEdit approach. We tokenize expert pose motion sequences and mask the key action phase as determined by task-specific kinematic criteria. We train a bidirectional transformer, **MotionInfiller**, to predict the expert pose tokens at the masked positions. At inference, we mask skill-critical action phases in a novice motion (see Sec. 4) and infill these regions with expert-like motion.

overall *body shape*. We operate in 3D pose space, abstracting away scene context to focus on execution quality encoded in body kinematics. While scene context (e.g., distance to the hoop/net, position of the ball in basketball) influences an action, the quality of its execution is concentrated in body pose dynamics.

Rather than replacing the motion entirely, we perform temporally localized pose edits centered on kinematically defined phases of the action, leaving the remaining frames unchanged. We discuss this procedure later. The result is a smooth motion sequence that preserves the source execution's trajectory and identity, while reflecting expert-level form at key phases of the action where skill differences are pronounced.

3.1 Task formulation

We represent a motion sequence as

$$\mathbf{X} = \{(\mathbf{r}_t, \mathbf{o}_t, \mathbf{p}_t)\}_{t=1}^{T},$$

where T is the sequence length, $\mathbf{r}_t \in \mathbb{R}^3$ denotes the global translation, $\mathbf{o}_t \in \mathbb{R}^3$ denotes the root orientation (in axis-angle form), and $\mathbf{p}_t \in \mathbb{R}^{3J}$ encodes the axis-angle rotations of J skeletal joints at time t. The number of joints J depends on the underlying pose representation (e.g., SMPL [36] or dataset-specific motion capture systems), but our formulation is agnostic to this choice.

Given a novice motion sequence $\mathbf{X}^{\text{nov}}$ demonstrating a particular skilled technique τ, our goal is to generate an edited motion sequence $\mathbf{X}^{\text{edit}}$ of the same duration that reflects expert-level quality in the same execution. Importantly, we preserve the source execution's global translation trajectory $\{\mathbf{r}_t\}_{t=1}^{T}$ and root orientation $\{\mathbf{o}_t\}_{t=1}^{T}$, restricting edits to the joint rotations $\mathbf{p}_t$ during selected motion phases. This ensures that the edited motion remains anchored to the

original actor's motion path and movement structure while improving the skill level of the execution. To do this, we learn a mapping

$$\mathcal{F}_\theta^\tau : \mathbb{R}^{T\times 3J} \to \mathbb{R}^{T\times 3J}, \qquad \hat{\mathbf{p}}_{1:T} = \mathcal{F}_\theta^\tau(\mathbf{p}_{1:T}), \tag{1}$$

where $\hat{\mathbf{p}}_t$ denotes the refined joint rotations for frame t, and J is the number of skeletal joints. For each frame, we construct the edited motion as

$$\mathbf{X}_t^{\text{edit}} = (\mathbf{r}_t, \mathbf{o}_t, \hat{\mathbf{p}}_t),$$

where the root translation $\mathbf{r}_t$ and root orientation $\mathbf{o}_t$ are directly copied from the input sequence. This preserves the actor's body shape, spatial trajectory, and orientation while allowing expert-like quality to emerge through learned body pose refinements across the motion sequence.

While temporal pacing and speed contribute to expertise, we preserve the input sequence's duration and overall rhythm when editing motion, focusing the model on correcting joint coordination at skill-critical phases rather than altering the global tempo of the movement.

Technique criteria. ExpertEdit operates on action-centric motions corresponding to goal-directed athletic techniques such as a reverse layup, penalty kick, or a martial arts strike. These actions exhibit a relatively consistent execution structure across performers, making them well suited for learning expert motion priors. A key property of techniques is the presence of an *operative moment*, a temporally localized event that determines the success of the action (e.g., ball release in a layup, or foot-ball contact in a penalty kick). The surrounding motion trajectory is organized around this moment, producing a canonical temporal structure shared across performers. We automate extraction of technique instances using text similarity between grounded narrations and technique operative moment descriptions (details in Supp.). In practice, technique instances can be identified within untrimmed recordings using grounded action annotations or narrations, or other temporal metadata commonly available in modern video datasets.

3.2 Approach

We approach skill-driven motion editing as the problem of learning an *expert motion manifold* from unpaired expert demonstrations, and using it to locally refine novice executions. Rather than generating motion from scratch, our goal is to project skill-critical phases of a novice sequence onto the manifold of expert motion while preserving the actor's global trajectory, orientation, and body configuration. We achieve this through *contextual motion infilling*: the model learns to reconstruct masked segments of expert motion during training, and at inference selectively replaces critical phases of a novice sequence with expert-like refinements conditioned on surrounding motion context. This formulation enables localized, personalized skill enhancement without requiring paired novice–expert data, text instructions, or reference demonstrations.

At a high level, ExpertEdit has three steps: (1) learn a discrete expert-motion vocabulary, (2) train a bidirectional infiller to reconstruct kinematically selected expert motion spans, and (3) apply the same mask-and-infill procedure to novice motions at inference. See Fig. 2.

We first train a **Pose Tokenizer** on expert motion sequences, learning a codebook of discrete, skilled motion tokens. We then train a bidirectional transformer, **MotionInfiller**, on tokenized expert sequences with a Masked Language Modeling (MLM) objective to infill kinematically selected windows of skill-critical moments conditioned on both past and future motion. At inference, we use the same kinematic criteria (defined below) to identify skill-critical moments in a novice sequence, and let the model infill those regions with expert-like poses accounting for the context of the surrounding novice poses. We retain the novice pose at unmasked regions, allowing the model to produce locally expert-like motion while retaining the actor's global trajectory and execution rhythm. Each component is described in detail below.

