informatics

MDPI

Article

Recognition-Guided Generative Skeleton Imputation for Robust Action Recognition Under Severe Occlusion in Construction Scenarios

Hechen Yun [1,*], Nobuhiko Kato [2], Ken Igarashi [2], Ken Kawamoto [2] and Yoichi Kageyama [1]

[1] Faculty of Informatics and Data Science, Akita University, Akita 010-8502, Japan; kageyama@ie.akita-u.ac.jp
[2] ADK Fuji System Co., Ltd., Akita 010-0851, Japan; nobuhiko@adf.co.jp (N.K.); igarashi@adf.co.jp (K.I.); kawamoto@adf.co.jp (K.K.)
* Correspondence: yun_hc@gipc.akita-u.ac.jp

Abstract

Skeleton-based human action recognition can support digital construction and renovation monitoring by converting worker motions into structured information for safety and process management. However, workers are often occluded by tools, materials, furniture, and temporary structures, resulting in incomplete skeleton sequences and reduced recognition accuracy. In this study, we propose a recognition-guided generative skeleton imputation approach for robust action recognition under severe occlusion. Our method first evaluates an incomplete skeleton and applies imputation only to samples with uncertain recognition results. For these samples, a pretrained OmniControl model generates multiple plausible motion candidates from the observed skeleton information and action descriptions. The candidate most consistent with the observed motion is selected and used only to fill missing regions while preserving reliable joints, and the completed skeleton is then re-evaluated by the same recognition model. Experiments on 565 collected indoor renovation clips and the NW-UCLA dataset, using four recognition models and nine occlusion conditions, demonstrated overall improvements in recognition robustness, with average accuracy increasing from 61.01% to 65.00% on the collected dataset and from 42.63% to 50.31% on NW-UCLA. These results demonstrate the potential of selective generative imputation for improving action recognition under incomplete skeletal observations.

Keywords: skeleton-based action recognition; skeleton imputation; occlusion; motion generation; OmniControl; construction monitoring

Check for updates

Academic Editors: Anuchit Jitpattanakul and Sakorn Mekruksavanich

Received: 18 June 2026
Revised: 20 August 2026
Accepted: 29 August 2026
Published: 3 September 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Human action recognition (HAR) has been extensively studied in recent years as a key technology for digital systems that interpret human behavior in healthcare, surveillance, industrial monitoring, and human–machine interaction. In Japan, the rapid aging of the population and the persistent shortage of skilled workers have further accelerated the need for artificial intelligence systems that can reduce human error, improve operational efficiency, and support safer working environments [1]. Construction-related industries are representative domains facing these challenges, as complex work processes require continuous supervision under safety-critical conditions. Vision-based monitoring systems have therefore attracted increasing attention for worker tracking, activity recognition, unsafe-posture detection, and safety-related behavior analysis [2–5]. Within such systems,

HAR provides structured information about worker behavior that can support both safety management and construction-process monitoring.

Among different HAR modalities, skeleton-based methods represent human motion using compact and structured joint trajectories rather than raw visual appearance [6–9]. These representations are less sensitive to background clutter, illumination changes, and appearance variations and can also provide greater privacy than RGB-based representations. Recent studies have extended skeleton-based HAR from controlled laboratory datasets toward more realistic and unconstrained environments, with increasing attention to noise, viewpoint variation, and domain shifts [10–12]. In construction applications, skeletal information is particularly relevant because many work activities and safety-related behaviors are characterized by body posture and joint movement patterns [13–15]. These studies demonstrate the potential of skeleton-based HAR for worker activity analysis and for understanding safety-related behavior.

Despite these advantages, deploying skeleton-based HAR in real-world construction scenarios remains highly challenging. Unlike benchmark datasets collected under controlled conditions, work sites contain changing layouts, temporary structures, equipment, materials, and constrained camera viewpoints. Workers may therefore experience partial occlusion or prolonged loss of critical body regions during actions. Such conditions produce incomplete or corrupted skeleton sequences in which joints can be missing, inaccurately localized, or temporally inconsistent. Because most skeleton-based recognition models rely on sufficiently complete and temporally coherent observations, the persistent loss of discriminative joints can substantially degrade recognition performance, particularly for fine-grained and safety-critical activities [16–18].

Existing approaches to recognition from incomplete skeletons generally follow two directions. The first improves the occlusion robustness of the recognizer itself through occlusion-aware architectures, data augmentation, or training with incomplete skeletons [18–20]. Although these approaches can improve recognition under missing observations, their effectiveness commonly depends on whether the training data adequately represent the occlusion patterns encountered during deployment [21]. They may also require model-specific retraining when a new recognition backbone or application environment is introduced. This requirement is restrictive in construction environments, where actions within the same category can vary with task phase, tool use, working posture, working height, and camera viewpoint. Collecting representative occluded samples across these variations can therefore become costly and difficult when HAR systems are adapted to different work sites.

The second direction attempts to recover or impute missing skeletal information before recognition. Conventional approaches estimate missing joints from spatial relationships among body parts or temporal continuity across frames [22,23]. Recent methods further combine spatial and temporal information, such as transformer-based architectures, graph-based temporal modeling, masked motion completion, or generative reconstruction frameworks [24–28]. Although these methods can improve the geometric consistency and temporal plausibility of incomplete skeleton sequences, reconstruction becomes more ambiguous when critical joints remain unobserved for extended periods, and only limited spatial or temporal context is available [29,30]. More importantly, while most completion methods are primarily optimized for reconstruction accuracy or motion plausibility, a physically plausible completion does not necessarily restore the action-discriminative information required by a downstream HAR model.

These limitations reveal a practical gap between occlusion-robust recognition and skeleton completion. Adapting an existing HAR system to severe skeletal missingness often requires either retraining the recognizer with representative occluded data or intro-

ducing a separately trained completion model. In contrast, a pretrained controllable motion generator can provide an external motion prior that is introduced at inference time to infer plausible motion beyond the locally observed skeleton context. If this capability is selectively activated according to recognition uncertainty, it can extend an existing HAR pipeline with missing-motion inference without requiring additional occlusion-specific retraining of the recognition backbone [31]. Based on this concept, we investigate a recognition-guided generative skeleton imputation framework for construction-related action recognition under severe occlusion.