Pose tokenizer. We adopt a Transformer-based VQ-VAE [37] with causal self-attention as our **Pose Tokenizer**, where each encoded latent depends on the temporal history of preceding poses. Unlike frame-wise quantization, this formulation preserves temporal coherence and encodes motion dynamics directly into the latent sequence. Given expert motion sequence

$$\mathbf{X}^{\text{exp}} = \{(\mathbf{r}_t^{\text{exp}}, \mathbf{o}_t^{\text{exp}}, \mathbf{p}_t^{\text{exp}})\}_{t=1}^{T},$$

we concatenate the root translation, root orientation, and joint rotations at each frame into a pose feature vector $\mathbf{x}_t^{\text{exp}} = [\mathbf{r}_t^{\text{exp}}, \mathbf{o}_t^{\text{exp}}, \mathbf{p}_t^{\text{exp}}] \in \mathbb{R}^{6+3J}$. The encoder E_ϕ produces a sequence of latent vectors $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_T]$ where each latent $\mathbf{z}_t$ attends only to the motion up to time t:

$$\mathbf{z}_t = E_\phi\big(\mathbf{x}_{\le t}^{\text{exp}}\big) = f_{\text{enc}}\big(\mathbf{x}_t^{\text{exp}},\, \text{Attn}_{\text{causal}}(\mathbf{x}_{<t}^{\text{exp}})\big), \tag{2}$$

where $\text{Attn}_{\text{causal}}$ denotes transformer blocks with causal masking that restrict attention to prior frames. Each latent is then quantized to its nearest code in the learned codebook $\mathcal{E} = \{\mathbf{e}_k\}_{k=1}^{K}$:

$$k_t^* = \arg\min_k \ \|\mathbf{z}_t - \mathbf{e}_k\|_2^2, \qquad \hat{\mathbf{x}}_t = D_\phi\big(\mathbf{e}_{k_t^*}\big), \tag{3}$$

where D_ϕ is a causal Transformer decoder that reconstructs each frame conditioned on the current and previously decoded tokens, $\hat{\mathbf{x}}_t = D_\phi(\mathbf{e}_{\le t})$. The model is trained using the standard VQ-VAE loss:

$$\mathcal{L}_{\text{VQ}} = \|\mathbf{x}_t^{\text{exp}} - \hat{\mathbf{x}}_t\|_2^2 + \|\text{sg}[E_\phi(\mathbf{x}_{\le t}^{\text{exp}})] - \mathbf{e}_{k_t^*}\|_2^2 + \beta\, \|E_\phi(\mathbf{x}_{\le t}^{\text{exp}}) - \text{sg}[\mathbf{e}_{k_t^*}]\|_2^2, \tag{4}$$

where $\text{sg}[\cdot]$ is the stop-gradient operator and β is the commitment weight. The tokenizer learns a discrete motion vocabulary in which each token encodes a short-term motion primitive, and a sequence of tokens is represented as $\mathbf{k} = [k_1, \ldots, k_T], \quad k_t \in \{1, \ldots, K\}$. The decoder D_ϕ provides a geometry-consistent inverse mapping from tokenized motion back to full motion space.

Motion infilling. To perform context-conditioned motion editing, we adopt a **masked language modeling (MLM)** objective. Given tokenized expert motion sequence $\mathbf{k}^{\text{exp}} = [k_1^{\text{exp}}, \ldots, k_T^{\text{exp}}]$, where $k_t^{\text{exp}} \in \{1, \ldots, K\}$ during training, we use kinematic criteria to select a skill-critical motion peak index t^*, described later.

We then define a masked span length $\ell = \max(2, \lfloor \alpha T \rfloor)$, where $\alpha \in (0, 1)$ controls the maximum proportion of the sequence to mask, and replace the corresponding token segment $[k_{t^*-h}^{\text{exp}}, \ldots, k_{t^*+h}^{\text{exp}}], \quad h = \lfloor \frac{\ell}{2} \rfloor$ centered at t^* with learned `[MASK]` tokens. This embedding is excluded from the output token vocabulary. The model is then trained to reconstruct the original tokens in the masked region given the unmasked temporal context on both sides:

$$\mathcal{L}_{\text{MLM}} = - \sum_{i=t^*-h}^{t^*+h} \log p_\theta\big(k_i^{\text{exp}} \mid \mathbf{k}_{\setminus[t^*-h:t^*+h]}^{\text{exp}}\big), \tag{5}$$

where $p_\theta(\cdot)$ is the transformer's predicted probability over the token vocabulary of size K. This cross-entropy loss encourages the network to infer contextually consistent motion that bridges masked spans with smooth and biomechanically valid transitions. We refer to this bidirectional transformer as **MotionInfiller**.

ExpertEdit is technique-specific by design. For each technique τ (e.g., reverse layup, penalty kick), we train a separate Pose Tokenizer and MotionInfiller using the same architecture and objective. This lets each model learn a focused expert prior over a coherent motion family. This specialization is lightweight in our setting: adding a new skill requires only unpaired expert demonstrations, not paired novice–expert samples or per-instance text edits. Thus, technique-specific training gives the benefit of sharper expert priors without the annotation burden of constructing a paired, heterogeneous multi-skill correction dataset.

Kinematic phase selection. Skill differences in physical actions typically concentrate around a small number of biomechanically critical phases (e.g., takeoff in a basketball layup, ball contact in a soccer penalty kick). To discover these phases automatically, we compute a scalar kinematic signal $h(t)$ over the motion sequence, derived from interpretable motion statistics such as vertical root velocity, joint acceleration, or jerk magnitude. The peak of this signal defines the skill-critical index

$$t^* = \arg \max_{t \in \{1, \ldots, T\}} h(t).$$

We then center the masked span at t^*, as described earlier. The choice of kinematic signal is skill-specific but requires no paired supervision: for each technique, we select a motion statistic that consistently identifies the execution phase where skill differences are most pronounced, based on visual inspection of expert sequences (see implementation details in Sec. 4). Alternatively, one could automate this statistic choice, e.g., using LLM-proposed body-part signals per drill name, analogous to our LLM-derived operative phrases and extraction offsets (see Supp.). The same criterion is applied during both training and inference, ensuring that the model learns to refine precisely those phases where execution quality is most strongly expressed.

Training and inference. For each technique, we first train a pose tokenizer on trimmed expert motion clips using Eq. (4). We then train MotionInfiller on tokenized expert sequences using the MLM objective in Eq. (5). At inference, we input a novice motion sequence $\mathbf{X}^{\text{nov}}$ to the pose tokenizer E_ϕ to produce a token sequence $\mathbf{k}^{\text{nov}} = [k_1^{\text{nov}}, \ldots, k_T^{\text{nov}}]$. We then identify the skill-critical phase index t^* using the technique-specific kinematic criterion and replace the corresponding centered span with learned `[MASK]` embeddings. MotionInfiller one-shot infills expert motion tokens in the masked region, producing edited token sequence $\hat{\mathbf{k}}$. Finally, D_ϕ reconstructs the edited motion sequence $\hat{\mathbf{X}} = \{(\mathbf{r}_t, \mathbf{o}_t, \hat{\mathbf{p}}_t)\}_{t=1}^{T}$ from $\hat{\mathbf{k}}$, where root translation $\mathbf{r}_t$ and orientation $\mathbf{o}_t$ are copied from the input sequence and joint rotations have been edited (cf. Sec. 3.2).