To implement this strategy, the proposed framework first evaluates an incomplete skeleton using the target HAR model and activates generative imputation only when the prediction is considered unreliable. A fixed pretrained OmniControl model [32] then generates multiple motion candidates from the partial skeleton observations and action-level semantic prompts. The candidates are ranked according to their consistency with the observed skeletal motion, and the selected sequence is used only to replace missing regions while reliable observations are preserved. The completed skeleton is subsequently re-evaluated by the same HAR model to obtain the final recognition result. The main contributions of this work are summarized as follows:

1. An inference-time robustness enhancement framework is introduced for skeleton-based HAR, enabling severe missing-skeleton conditions to be addressed without additional occlusion-specific retraining of the target recognition model.
2. A model-adaptive selective generation mechanism is developed to activate generative imputation according to recognition confidence and ambiguity, thereby focusing generation on samples that are difficult for the original HAR model.
3. A multi-candidate generation and selection strategy is combined with missing-region-only integration, allowing plausible motion hypotheses to be evaluated while preserving reliable skeletal observations.
4. The framework is evaluated on the collected dataset and NW-UCLA, including four graph-based recognition models, nine occlusion conditions, trigger and sampling-step ablations, computational-cost measurements, and comparison with a learned skeleton-completion baseline.

The remainder of this paper is organized as follows: Section 2 reviews related work and clarifies the research gap; Section 3 outlines the proposed methodology; Section 4 presents our experimental results and discussion; and Section 5 summarizes our conclusions and future research directions.

2. Related Work

2.1. Skeleton-Based Human Action Recognition

Skeleton-based HAR represents human motion using compact and structured joint sequences and has been extensively studied through graph-based modeling, alternative pose representations, and discriminative representation learning. Graph-based methods remain a dominant direction because they can explicitly model spatial and temporal dependencies among body joints. CTR-GCN [33] introduces channel-wise topology refinement to learn adaptive joint relationships, ST-GCN++ [34] establishes a stronger and more efficient graph-convolutional baseline, and HD-GCN [35] further models local and distant joint dependencies through hierarchical graph decomposition. Beyond coordinate-based graph representations, PoseC3D [36] converts skeleton sequences into 3D heatmap volumes for convolutional recognition, whereas InfoGCN [37] learns informative latent representations of skeletal motion. While these approaches substantially improve recognition from observed skeleton sequences, they do not explicitly recover motion information that is absent

from the input. When action-discriminative joints remain unobserved for extended periods, improved feature extraction alone may therefore be insufficient for reliable recognition.

2.2. Vision-Based Construction Site Monitoring and Worker Action Recognition

Vision-based construction monitoring has increasingly incorporated worker activity recognition to support safety management, process assessment, and productivity analysis. Construction 5.0 research has identified worker action understanding as an important component of intelligent construction systems that connect visual perception with operational and safety information [38]. To address the domain gap between general HAR benchmarks and construction activities, Tian et al. [39] developed a construction motion dataset and subsequently introduced a multi-stream skeleton-based recognition model using body-segment representations [40]. Other studies have investigated unsafe-behavior detection through spatio-temporal video analysis [41] and worker tracking under complex construction site conditions [42]. Real-world fixed-camera monitoring has also been demonstrated in other safety-related environments, such as highway vehicle monitoring [43], confirming the practical feasibility of visual sensing under field conditions. Although these studies demonstrate the practical value of vision-based monitoring, they primarily address activity modeling, behavior detection, or tracking when sufficient visual evidence remains available. The recovery and recognition of worker actions from severely incomplete skeletal observations remain comparatively underexplored.

2.3. Occlusion-Aware Skeleton-Based Action Recognition and Skeleton Imputation

The existing approaches to incomplete-skeleton HAR can be broadly divided into robustness-oriented recognition and reconstruction-oriented imputation. Robustness-oriented methods improve the recognizer itself through occlusion-aware training, self-supervised representation learning, or part-aware modeling. Chen et al. [44] introduced self-supervised learning with imputation-assisted training to improve representations under skeletal occlusion. Part-aware and multi-stream graph models [45] exploit information from visible body regions when specific joints are unavailable. Reconstruction-oriented methods instead recover missing skeletal observations before recognition. Lee et al. [23] employed a denoising graph autoencoder that exploits skeletal graph structure to estimate missing joints, while Vernikos et al. [26] used adversarial generation for skeleton reconstruction under occlusion. More recent masked and generative completion approaches further indicate the potential of learned motion priors for recovering incomplete human motion sequences [27–30]. Temporal sequence modeling has also been studied extensively in sensor-based HAR. Tehrani et al. [46] employed Bi-LSTM to capture bidirectional temporal dependencies in wearable-sensor sequences, while Kia et al. [47] combined bidirectional generative modeling with frequency-domain analysis for body-worn sensor HAR. Although these studies operate on sensor signals rather than skeletal coordinates, they demonstrate the importance of temporal context and generative sequence modeling for robust human activity understanding.

Despite these advances, these two directions retain different limitations: robustness-oriented methods commonly require retraining with representative missing-data patterns, whereas reconstruction-oriented methods primarily optimize coordinate recovery, motion consistency, or reconstruction plausibility. Their performance can also become less reliable when prolonged or structured missingness leaves insufficient contextual information for recovery. More importantly, a plausible reconstructed motion is not necessarily the motion representation that best supports a fixed downstream recognizer. This leaves an opportunity for an inference-time framework that combines the robustness of an exist-

ing HAR model with an external pretrained motion prior, without requiring additional occlusion-specific retraining.

2.4. Motion Generation

Human motion generation has advanced rapidly through text-conditioned generation, masked modeling, and diffusion-based frameworks. T2M-GPT [48] represents motion using discrete tokens for autoregressive text-to-motion generation, while Fg-T2M [49] improves fine-grained language–motion alignment through diffusion modeling. Masked approaches such as MoMask [50] further enable motion synthesis from incomplete representations, whereas MotionLCM [51] investigates more efficient controllable generation through latent consistency modeling. Attention-based [52] and contact-aware [53] approaches additionally improve semantic conditioning and human–object interaction modeling. Together, these developments demonstrate that modern generative models can provide rich semantic and structural priors for human motion.

Among controllable motion models, OmniControl [32] is particularly relevant to incomplete-skeleton reconstruction because it supports spatial constraints on arbitrary joints at arbitrary timesteps. This enables partial skeletal observations to serve as control information, while unconstrained regions are inferred from the learned motion prior. However, the existing motion generation studies primarily evaluate these models in terms of realism, text–motion alignment, diversity, controllability, or generation efficiency rather than downstream action-recognition performance. Therefore, motion plausibility alone does not guarantee that generated motion restores the discriminative skeletal evidence required for HAR. This distinction motivates the use of a pretrained controllable motion generator as an external motion prior within the recognition pipeline, rather than as a standalone motion synthesis model.

3. Methodology

Our proposed framework extends an existing skeleton-based HAR model with a recognition-guided generative imputation capability without requiring occlusion-specific retraining of the recognition backbone. An overview of the framework is shown in Figure 1. The target HAR model is first trained using non-occluded skeleton sequences. During evaluation under missing-skeleton conditions, an incomplete sequence is first classified by the fixed HAR model, and a validation-calibrated uncertainty gate determines whether generative imputation should be activated. For triggered samples, a fixed pretrained OmniControl model is used as an external motion prior to generating multiple motion candidates directly at the same temporal length as the HAR input. The generated candidates are evaluated according to their compatibility with the observed skeleton, and the selected candidate is spatially aligned to the same coordinate system. Only missing regions are then replaced by generated motion. Temporal refinement is subsequently applied to suppress local jitter, after which the original visible joints are restored. The completed sequence is finally re-evaluated using the same HAR model to obtain the final recognition result.

3.1. Data Acquisition and Skeleton Sequence Estimation

Video data were acquired using a single monocular RGB camera (Panasonic HC-VX3; 1920 × 1080 pixels; 60 fps). One participant was recorded while performing real construction-related renovation tasks in an indoor environment. The raw video recordings were manually segmented into individual action clips, and each clip was assigned one of five action labels according to the dominant activity: walking, carrying, climbing (ascending/descending) a stepladder, stepladder work, and tabletop work. The data acquisition environment and video samples are shown in Figure 2.

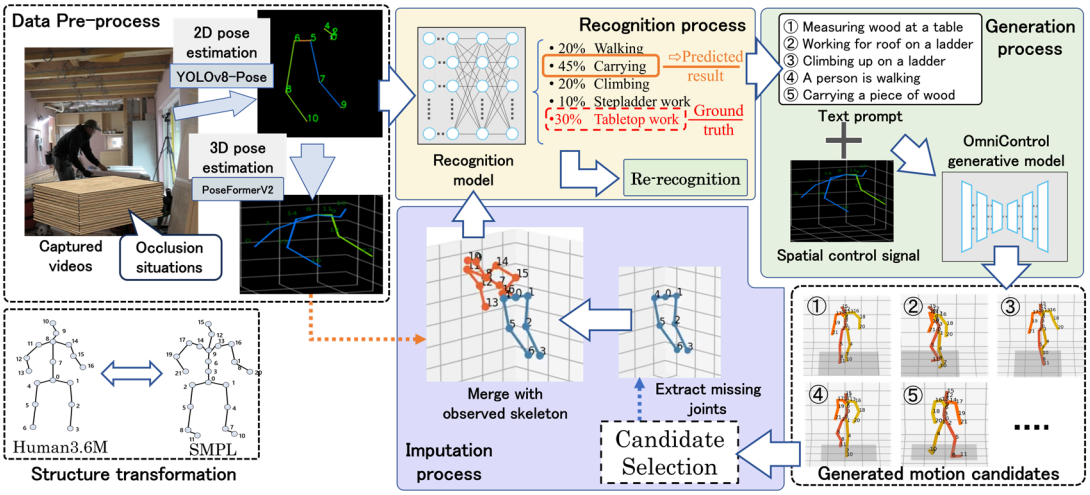


Figure 1. Overview of the proposed method.

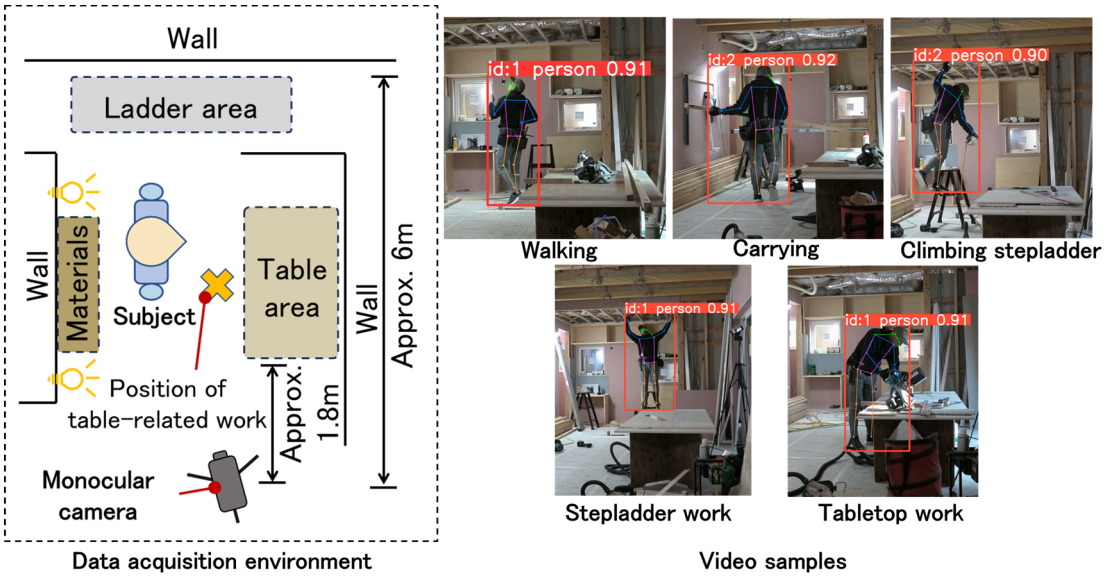


Figure 2. Data acquisition environment and video samples.

RGB frame sequences were sampled from each video clip at 29.97 fps for subsequent pose estimation and skeleton sequence transformation. YOLOv8 Pose [54] was applied to each sampled frame to estimate 2D human keypoints in the COCO 17-joint (COCO17) representation. When multiple persons appeared in a frame, the worker corresponding to the annotated action was identified during preprocessing. The estimated COCO17 keypoints were then mapped to the Human3.6M-style 17-joint (H36M17) representation and subsequently lifted to 3D coordinates using PoseFormerV2 [55]. The resulting single-person 3D skeleton sequences were temporally normalized to a fixed length of 64 frames. Each processed sample was therefore represented as an H36M17 skeleton sequence in $\mathbb{R}^{64\times17\times3}$, which served as the common skeleton representation for the subsequent recognition and generative imputation pipeline. A total of 565 processed action clips were obtained before simulated occlusion and data augmentation, and these were used in the experiments.