4 Experiments

Datasets. We evaluate ExpertEdit on eight diverse techniques across three sports—Basketball, Soccer, and Karate—from two datasets, Ego-Exo4D [15] and Kyokushin Karate [52]. We focus on sports datasets that capture wide variation in skill levels within a technique, enabling us to isolate expert executions for training and novice executions for evaluation.

Ego-Exo4D [15] is a large-scale dataset of skilled human activity covering diverse physical activities and skill levels, with timestamped, free-form narrations. Ego-Exo4D offers person-level proficiency score annotations in addition to grounded action narrations. We use "Late Expert" videos for our expert training set and "Novice" actor clips in the pairs for evaluation. We leverage all activities in Ego-Exo4D that exhibit a canonical motion path (see technique criteria, Sec. 3.2): Mikan layup, Reverse layup, Mid-range jumpshot (basketball), and Penalty kick (soccer). We curate a training set of over 24k expert video clips (see Supp.) All motion sequences are extracted from exocentric views.

Kyokushin Karate [52] contains 1411 mocap recordings of 37 subjects performing four karate techniques: Reverse Punch, Spinning Back Kick, Front Kick, and Roundhouse Kick. The dataset contains 3229 single kicks and punches captured at 250 Hz. Each performer has a proficiency grade ranging from 9th kyu (least skilled) to 3rd dan (most skilled). We reclassify each sample as Expert (1st-3rd dan and 1st-3rd kyu) or Novice (6th-9th kyu). The train/val expert set is partitioned at the subject level to prevent the model from exploiting actor-specific motion idiosyncrasies.

4.1 Constructing a paired test set for evaluation

To evaluate expert-like motion editing, we construct a high-quality testbed of temporally-aligned novice-expert pairs across both datasets. This allows us to compute frame-level motion quality metrics. Constructing skill-aligned pairs for evaluation follows recent precedent in prior coaching feedback work: ExpertAF [3] similarly builds novice–expert video pairs (there for training). In contrast, we use manually verified novice–expert pairs only for evaluation.

Fig. 3: Example novice–expert pseudo-pair sequence in our testbed. See text.

For each trimmed novice clip, we retrieve $k = 3$ expert clips within the same technique using similarity computed over available metadata (temporally annotated narrations e.g., "*C shoots a layup into the net with his right hand*" for Ego-Exo4D and motion similarity for Kyokushin Karate). Pairing each novice with multiple expert references mitigates the effect of any single outlier pairing and captures natural variation in expert execution. While Karate techniques follow highly canonical forms such that pairs exhibit natural temporal alignment, basketball and soccer techniques exhibit greater variation in player distance to the ball or basket and starting position. To account for this variation, we temporally align each novice–expert pair using Dynamic Time Warping (DTW) on pose distance and resample the expert sequence to match the novice length. To address chirality asymmetries (e.g., opposite shooting hand), we compute alignment cost for both the original and sagittal-mirrored expert motion and select the lower-cost alignment.

While we use these pairs only for evaluation, when fine-tuning supervised baselines, we separately construct a larger set of pseudo-pairs from the target datasets using the same retrieval procedure as above, but without expert verification or DTW alignment. This ensures baselines have target-domain paired exposure while avoiding expensive and unrealistically clean training samples–dense, human-verified, frame-aligned corrections–that ExpertEdit never receives.

Human validation of evaluation pairs. The pre-processed Karate Kyokushin dataset clips are trimmed at the atomic action level [48], have formally defined proficiency ranks, and exhibit highly canonical forms, allowing us to form high-quality pairs from the existing trimmed clips. For our pseudo-pairs constructed with Ego-Exo4D, we recruit two basketball experts (28 combined years of playing experience) and two soccer experts (23 combined years) to independently assess the quality of the constructed pairs. Each rater evaluates whether (1) the novice and expert clips are temporally aligned, and (2) the expert clip clearly demonstrates higher skill and is correctly identifiable within the pair. A pair is retained if at least one rater judges it valid under both criteria. Using this protocol, 83% of basketball pairs and 77% of soccer pairs are deemed valid. These verified pairs constitute our final evaluation test set, which is now publicly available and constitutes the first firm benchmark for this new task.

This procedure yields high-quality, temporally-aligned novice-expert evaluation pairs for each technique: Reverse layup (499), Penalty kick (173), Jumpshot

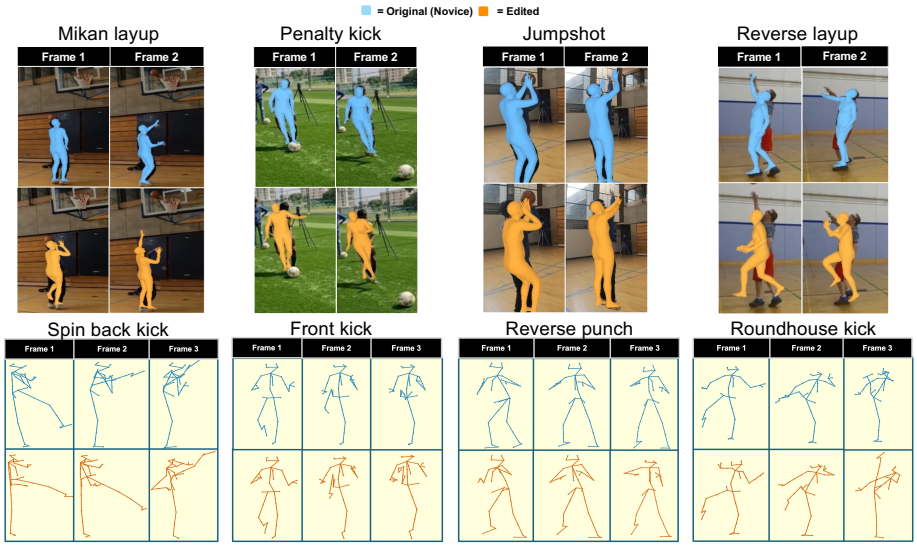

Fig. 4: ExpertEdit sequence visualization: We show novice source pose (blue) and edited pose (orange) at several frames for all techniques. ExpertEdit makes subtle pose refinements that improve form at skill-critical action moments, including raising the knee on the shooting hand-side higher during layups (Mikan, reverse), extending legs further on kicks (spin back, roundhouse), moving the shooting hand under the ball during jumpshots, and improving follow-through from the kicking leg on penalty kicks.