3.2. Occlusion Simulation and Validation-Calibrated Imputation Trigger

To evaluate recognition robustness under controlled severe-occlusion conditions, nine simulated missing-skeleton patterns were applied to the 64-frame H36M17 sequences. These conditions consisted of eight spatial occlusion patterns—upper body, bottom body,

left side, right side, left hand, right hand, left leg, and right leg—and one temporal interruption. The spatial masks represent the persistent loss of predefined body regions throughout an action sequence, whereas the temporal interruption represents a continuous period in which the entire skeleton is unavailable. These patterns were designed as controlled stress conditions motivated by common visibility loss in construction-related environments rather than as exact reproductions of naturally occurring site occlusions.

Let $X_i \in \mathbb{R}^{T\times J\times 3}$ denote the original 3D skeleton sequence of sample i, where $T = 64$ and $J = 17$. For each occlusion condition o, a binary missing mask $M_i^o \in \{0,1\}^{T\times J}$ is constructed, where $M_{i,t,j}^o = 1$ indicates that joint j is missing at frame t, and $M_{i,t,j}^o = 0$ indicates that it remains observed. The corresponding visible mask and simulated occluded sequence are defined as follows:

$$V_i^o = 1 - M_i^o, \tag{1}$$

$$X_i^o = V_i^o \odot X_i, \tag{2}$$

where V_i^o is the visible mask under the occlusion condition o, and $\odot$ denotes element-wise multiplication. Missing coordinates are represented as zero after masking. The visibility mask was retained for subsequent imputation and integration.

After occlusion simulation, each skeleton sequence is first evaluated by the fixed target HAR model. The first-stage recognition result serves two purposes: it provides the baseline prediction under occlusion and supplies the uncertainty information used to determine whether generative imputation should be activated. The trigger does not directly determine whether the prediction is correct, but instead estimates whether the prediction is sufficiently reliable to bypass additional generation. For each sample, the recognition model outputs a class-score vector. Let $s_{i,1}$ and $s_{i,2}$ denote the highest and second-highest class confidence scores, respectively. The confidence margin is defined as follows:

$$\Delta_i = s_{i,1} - s_{i,2}. \tag{3}$$

The top-1 confidence $s_{i,1}$ reflects how strongly the model supports its predicted class, whereas the margin Δ_i reflects how clearly the prediction is separated from the second-best class. A sample bypasses generative imputation only when both quantities exceed their calibrated thresholds. The imputation trigger is defined as follows:

$$g_i = \begin{cases} 0, & s_{i,1} > \theta_{\text{abs}} \text{ and } \Delta_i > \theta_{\text{margin}}, \\ 1, & \text{otherwise}, \end{cases} \tag{4}$$

where $g_i = 1$ indicates that the sample is selected for generative imputation, and $g_i = 0$ indicates that the occluded skeleton sequence is retained without additional generation. This design reduces unnecessary OmniControl inference and avoids modifying samples that are already recognized with sufficient confidence.

Because recognition backbones and occlusion conditions can produce different confidence distributions, the trigger thresholds are calibrated independently for each recognizer–occlusion pair using only the validation subset. The absolute-confidence threshold is searched over $\theta_{\text{abs}} \in \{0.00, 0.05, \ldots, 0.95\}$, and the confidence-margin threshold is searched over $\theta_{\text{margin}} \in \{0.00, 0.05, \ldots, 0.55\}$. After calibration, the selected parameters are fixed and applied to the held-out test subset.

For each threshold pair, the trigger decisions were compared with the correctness of the first-stage recognition results on the validation split. Misclassified samples selected for imputation were treated as successful catches, whereas correctly classified samples selected

for imputation were treated as over-imputation cases. To balance these two effects, the thresholds were selected by maximizing, as follows:

$$J = R_{\text{catch}} - \lambda \cdot R_{\text{over}}, \tag{5}$$

where R_{catch} measures the proportion of recognition failures selected for imputation, R_{over} measures the proportion of correctly recognized samples unnecessarily selected for imputation, and $\lambda = 1.5$ penalizes over-imputation. The calibrated confidence-margin gate, applied to each test sample before motion generation, estimates predictive uncertainty rather than prediction correctness; consequently, incorrect predictions with both high top-1 confidence and a large confidence margin may remain untriggered.

3.3. OmniControl-Based Motion Candidate Generation with Timestep-Respaced Inference

Samples selected by the validation-calibrated trigger are processed using the fixed pretrained OmniControl model. The skeleton sequences used by the HAR models follow the H36M17 representation, whereas OmniControl uses a HumanML3D/SMPL-compatible 22-joint (SMPL22) motion representation. The triggered sequences are therefore converted from H36M17 to SMPL22 through a fixed joint-topology mapping and coordinate-axis transformation.

Let $X_i^H \in \mathbb{R}^{T\times 17\times 3}$ denote the H36M17 skeleton sequence of the i-th selected sample. The sequence was transformed into an SMPL-style representation by applying coordinate-axis conversion and joint-topology mapping:

$$X_i^S = \Phi_{H\to S}(X_i^H), X_i^S \in \mathbb{R}^{T\times 22\times 3}, \tag{6}$$

where $\Phi_{H\to S}$ denotes the fixed coordinate and joint-topology conversion. The corresponding missing mask and visible mask are transformed using the same joint correspondence.

During this conversion, joints that were not directly available in the H36M17 format were either estimated from neighboring joints or treated as missing. The converted skeleton was then normalized to match the SMPL22-style motion distribution by applying scale normalization and root-centered coordinate transformation. Missing joints were kept as zero-valued entries after normalization so that they would not be interpreted as valid spatial constraints.

The converted SMPL22 sequence retained the 64-frame temporal length used by the HAR models. It was then flattened into a spatial hint representation $H_i \in \mathbb{R}^{64\times 66}$, where the 66 dimensions correspond to the three-dimensional coordinates of 22 SMPL joints. A valid spatial control mask was derived from the transformed missing mask and used to ensure that only observed joints were provided as spatial control hints. S_i After standardization using SMPL spatial statistics derived from HumanML3D [56], the hint representation was multiplied by the valid-point mask so that missing joints remained zero in the final OmniControl input.