(499), Mikan layup (436), Reverse punch (248), Spinning back kick (224), Front kick (231), and Roundhouse kick (239). Fig. 3 shows one such pair, a mid-range jumpshot. The expert follows the same technique and action-phase progression as the novice, while exhibiting expert-level differences in pose and form.

Metrics. We evaluate whether generated motion exhibits expert-like characteristics using two expert-motion quality metrics. Metrics are computed only on frames within the kinematic-peak defined mask region for all methods. We also provide **human-preference rates** as a quantitative metric of perceived quality.

Pose Improvement (P) measures expert alignment using Procrustes-Aligned Mean Per-Joint Position Error (PA-MPJPE), a standard metric for evaluating 3D human pose accuracy [3, 31, 64, 65, 75]. PA-MPJPE computes the distance between predicted and target 3D joint locations averaged over all joints and frames after scale and rigid alignment. We report relative improvement over the input novice motion:

$$P(\%) = \frac{\text{PA-MPJPE}_{\text{novice}} - \text{PA-MPJPE}_{\text{gen}}}{\text{PA-MPJPE}_{\text{novice}}} \times 100.$$

Higher values indicate stronger movement toward expert-level execution. For each novice motion we generate $m = 3$ edits and compute metrics against the $k = 3$ paired experts to reduce reliance on any single expert trajectory. We report the minimum error across the m generations.

FID Improvement (F) measures whether generated motions align with the expert motion distribution using a Fréchet distance metric analogous to FID [19] used in generative image [18, 20, 28] and pose [37, 44, 45, 54, 71, 72] evaluation. We train technique-level classifiers to distinguish novice from expert motions and use their intermediate feature representations to compute a Fréchet distance between generated motions and expert motions in feature space. As in Pose Improvement, we report relative FID improvement over the source novice distribution. DTW-based warping for frame-level alignment is only used when computing P; F does not require frame-wise novice–expert correspondence.

Baselines. We evaluate ExpertEdit against several state-of-the-art motion editing methods, listed below. These approaches are designed for supervised motion editing and typically require paired demonstrations or explicit editing instructions during training. We fine-tune these baselines on a curated training set of 16k novice–expert pseudo-pairs derived using the same pairing procedure as our evaluation set, but without expensive frame-level alignment or human validation. These pairs constitute *privileged supervision* provided only to the baselines: while ExpertEdit is trained exclusively on unpaired expert demonstrations, the baselines are allowed to learn directly from novice–expert pairs constructed from the target datasets. This gives the baselines training exposure to the target datasets for fair comparison. During both fine-tuning and inference, we use the text prompt "Make the `[TECHNIQUE_NAME]` motion smoother and more controlled". Without access to expert-provided correctional text instructions for each novice sample, this prompt reflects a typical use case for skill-driven motion editing. We arrived at this prompt after evaluating multiple prompts to elicit the best performance from the baselines; see Supp. for a prompt study. We found it produces more stable motion edits than prompts referring explicitly to "expert motion," and more closely matches the style of text edits used in the datasets on which the baselines are pretrained [5, 17].

- **TMED [5]:** A diffusion transformer (DiT)–based motion editing model trained on MotionFix [5]. TMED performs text-conditioned motion editing by denoising motion tokens using cross-attention to the conditioning prompt.
- **SimMotionEdit [32]:** A text-conditioned motion editing model that extends the TMED architecture with an auxiliary temporal guidance module. In addition to generating edited motion, SimMotionEdit predicts a 1D temporal signal indicating where edits should be applied, allowing the model to focus modifications on frames most relevant to the text prompt.
- **FLAME [30]:** A diffusion-based model designed for inference-time motion editing. FLAME generates edited motion conditioned on a source motion sequence and a text instruction. Unlike TMED and SimMotionEdit, which we fine-tune on our pseudo-paired supervision, FLAME does *not* support paired fine-tuning, and is evaluated strictly in its inference-time editing mode.

We fine-tune TMED and SimMotionEdit from MotionFix [5] pretrained checkpoints and evaluate FLAME from a HumanML3D [17] checkpoint for Ego-Exo4D techniques. For the Kyokushin Karate dataset, we evaluate only TMED

Table 1: Results on four basketball and soccer techniques. P and F represent relative improvement in PA-MPJPE and alignment with the expert distribution respectively over the source motion. Higher is better for both (↑).*Indicates no access to paired supervision for training. PA-MPJPE is averaged over $k = 3$ expert reference pairs to account for natural variation in expert behavior. See Supp. for absolute metrics.

	Basketball and Soccer							
	Mikan		**Rev. Layup**		**Jumpshot**		**Pen. Kick**	
Method	$P(\uparrow)$	$F(\uparrow)$	$P(\uparrow)$	$F(\uparrow)$	$P(\uparrow)$	$F(\uparrow)$	$P(\uparrow)$	$F(\uparrow)$
SimMotionEdit [32]	2.26	1.28	2.31	4.88	3.95	1.24	3.20	2.38
TMED [5]	1.87	1.00	2.20	4.30	2.58	2.03	2.57	1.95
FLAME* [30]	2.35	1.45	2.09	6.13	3.13	1.54	2.90	2.22
ExpertEdit* (Ours)	**6.18**	**5.72**	**5.87**	**12.08**	**5.34**	**7.66**	**6.03**	**9.14**

and SimMotionEdit. FLAME is not included in this setting because its pretrained model operates on SMPL-based pose representations, whereas the karate dataset uses a custom MoCap joint parameterization. Applying FLAME would therefore require retraining the model from scratch in a new representation, which falls outside the inference-time editing paradigm for which the method was designed.

Implementation details. We extract SMPL [36] parameters from Ego-Exo4D videos using WHAM [49] at 30 fps. For Kyokushin karate, we use the provided axis-angle joint rotations ($J = 39$). For Ego-Exo4D techniques, we train pose tokenizers on fixed-length clips of 120 frames (reverse layup, penalty kick) and 90 frames (mikan layup, jumpshot). Karate sequences are uniformly resampled to 64 frames to normalize action duration while preserving relative motion timing. Our pose tokenizer follows the Transformer VQ-VAE architecture of [37]. MotionInfiller uses a bidirectional Transformer trained with span masking with maximum span fraction $\alpha = 0.3$. See Supp. for an ablation on α. Kinematic peak selection uses vertical velocity for basketball, maximum jerk for soccer, postural extremeness for reverse punch, and foot acceleration for karate kicks. Additional architecture, preprocessing, and training details are provided in Supp.

4.2 Results

Basketball and soccer techniques. Results on basketball and soccer are shown in Table 1. Despite lacking access to privileged supervision provided to SimMotionEdit [32] and TMED [5], ExpertEdit outperforms state-of-the-art motion editing baselines across all four techniques, achieving roughly 2–4× larger gains in pose improvement (P) and expert FID score improvement (F) compared to the strongest baselines. The largest gains occur for reverse layups and penalty kicks, where ExpertEdit achieves $F = 12.08\%$ and 9.14% respectively, compared to at most 6.13% and 2.38% for competing approaches.

Fig. 4 (top row) illustrates these corrections on sample edited frames. On the reverse layup (right), the edited motion lifts the shooting side knee higher

Table 2: Results on Kyokushin Karate Dataset, four techniques. P denotes relative improvement in PA-MPJPE. F denotes relative improvement in FID over novice motion. Higher is better for both (↑).

	Kyokushin Karate							
	Rev. Punch		Spin BK		Front		High RH	
Method	P(↑)	F(↑)	P(↑)	F(↑)	P(↑)	F(↑)	P(↑)	F(↑)
SimMotionEdit [32]	**2.20**	1.30	1.43	3.30	1.45	5.05	**3.30**	5.88
TMED [5]	0.92	1.08	0.86	2.25	0.38	2.90	2.25	3.37
ExpertEdit (Ours)	2.07	**1.36**	**1.79**	**4.23**	**1.88**	**9.73**	2.96	**6.18**

during the jump and release phase, while in penalty kicks (middle left) the model exaggerates follow through on the striking leg after ball contact. These localized corrections lead to motions that better match both expert pose trajectories and the overall expert motion distribution.

These gains reflect the fact that skill differences often manifest in short, temporally localized moments of an action. While SimMotionEdit predicts a temporal signal to identify frames requiring stronger edits, it relies on descriptive text prompts to determine this signal and how the motion should change, which can suffer under the general text edit prompt required in this setting. In contrast, ExpertEdit targets skill-critical motion phases through kinematic masking identified from the source motion itself, allowing the model to focus edits on the moments where expert execution differs most strongly.

Kyokushin karate techniques. Results on the Kyokushin Karate techniques are shown in Table 2. ExpertEdit consistently improves alignment with the expert motion distribution, achieving the largest gain in FID Improvement (F) across all techniques. While SimMotionEdit achieves slightly higher Pose Improvement (P) on two techniques (Rev. Punch and High Roundhouse), these gains do not translate into stronger expert distribution alignment: ExpertEdit still achieves higher F scores on both techniques. This suggests that while supervised models may produce edits that reduce pose error relative to individual expert examples, they do not necessarily generate motions that better match the overall expert motion distribution. While the absolute scale of improvements is smaller than in the basketball and soccer scenarios, this reflects characteristics of the dataset: karate techniques follow highly canonical motion patterns, and our expert pool includes both dan practitioners and upper-level kyu ranks (1st–3rd kyu) to ensure sufficient training data. This produces a narrower skill gap between actors and a less sharply defined expert motion manifold.

Fig. 4 (bottom row) illustrates our results on karate techniques. Across multiple kick techniques, the edited motion (orange) consistently produces better leg extension than the source clip. These corrections correspond to the large improvements in expert distribution alignment observed for these techniques. See Supp. video for examples with motion.

Human preference study. Finally, to evaluate whether generated edits are perceived as useful visual feedback, we conduct a blinded A/B preference study

Table 3: Blinded human preference study. Participants compare anonymized ExpertEdit and SimMotionEdit (strongest baseline) rendered edits and choose which would be more useful for improving form.

Preferred edit	Basketball (78)	Soccer (34)	Overall (112)
ExpertEdit	**42.3%**	**47.1%**	**43.8%**
SimMotionEdit [32]	26.9%	23.5%	25.9%
Neither	30.8%	29.4%	30.4%
ExpertEdit among non-neutral	**61.1%**	**66.7%**	**62.8%**

on rendered edits. We focus this study on the basketball and soccer domains, where we could recruit target users with relevant playing experience to judge form improvements. For each trial, participants view the source motion and two anonymized edited motions from ExpertEdit and a strong fine-tuned baseline, SimMotionEdit [32], with method order randomized. We ask: "Which edited motion would be more useful to improve your form if you were the performer in the video?" Participants choose from "A", "B", or "Neither".

Table 3 summarizes the results. Across 24 raters and 112 total pairwise judgments, ExpertEdit is preferred in 49 judgments versus 29 for SimMotionEdit, with 34 "Neither" responses. Among non-neutral judgments, ExpertEdit is preferred 62.8% of the time, suggesting that our localized and skill-driven edits are perceived as more useful corrective feedback than the compared text-driven motion editor. These results augment the metrics reported above by directly measuring perceived feedback usefulness.

5 Conclusion

We introduce **ExpertEdit**, a framework for skill-driven motion editing that refines novice demonstrations into expert-like executions by projecting skill-critical phases in a novice motion onto a learned manifold of expert motion. Unlike prior approaches that rely on text guidance, reference clips, or heavy paired supervision, ExpertEdit learns an expert motion prior directly from unpaired expert performances and applies it to novice inputs at inference without additional conditioning. Across diverse techniques spanning basketball, soccer, and karate from two datasets, ExpertEdit outperforms state-of-the-art motion editing methods on multiple expert-quality metrics, including supervised approaches that rely on privileged paired demonstrations. These results establish skill-driven motion refinement as a previously unmet capability in motion editing. By forgoing expensive skill-annotated motion pairs, ExpertEdit unlocks the potential to tackle diverse motion domains simply by watching how experts perform.