OmniControl was used as a pretrained controllable motion generation backbone, and no additional training of the diffusion model was performed. For each selected sample, action-level prompts corresponding to the five action categories were used: moving, carrying, climbing, ladder work, and table work. The action-level prompts are listed in Table 1, and all five prompts were used for each selected sample. For each prompt, OmniControl generated K candidate motions, producing multiple possible interpretations of the occluded action. The generation process can be represented as follows:

$$\hat{Y}_{i,c,k}^S = G_\phi(H_i, S_i, p_c, \epsilon_k), c = 1, \ldots, C, k = 1, \ldots, K, \tag{7}$$

where G_ϕ denotes the pretrained OmniControl generator, S_i is the valid spatial control hint, p_c is the c-th action-level prompt, the number of prompts C is set to 5, and ϵ_k denotes the stochastic sampling noise used to generate diverse candidates. In our implementation, the repeat generation number K is set to 5, which resulted in 25 generated candidates for each selected sample.

Table 1. Action-level prompts used for OmniControl generation.

Action Label	Prompt
Moving	A person is walking forward normally.
Carrying	A person is walking forward while holding a heavy object in front of their chest with both hands.
Climbing	A person is climbing up a vertical stepladder, moving their arms and legs upwards alternately.
Ladder work	A person is standing still with their legs apart, reaching both arms upwards to interact with something above their head.
Table work	A person is standing in place, bending slightly forward, and moving their hands back and forth in front of them.

However, the original OmniControl diffusion process performs denoising over 1000 timesteps. In the proposed framework, each triggered sample is conditioned on the complete action-prompt bank, and multiple stochastic candidates are generated for every prompt. Applying the full 1000-step sampling process to every candidate would therefore result in substantial computational cost. To reduce this cost without modifying or retraining the pretrained motion generator, an inference-time timestep-respacing strategy inspired by denoising diffusion implicit models (DDIMs) [57] was introduced.

The proposed modification retains the OmniControl network architecture, pretrained checkpoint, text and spatial conditioning, classifier-free guidance, and stochastic ancestral denoising mechanism, and changes only the diffusion timestep schedule used during inference. Given the original 1000-step diffusion process, in this study, the DDIM-based inference method retains 50 uniformly spaced timesteps with a stride of 20:

$$\mathcal{T}_{50} = \{0, 20, 40, \ldots, 980\}. \tag{8}$$

A spaced diffusion process reconstructs the transition schedule between the retained timesteps so that their cumulative noise levels remain consistent with the original 1000-step diffusion trajectory. Each retained timestep is mapped back to its original index before being provided to the pretrained network. The model architecture, checkpoint, text and spatial conditioning, classifier-free guidance, and stochastic denoising rule therefore remain unchanged, and no additional training is required. This modification reduces the number of denoising-network evaluations for each candidate and accelerates the generation process of the proposed method.

Each generated SMPL22 motion is converted back into the H36M17 representation through inverse joint-topology and coordinate-axis transformations. The converted candidates are not used directly as final HAR inputs, but are instead passed to the candidate-evaluation and frame-wise spatial-alignment stage, where the motion hypothesis most compatible with the observed skeleton is selected before missing-region integration.

3.4. Candidate Evaluation and Missing-Region Integration

The generated candidates were first converted back into the H36M17 representation and then evaluated before missing-region integration. Candidate selection aimed to identify

the motion hypothesis that was most compatible with the reliable observations while maintaining a stable skeletal structure and temporal motion. Four complementary scores were used: visible-joint distance consistency, bone-vector directional consistency, bone-length ratio consistency, and temporal motion smoothness.

Let $X_i^{occ} \in \mathbb{R}^{T\times J\times 3}$ denote the occluded H36M17 skeleton sequence of the i-th selected sample, where $T = 64$ and $J = 17$. Let $Y_{i,n} \in \mathbb{R}^{T\times J\times 3}$ denote the n-th generated candidate, where $n = 1, \ldots, N$ and N denotes the total number of candidates. The visibility mask is denoted by $V_i \in \{0,1\}^{T\times J}$, where $V_{i,t,j} = 1$ indicates that joint j at frame t is visible, and $V_{i,t,j} = 0$ indicates that it is missing.

First, visible-joint distance consistency measures how closely a generated candidate matches the observed skeleton over visible joints. To reduce the influence of global coordinate differences during candidate scoring, each generated candidate was temporarily fitted to the visible observations using a similarity transformation estimated from the available joints. This temporary fitting was used only for candidate ranking. After selection, the best candidate underwent a separate frame-wise spatial-alignment procedure before missing-region integration. Let $\hat{Y}_{i,n}$ denote the fitted candidate. The visible-joint distance score is computed as follows:

$$D_{i,n} = \frac{\sum_{t=1}^{T}\sum_{j=1}^{J} V_{i,t,j}\left\|\hat{Y}_{i,n,t,j} - X^{o}_{i,t,j}\right\|}{\sum_{t=1}^{T}\sum_{j=1}^{J} V_{i,t,j} + \epsilon}, \tag{9}$$

where a smaller $D_{i,n}$ indicates better agreement with the reliable visible joints.

Second, bone-vector directional consistency evaluates whether the candidate preserves bone directions consistently with the visible parts of the observed skeleton. For each bone in the skeletal edge set, unit bone vectors are computed for both the fitted candidate and the observed skeleton. Their cosine dissimilarity is averaged only over bones whose two endpoints are visible. This score is denoted by $C_{i,n}$, where a smaller value indicates better directional consistency.

Third, bone-length ratio consistency evaluates structural plausibility. For each candidate, bone lengths are normalized by the total body bone length in each frame, and the temporal variation in these normalized bone-length ratios is measured. Candidates whose body proportions fluctuate unnaturally over time receive larger penalties. This score is denoted by $P_{i,n}$.

Fourth, temporal motion smoothness is measured using the mean magnitude of the second-order temporal difference:

$$S_{i,n} = \frac{1}{(T-2)J}\sum_{t=1}^{T-2}\sum_{j=1}^{J}\left\|\tilde{Y}_{i,n,t+2,j} - 2\tilde{Y}_{i,n,t+1,j} + \tilde{Y}_{i,n,t,j}\right\|, \tag{10}$$

where $\tilde{Y}_{i,n}$ denotes the temporally smoothed candidate, and a smaller $S_{i,n}$ indicates smoother motion with fewer abrupt fluctuations.