Acknowledgements

This research is supported in part by the UT Austin IFML NSF AI Institute, Lyda Hill Philanthropies, and a gift from Amazon. We thank Ashutosh Kumar and Joungbin An for helpful discussions.

References

1. Aberman, K., Weng, Y., Lischinski, D., Cohen-Or, D., Chen, B.: Unpaired motion style transfer from video to animation. ACM Transactions on Graphics **39**(4) (Aug 2020). https://doi.org/10.1145/3386569.3392469
2. Ashutosh, K., Grauman, K.: Learning skill-attributes for transferable assessment in video (2025), arXiv preprint arXiv:2511.13993
3. Ashutosh, K., Nagarajan, T., Pavlakos, G., Kitani, K., Grauman, K.: Expertaf: Expert actionable feedback from video (2025), arXiv preprint arXiv:2408.00672
4. Ashutosh, K., Wu, C.H., Grauman, K.: Sportskills: Physical skill learning from sports instructional videos (2026), arXiv preprint arXiv:2603.25163
5. Athanasiou, N., Cseke, A., Diomataris, M., Black, M.J., Varol, G.: Motionfix: Text-driven 3d human motion editing (2024), arXiv preprint arXiv:2408.00712
6. Chan, C., Ginosar, S., Zhou, T., Efros, A.A.: Everybody dance now (2019), arXiv preprint arXiv:1808.07371
7. Cheng, L., Xie, X., Peng, Y., Feng, M., He, Y., Cao, A., Wu, Y., Zhang, H., Wu, Y.: Vismimic: Integrating motion chain in feedback video generation for motor coaching. In: Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology. UIST '25, Association for Computing Machinery, New York, NY, USA (2025). https://doi.org/10.1145/3746059.3747794
8. Delmas, G., Weinzaepfel, P., Moreno-Noguer, F., Rogez, G.: Posefix: Correcting 3d human poses with natural language (2024), arXiv preprint arXiv:2309.08480
9. Ding, Y., Zhang, S., Shenglan, L., Zhang, J., Chen, W., Haifei, D., dong, b., Sun, T.: 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. p. 1564–1572. MM '24, Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3664647.3681084
10. Dittakavi, B., Bavikadi, D., Desai, S.V., Chakraborty, S., Reddy, N., Balasubramanian, V.N., Callepalli, B., Sharma, A.: Pose tutor: An explainable system for pose correction in the wild. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 3539–3548 (2022). https://doi.org/10.1109/CVPRW56347.2022.00398
11. Doughty, H., Damen, D., Mayol-Cuevas, W.: Who's better? who's best? pairwise deep ranking for skill determination (2018), arXiv preprint arXiv:1703.09913
12. Du, H., Herrmann, E., Sprenger, J., Fischer, K., Slusallek, P.: Stylistic locomotion modeling and synthesis using variational generative models. In: Proceedings of the 12th ACM SIGGRAPH Conference on Motion, Interaction and Games. MIG '19, Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3359566.3360083
13. Fieraru, M., Zanfir, M., Pirlea, S.C., Olaru, V., Sminchisescu, C.: Aifit: Automatic 3d human-interpretable feedback models for fitness training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9919–9928 (June 2021)
14. Goel, P., Wang, K.C., Liu, C.K., Fatahalian, K.: Iterative motion editing with natural language. In: Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers. p. 1–9. SIGGRAPH '24, ACM (Jul 2024). https://doi.org/10.1145/3641519.3657447
15. Grauman, K., Westbury, A., Torresani, L., Kitani, K., Malik, J., Afouras, T., Ashutosh, K., Baiyya, V., Bansal, S., Boote, B., Byrne, E., Chavis, Z., Chen, J.,

Cheng, F., Chu, F.J., Crane, S., Dasgupta, A., Dong, J., Escobar, M., Forigua, C., Gebreselasie, A., Haresh, S., Huang, J., Islam, M.M., Jain, S., Khirodkar, R., Kukreja, D., Liang, K.J., Liu, J.W., Majumder, S., Mao, Y., Martin, M., Mavroudi, E., Nagarajan, T., Ragusa, F., Ramakrishnan, S.K., Seminara, L., Somayazulu, A., Song, Y., Su, S., Xue, Z., Zhang, E., Zhang, J., Castillo, A., Chen, C., Fu, X., Furuta, R., Gonzalez, C., Gupta, P., Hu, J., Huang, Y., Huang, Y., Khoo, W., Kumar, A., Kuo, R., Lakhavani, S., Liu, M., Luo, M., Luo, Z., Meredith, B., Miller, A., Oguntola, O., Pan, X., Peng, P., Pramanick, S., Ramazanova, M., Ryan, F., Shan, W., Somasundaram, K., Song, C., Southerland, A., Tateno, M., Wang, H., Wang, Y., Yagi, T., Yan, M., Yang, X., Yu, Z., Zha, S.C., Zhao, C., Zhao, Z., Zhu, Z., Zhuo, J., Arbelaez, P., Bertasius, G., Crandall, D., Damen, D., Engel, J., Farinella, G.M., Furnari, A., Ghanem, B., Hoffman, J., Jawahar, C.V., Newcombe, R., Park, H.S., Rehg, J.M., Sato, Y., Savva, M., Shi, J., Shou, M.Z., Wray, M.: Ego-exo4d: Understanding skilled human activity from first- and third-person perspectives (2024), arXiv preprint arXiv:2311.18259