Because the four scores have different numerical scales, each score type was normalized across all candidates of the same sample using local min–max normalization:

$$\bar{q}_{i,n} = \frac{q_{i,n} - \min_k q_{i,k}}{\max_k q_{i,k} - \min_k q_{i,k} + \epsilon}, q \in \{D, C, P, S\}. \tag{11}$$

The final candidate score was computed as a weighted sum:

$$\mathcal{L}_{i,n} = 1.0\bar{D}_{i,n} + 1.5\bar{C}_{i,n} + 2.0\bar{P}_{i,n} + 0.5\bar{S}_{i,n}. \tag{12}$$

The candidate with the lowest total score was selected for missing-region completion:

$$n_i^* = \arg\min_n \mathcal{L}_{i,n}. \tag{13}$$

After candidate selection, the selected sequence was aligned frame by frame to the coordinate system of the observed skeleton. Visible joints in each frame were used as spatial anchors. When at least three valid and non-degenerate anchors were available, a 3D similarity transformation was estimated, and the scale factor was constrained to the range from 0.5 to 2.0 to prevent unstable enlargement or shrinkage. When the anchor configuration was insufficient, the procedure successively fell back to scale–translation alignment, translation-only alignment, or the original generated coordinates. The resulting aligned candidate is denoted by $\hat{X}_i^{gen}$. After frame-wise alignment, the aligned candidate was integrated with the occluded skeleton using the explicit missing mask retained from the occlusion simulation stage. Generated coordinates were inserted only at locations marked as missing, while the observed coordinates were retained from the original input. This integration can be expressed as follows:

$$X_i^{imp} = V_i^o \odot X_i + M_i^o \odot \hat{X}_i^{gen}, \tag{14}$$

where X_i^{imp} is the completed skeleton sequence. The mask is applied to the coordinate dimension during integration. To suppress local high-frequency jitter in the integrated motion, a Savitzky–Golay filter [58] was applied along the temporal dimension of each joint-coordinate trajectory. A window length of 9 frames and a second-order polynomial were used.

Finally, the completed skeleton sequence is used as the input for second-stage action recognition. For samples selected by the recognition-guided trigger, the imputed sequence is fed back into the same action recognition model. For samples not selected for imputation, the original occluded sequence is retained. The recognition output obtained in this second stage is used as the final action recognition result. In this way, the proposed approach forms a closed-loop process in which the first recognition stage determines whether imputation is necessary, and the second recognition stage evaluates the skeleton sequence after selective missing-region completion.

4. Experiments, Results, and Discussion

4.1. Experimental Settings, Comparison Methods, and Evaluation Protocol

Datasets and data partitioning: The collected indoor renovation dataset contained 565 original action clips from five action categories: walking, carrying, climbing stepladder, ladder work, and tabletop work. Each clip was represented as an H36M17 skeleton sequence estimated from monocular RGB videos. For recognition, all skeleton sequences were uniformly sampled or linearly interpolated to 64 frames and treated as single-person skeleton inputs. The clips were partitioned at the original-sample level using label-stratified sampling. The training, validation, and test subsets contained 395, 85, and 85 clips, respectively, corresponding to an approximate ratio of 70:15:15 while preserving similar action-class proportions across the subsets. Left–right skeleton swapping was applied only to the training subset after partitioning, increasing the training set from 395 to 790 sequences. The validation and test subsets contained only original, non-augmented sequences.

External validation was conducted using the processed NW-UCLA dataset [59] with ten action classes. Kinect-v1 20-joint skeletons were mapped to the H36M17 representation and temporally normalized to 64 frames. View 3 was reserved for testing, while the remaining views were used to construct the training and validation subsets. To reduce duplicate leakage, 95 View-2 sequences that were byte-identical to View-3 test samples

were removed from the training and validation pool. For NW-UCLA samples containing pre-existing missing skeletal observations, the natural missing mask was retained and combined with each simulated occlusion mask.

All HAR backbones were trained exclusively using non-occluded training sequences. The validation subsets were used for checkpoint selection and calibration of the confidence-margin trigger, whereas the held-out test subsets were reserved for final recognition and imputation evaluation.

Recognition models and training settings: Four representative graph-based skeleton action recognition backbones were evaluated: ST-GCN, ST-GCN++, CTR-GCN, and MSG3D. Their original network architectures were retained without modification. The same fixed checkpoint was used for first-stage recognition and second-stage recognition after imputation, and no HAR model was fine-tuned using occluded or generated skeletons. The output classification layer was configured for five classes on the collected dataset and ten classes on the NW-UCLA dataset.

The pose-estimation pipeline was implemented using Ultralytics 8.0.123, PoseFormerV2, and PyTorch 2.0.0, while the HAR models were implemented using MMAction2 1.2.0. The recognition backbones were optimized using stochastic gradient descent with a momentum of 0.9, a weight decay of 0.0005, and cosine learning-rate scheduling. The initial learning rate was 0.1 for ST-GCN, ST-GCN++, and CTR-GCN and 0.05 for MSG3D. Each model was trained for 200 epochs, and the checkpoint with the highest top-1 accuracy on the non-occluded validation subset was used in all subsequent evaluations. The proposed method executions were distributed across Windows 11 single-GPU workstations. The software configuration and model parameters were fixed within each controlled comparison.

Motion generation settings: A pretrained OmniControl model was used as the external motion generation prior without additional training or fine-tuning on either evaluation dataset. Motion was generated directly at the 64-frame temporal length used by the HAR models. The main configuration employed the DDIM50-style timestep-respacing profile, which retained 50 effective stochastic ancestral denoising steps from the original 1000-step diffusion process. Five stochastic repetitions were generated for each prompt. The collected dataset therefore produced 25 candidates from five prompts, whereas NW-UCLA produced 50 candidates from ten prompts.

Comparison and ablation settings: The no-generation setting directly used the first-stage HAR prediction of the occluded skeleton and served as the recognition baseline. In the force-generation setting, every test sample was sent to OmniControl, while candidate ranking, alignment, and integration were kept identical to the main configuration. The proposed setting used the validation-calibrated confidence-margin gate and generated motion only for triggered samples.

Skeleton-in-Context (SiC) [27] was evaluated as an external learned completion baseline using the available motion-completion-only checkpoint. SiC was applied to the masked test sequences for each occlusion condition, and its completed skeletons were evaluated using the same fixed HAR checkpoints. Because this checkpoint and completion scope differ from the full multi-task model reported in the original SiC study, the comparison is interpreted as an external motion-completion baseline rather than a reproduction of the complete SiC framework. SiC completed all masked test samples, whereas the proposed method generated motion only for samples selected by the confidence-margin trigger. The influence of diffusion-step reduction was evaluated by comparing DDIM50 and DDIM100 timestep-respacing profiles. Only the number of effective denoising steps was changed from 50 to 100.