16. Guo, C., Mu, Y., Javed, M.G., Wang, S., Cheng, L.: Momask: Generative masked modeling of 3d human motions (2023), arXiv preprint arXiv:2312.00063
17. Guo, C., Zou, S., Zuo, X., Wang, S., Ji, W., Li, X., Cheng, L.: Generating diverse and natural 3d human motions from text. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5152–5161 (June 2022)
18. Hartwig, S., Engel, D., Sick, L., Kniesel, H., Payer, T., Poonam, P., Glöckler, M., Bäuerle, A., Ropinski, T.: A survey on quality metrics for text-to-image generation (2025), arXiv preprint arXiv:2403.11821
19. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium (2018), arXiv preprint arXiv:1706.08500
20. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models (2020), arXiv preprint arXiv:2006.11239
21. Holden, D., Saito, J., Komura, T., Joyce, T.: Learning motion manifolds with convolutional autoencoders. In: SIGGRAPH Asia 2015 Technical Briefs. SA '15, Association for Computing Machinery, New York, NY, USA (2015). https://doi.org/10.1145/2820903.2820918
22. Hu, L., Zhang, Z., Ye, Y., Xu, Y., Xia, S.: Diffusion-based human motion style transfer with semantic guidance (2024), arXiv preprint arXiv:2405.06646
23. Hu, L., Gao, X., Zhang, P., Sun, K., Zhang, B., Bo, L.: Animate anyone: Consistent and controllable image-to-video synthesis for character animation (2024), arXiv preprint arXiv:2311.17117
24. Huh, M., Xue, Z., Das, U., Ashutosh, K., Grauman, K., Pavel, A.: Vid2coach: Transforming how-to videos into task assistants (2025), arXiv preprint arXiv:2506.00717
25. Jang, D.K., Park, S., Lee, S.H.: Motion puzzle: Arbitrary motion style transfer by body part. ACM Transactions on Graphics **41**(3), 1–16 (Jun 2022). https://doi.org/10.1145/3516429
26. Jiang, B., Chen, X., Liu, W., Yu, J., Yu, G., Chen, T.: Motiongpt: Human motion as a foreign language (2023), arXiv preprint arXiv:2306.14795
27. Jiang, N., Li, H., Yuan, Z., He, Z., Chen, Y., Liu, T., Zhu, Y., Huang, S.: Dynamic motion blending for versatile motion editing (2025), arXiv preprint arXiv:2503.20724
28. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation (2018), arXiv preprint arXiv:1710.10196
29. Kim, B., Kim, J., Chang, H.J., Choi, J.Y.: Most: Motion style transformer between diverse action contents. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1705–1714 (June 2024)

30. Kim, J., Kim, J., Choi, S.: Flame: Free-form language-based motion synthesis & editing (2023), arXiv preprint arXiv:2209.00349
31. Li, J., Cao, J., Zhang, H., Rempe, D., Kautz, J., Iqbal, U., Yuan, Y.: Genmo: A generalist model for human motion (2025), arXiv preprint arXiv:2505.01425
32. Li, Z., Cheng, K., Ghosh, A., Bhattacharya, U., Gui, L., Bera, A.: Simmotionedit: Text-based human motion editing with motion similarity prediction (2025), arXiv preprint arXiv:2503.18211
33. Liu, J., Saquib, N., Zhutian, C., Kazi, R.H., Wei, L.Y., Fu, H., Tai, C.L.: Posecoach: A customizable analysis and visualization system for video-based running coaching. IEEE Transactions on Visualization and Computer Graphics **30**(7), 3180–3195 (Jul 2024). https://doi.org/10.1109/tvcg.2022.3230855
34. Liu, S.L., Ding, Y.N., Yan, G., Zhang, S.F., Zhang, J.R., Chen, W.Y., Xu, X.H.: Fine-grained action analysis: A multi-modality and multi-task dataset of figure skating (2024), arXiv preprint arXiv:2307.02730
35. Liu, W., Piao, Z., Min, J., Luo, W., Ma, L., Gao, S.: Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis (2019), arXiv preprint arXiv:1909.12224
36. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. ACM Trans. Graphics (Proc. SIGGRAPH Asia) **34**(6), 248:1–248:16 (Oct 2015)
37. Lucas, T., Baradel, F., Weinzaepfel, P., Rogez, G.: Posegpt: Quantization-based 3d human motion generation and forecasting (2022), arXiv preprint arXiv:2210.10542
38. Majeedi, A., Gajjala, V.R., GNVV, S.S.S.N., Li, Y.: Rica2: Rubric-informed, calibrated assessment of actions (2024), arXiv preprint arXiv:2408.02138
39. Noworolnik, F., Jaworek-Korjakowska, J.: Assessing the quality of soccer shots from single-camera video with vision-language models and motion features. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops. pp. 2733–2740 (October 2025)
40. Pan, Y., Zhang, C., Bertasius, G.: Basket: A large-scale video dataset for fine-grained skill estimation (2025), arXiv preprint arXiv:2503.20781
41. Parmar, P., Gharat, A., Rhodin, H.: Domain knowledge-informed self-supervised representations for workout form assessment. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVIII. pp. 105–123. Springer (2022)
42. Parmar, P., Morris, B.T.: Learning to score olympic events (2017), arXiv preprint arXiv:1611.05125
43. Parmar, P., Morris, B.T.: What and how well you performed? a multitask learning approach to action quality assessment (2019), arXiv preprint arXiv:1904.04346
44. Petrovich, M., Black, M.J., Varol, G.: Action-conditioned 3d human motion synthesis with transformer vae (2021), arXiv preprint arXiv:2104.05670
45. Petrovich, M., Black, M.J., Varol, G.: Temos: Generating diverse human motions from textual descriptions (2022), arXiv preprint arXiv:2204.14109
46. Ram, N., McCullagh, P.: Self-modeling: Influence on psychological responses and physical performance. Sport Psychologist **17**, 220–241 (06 2003). https://doi.org/10.1123/tsp.17.2.220
47. Rao, J., Wu, H., Jiang, H., Zhang, Y., Wang, Y., Xie, W.: Towards universal soccer video understanding (2025), arXiv preprint arXiv:2412.01820
48. Richardson, A., Putze, F.: Motion diffusion autoencoders: Enabling attribute manipulation in human motion demonstrated on karate techniques. In: Proceedings of the 27th International Conference on Multimodal Interaction. p. 372–380. ICMI '25, ACM (Oct 2025). https://doi.org/10.1145/3716553.3750773