Evaluation protocol: Top-1 recognition accuracy was used as the primary evaluation metric. For each model–occlusion condition, Acc_{before} denotes the first-stage recognition

accuracy of the occluded skeleton, and Acc_{after} denotes the final accuracy after selective imputation and second-stage recognition. ΔAcc denotes the change in accuracy from Acc_{before} to Acc_{after}, and is reported in percentage points. For each recognition model and occlusion condition, changes in paired first-stage and final predictions were evaluated using an exact McNemar test. Holm correction was subsequently applied across the 36 model–occlusion combinations of each dataset. The trigger rate denotes the proportion of test samples forwarded to the generation module. A rescue is defined as a sample changing from an incorrect first-stage prediction to a correct final prediction, whereas a harm is a sample changing from correct to incorrect. Their difference provides the descriptive net correction count. Intervals were estimated using a base-sample-level cluster bootstrap with 10,000 iterations. The 85 test clips of the collected dataset or the 464 NW-UCLA test sequences were resampled as clusters, thereby preserving the dependence among repeated decisions obtained from the same sample across recognition models and occlusion conditions.

4.2. Evaluation Results and Discussion

Table 2 summarizes the average recognition performance of the proposed method on the collected dataset and NW-UCLA. Each value is averaged over the 36 combinations of four recognition models and nine occlusion conditions. For the collected dataset, the average accuracy increased from 61.01% before imputation to 65.00% after imputation, resulting in an improvement of 3.99 percentage points. Generation was activated for 32.29% of the recognition decisions. The base-sample-level cluster-bootstrap confidence interval ranged from 2.48 to 5.46 percentage points, supporting a positive overall effect of the proposed method.

Table 2. Average recognition accuracy before and after recognition-guided imputation on two datasets.

Dataset	Before (%)	After (%)	ΔAcc (pp)	Trigger Rate (%)	Cluster-Bootstrap 95% CI (pp)
Collected	61.01	65.00	3.99	32.29	2.48–5.46
NW-UCLA	42.63	50.31	7.67	28.24	6.92–8.37

A larger improvement was obtained for NW-UCLA, where the average accuracy reached 50.31% compared with 42.63% before imputation. The gain was 7.67 percentage points, despite a lower trigger rate of 28.24%. The corresponding confidence interval extended from 6.92 to 8.37 percentage points, and 19 of the 36 model–occlusion combinations remained significant after Holm correction. The greater improvement cannot be explained simply by more frequent generation. Instead, the lower initial accuracy of NW-UCLA provided more recognition failures that could potentially be corrected. In addition, the action distribution of NW-UCLA may contain motion patterns that are more compatible with the pretrained motion prior. In contrast, the collected data include real-world and comparatively complex construction-related action variations. For these motions, the action-level prompts used in this study need to balance generality and completion precision. We adopted broadly defined prompts that can cover diverse samples but may provide insufficient guidance for fine-grained motion completion, which may partly explain the smaller improvement observed on the collected dataset. The external result therefore supports the applicability of the proposed framework beyond the collected construction-related data.

Table 3 compares the results for the four recognition models. Positive average changes were obtained for every model on both datasets. For the collected dataset, MSG3D showed the strongest response, improving from 56.86% to 63.79% and gaining 6.93 percentage

points. ST-GCN and CTR-GCN improved by 3.66 and 3.40 percentage points, respectively, whereas ST-GCN++ showed the smallest gain of 1.96 percentage points.

Table 3. Average recognition performance before and after recognition-guided imputation across the nine occlusion conditions for each HAR model.

Dataset	Model	Before (%)	After (%)	ΔAcc (pp)	Trigger Rate (%)
Collected	ST-GCN	65.36	69.02	3.66	31.90
	ST-GCN++	56.86	58.82	1.96	35.69
	CTR-GCN	64.97	68.37	3.40	27.97
	MSG3D	56.86	63.79	6.93	33.59
NW-UCLA	ST-GCN	43.22	51.92	8.69	34.53
	ST-GCN++	40.90	47.03	6.13	23.59
	CTR-GCN	47.27	55.03	7.76	33.52
	MSG3D	39.13	47.25	8.12	21.34

The relatively large improvement of MSG3D on the collected dataset may be related to its multi-scale spatio-temporal graph modeling. Reconstructed local joint trajectories can influence motion representations at several spatial and temporal scales, allowing recovered information to contribute to higher-level action features. In comparison, ST-GCN++ achieved a smaller improvement even though its trigger rate was the highest among the four models. This result indicates that increasing the number of generated samples does not necessarily produce a larger benefit. Some triggered sequences may remain intrinsically ambiguous, while generated trajectories may also differ in their compatibility with the features learned by each recognition model.

The four models responded more uniformly on NW-UCLA. ST-GCN obtained the largest gain of 8.69 percentage points, followed by MSG3D with 8.12 and CTR-GCN with 7.76. ST-GCN++ also improved by 6.13 percentage points. Notably, MSG3D obtained its improvement while triggering only 21.34% of the recognition decisions. These results further indicate that trigger coverage alone is not a sufficient indicator of effectiveness; instead, whether the triggered samples contain recoverable recognition errors and whether the generated motion restores information that is useful for the corresponding recognition model are more important factors.

Table 4 compares the nine occlusion conditions, with values averaged across the four recognition models. Considerable differences were observed among the masks, indicating that the effectiveness of imputation depends strongly on both the amount and type of missing motion.

For the collected dataset, all eight spatial occlusion conditions improved. Left-hand occlusion produced the largest gain of 7.35 percentage points. In this condition, hand-related motion is lost, while the torso, opposite side of the body, and bottom-body movement remain available. These observations provide useful constraints for motion generation, while reconstruction restores motion that is important for distinguishing manipulation-related activities such as carrying and table work. The combination of discriminative missing information and sufficient visible context may therefore explain the relatively large improvement.

Left-leg and right-leg occlusions each gained 5.88 percentage points, while bottom-body occlusion improved by 5.59 percentage points. Walking and climbing a stepladder include characteristic bottom-body motion, and the remaining torso and upper-body trajectories provide additional information for estimating the missing sequence. The results

suggest that coordinated or periodic joint motion can be effectively reconstructed when a sufficient portion of the body remains observable.

Table 4. Average recognition performance under each occlusion condition.