49. Shin, S., Kim, J., Halilaj, E., Black, M.J.: Wham: Reconstructing world-grounded humans with accurate 3d motion (2024), arXiv preprint arXiv:2312.07531
50. Ste-Marie, D.M., Vertes, K., Rymal, A.M., Martini, R.: Feedforward self-modeling enhances skill acquisition in children learning trampoline skills. Frontiers in Psychology **2**, 155 (2011). https://doi.org/10.3389/fpsyg.2011.00155
51. Steel, K.A., Mudie, K., Sandoval, R., Anderson, D., Dogramaci, S., Rehmanjan, M., Birznieks, I.: Can video self-modeling improve affected limb reach and grasp ability in stroke patients? Journal of Motor Behavior **50**(2), 117–126 (2018). https://doi.org/10.1080/00222895.2017.1306480, epub 2017 May 19
52. Szczęsna, A., Błaszczyszyn, M., Pawlyta, M.: Optical motion capture dataset of selected techniques in beginner and advanced kyokushin karate athletes. Scientific Data **8** (01 2021). https://doi.org/10.1038/s41597-021-00801-5
53. Tao, T., Zhan, X., Chen, Z., van de Panne, M.: Style-erd: Responsive and coherent online motion style transfer (2022), arXiv preprint arXiv:2203.02574
54. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-Or, D., Bermano, A.H.: Human motion diffusion model (2022), arXiv preprint arXiv:2209.14916
55. Tu, S., Dai, Q., Cheng, Z.Q., Hu, H., Han, X., Wu, Z., Jiang, Y.G.: Motioneditor: Editing video motion via content-aware diffusion (2023), arXiv preprint arXiv:2311.18830
56. Tu, S., Dai, Q., Zhang, Z., Xie, S., Cheng, Z.Q., Luo, C., Han, X., Wu, Z., Jiang, Y.G.: Motionfollower: Editing video motion via lightweight score-guided diffusion (2024), arXiv preprint arXiv:2405.20325
57. Villegas, R., Yang, J., Ceylan, D., Lee, H.: Neural kinematic networks for unsupervised motion retargetting (2018), arXiv preprint arXiv:1804.05653
58. Wang, Y., Huang, D., Zhang, Y., Ouyang, W., Jiao, J., Feng, X., Zhou, Y., Wan, P., Tang, S., Xu, D.: Motiongpt-2: A general-purpose motion-language model for motion generation and understanding (2024), arXiv preprint arXiv:2410.21747
59. Wu, C.H., Ashutosh, K., Grauman, K.: Skillsight: Efficient first-person skill assessment with gaze (2026), arXiv preprint arXiv:2511.19629
60. Xu, H., Ke, X., Li, Y., Xu, R., Wu, H., Lin, X., Guo, W.: Vision-language action knowledge learning for semantic-aware action quality assessment. In: Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XLII. p. 423–440. Springer-Verlag, Berlin, Heidelberg (2024). https://doi.org/10.1007/978-3-031-72946-1_24
61. Xu, J., Rao, Y., Yu, X., Chen, G., Zhou, J., Lu, J.: Finediving: A fine-grained dataset for procedure-aware action quality assessment (2022), arXiv preprint arXiv:2204.03646
62. Xu, J., Yin, S., Zhao, G., Wang, Z., Peng, Y.: Fineparser: A fine-grained spatio-temporal action parser for human-centric action quality assessment (2024), arXiv preprint arXiv:2405.06887
63. Xu, Z., Zhang, J., Liew, J.H., Yan, H., Liu, J.W., Zhang, C., Feng, J., Shou, M.Z.: Magicanimate: Temporally consistent human image animation using diffusion model (2023), arXiv preprint arXiv:2311.16498
64. Yang, C., Song, H., Choi, S., Lee, S., Kim, J., Do, H.: Posesyn: Synthesizing diverse 3d pose data from in-the-wild 2d data (2025), arXiv preprint arXiv:2503.13025
65. Yang, C., Tkach, A., Hampali, S., Zhang, L., Crowley, E.J., Keskin, C.: Egoposeformer: A simple baseline for stereo egocentric 3d human pose estimation (2024), arXiv preprint arXiv:2403.18080
66. Yang, Z., Zhu, W., Wu, W., Qian, C., Zhou, Q., Zhou, B., Loy, C.C.: Transmomo: Invariance-driven unsupervised video motion retargeting (2020), arXiv preprint arXiv:2003.14401

67. Yeh, W.H., Su, Y.A., Chen, C.N., Lin, Y.H., Ku, C., Chiu, W., Hu, M.C., Ku, L.W.: Coachme: Decoding sport elements with a reference-based coaching instruction generation model. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). p. 29126–29151. Association for Computational Linguistics (2025). https://doi.org/10.18653/v1/2025.acl-long.1413
68. Yeung, C., Ide, K., Fujii, K.: Autosoccerpose: Automated 3d posture analysis of soccer shot movements (2024), arXiv preprint arXiv:2405.12070
69. Yi, H., Pan, Y., He, F., Liu, X., Zhang, B., Oguntola, O., Bertasius, G.: Exact: A video-language benchmark for expert action analysis (2025), arXiv preprint arXiv:2506.06277
70. Yin, H., Gu, L., Parmar, P., Xu, L., Guo, T., Fu, W., Zhang, Y., Zheng, T.: Flex: A large-scale multi-modal multi-action dataset for fitness action quality assessment (2025), arXiv preprint arXiv:2506.03198
71. Zhang, J., Zhang, Y., Cun, X., Huang, S., Zhang, Y., Zhao, H., Lu, H., Shen, X.: T2m-gpt: Generating human motion from textual descriptions with discrete representations (2023), arXiv preprint arXiv:2301.06052
72. Zhang, M., Cai, Z., Pan, L., Hong, F., Guo, X., Yang, L., Liu, Z.: Motiondiffuse: Text-driven human motion generation with diffusion model (2022), arXiv preprint arXiv:2208.15001
73. Zhang, M., Li, H., Cai, Z., Ren, J., Yang, L., Liu, Z.: Finemogen: Fine-grained spatio-temporal motion generation and editing (2023), arXiv preprint arXiv:2312.15004
74. Zhang, S., Bai, S., Chen, G., Chen, L., Lu, J., Wang, J., Tang, Y.: Narrative action evaluation with prompt-guided multimodal interaction (2024), arXiv preprint arXiv:2404.14471
75. Zhao, S., Wang, Z., Luan, T., Jia, J., Zhu, W., Luo, J., Yuan, J., Xi, N.: Pp-motion: Physical-perceptual fidelity evaluation for human motion generation (2026), arXiv preprint arXiv:2508.08179
76. Zhong, X., Huang, X., Yang, X., Lin, G., Wu, Q.: Deco: Decoupled human-centered diffusion video editing with motion consistency (2024), arXiv preprint arXiv:2408.07481
77. Zhu, S., Chen, J.L., Dai, Z., Su, Q., Xu, Y., Cao, X., Yao, Y., Zhu, H., Zhu, S.: Champ: Controllable and consistent human image animation with 3d parametric guidance (2024), arXiv preprint arXiv:2403.14781