Dataset	**Collected**			**NW-UCLA**		
Occlusion Condition	**Before (%)**	**After (%)**	**ΔAcc (pp)**	**Before (%)**	**After (%)**	**ΔAcc (pp)**
Upper body	27.06	32.35	5.29	10.29	11.48	1.19
Bottom body	62.94	68.53	5.59	40.25	44.23	3.99
Left side	54.41	56.47	2.06	34.32	48.28	13.95
Right side	60.00	63.82	3.82	49.62	56.47	6.84
Left hand	56.18	63.53	7.35	25.43	42.67	17.24
Right hand	76.47	78.24	1.76	63.85	65.68	1.83
Left leg	69.71	75.59	5.88	71.23	73.17	1.94
Right leg	69.12	75.00	5.88	64.28	71.12	6.84
Temporal interruption	73.24	71.47	−1.76	24.41	39.66	15.25
Average	61.01	65.00	3.99	42.63	50.31	7.67

While upper-body occlusion also improved by 5.29 percentage points, its final accuracy remained low, at 32.35%. This mask removes a much larger amount of information than an individual hand or leg condition. The remaining bottom-body motion is often insufficient to uniquely determine work actions whose differences are mainly expressed through the arms and torso. Generative imputation can therefore recover part of the missing information, though it cannot fully compensate for the simultaneous loss of several highly discriminative joints.

Right-hand occlusion showed only a small improvement of 1.76 percentage points. Its accuracy before imputation was already 76.47%, leaving considerably less room for correction. The difference between left-hand and right-hand occlusion may also reflect asymmetric work motions, camera viewpoint, or pose-estimation quality in the collected dataset. Since these factors were not independently controlled, the observed left–right difference should not be interpreted as a general property of human motion.

Temporal interruption was the only condition that decreased in the collected dataset, with its accuracy changing from 73.24% to 71.47%. Unlike spatial occlusion, temporal interruption removes the complete skeleton over a continuous interval. The generated segment must therefore reproduce not only plausible joint configurations but also a motion phase that remains consistent with the observed frames before and after the missing period. While the revised pipeline avoids the previous 64-to-196-to-64 temporal conversion and applies temporal refinement after integration, it does not explicitly constrain velocity, phase, or boundary continuity. A generated segment can consequently remain physically plausible while introducing temporal information that is unfavorable for recognition.

The NW-UCLA results showed a different tendency. Left-hand occlusion again provided the largest improvement, reaching 42.67% from an initial accuracy of 25.43%. Left-side occlusion also gained 13.95 percentage points. In both cases, substantial motion information remains available from the opposite side and central body structure, enabling the generator to infer the missing motion from coordinated body dynamics.

Temporal interruption also improved markedly on NW-UCLA, increasing from 24.41% to 39.66%. The difference from the collected dataset indicates that temporal completion is strongly dependent on the target's motion distribution. The considerably lower accuracy

before imputation on NW-UCLA provided more recognition errors that could potentially be rescued, and the action patterns in NW-UCLA may also be more compatible with the motion prior learned by OmniControl. Therefore, temporal interruption should not be regarded as intrinsically unsuitable for generative completion; its effectiveness rather depends on the relationship between the missing interval, the available context, and the target actions.

Upper-body occlusion remained the most difficult NW-UCLA condition, with its accuracy improving only from 10.29% to 11.48%. In addition to the extensive loss of skeletal information, the trigger rate for this condition was only 9.38%. Many incorrect predictions were therefore not forwarded to the generation module. This observation illustrates a limitation of the current confidence-margin trigger: it estimates prediction uncertainty rather than prediction correctness, and confidently incorrect predictions can bypass imputation.

To further examine the source of these changes, Figures 3 and 4 present confusion matrices for the individual model–occlusion combinations with the best positive and worst negative changes on each dataset. On the collected dataset, CTR-GCN under left-hand occlusion improved from 55.3% to 67.1%. The recall of carry increased from 0.17 to 0.42, while that of table work rose from 0.37 to 0.67. Before imputation, 75% of carry samples and 43% of table-work samples were classified as climb. These values decreased to 33% and 10% after imputation. Restoring the missing hand-related motion therefore appears to recover information that helps distinguish manipulation-oriented actions from climbing.

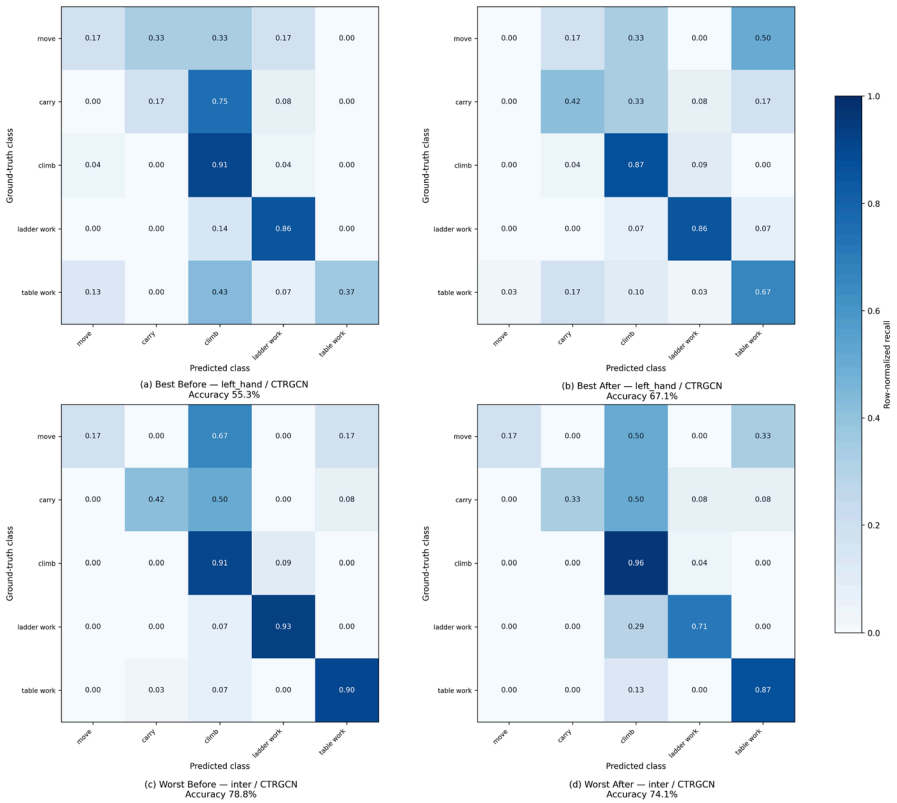


Figure 3. Row-normalized confusion matrices for the model–occlusion combinations with the best and worst accuracy changes on the collected dataset.

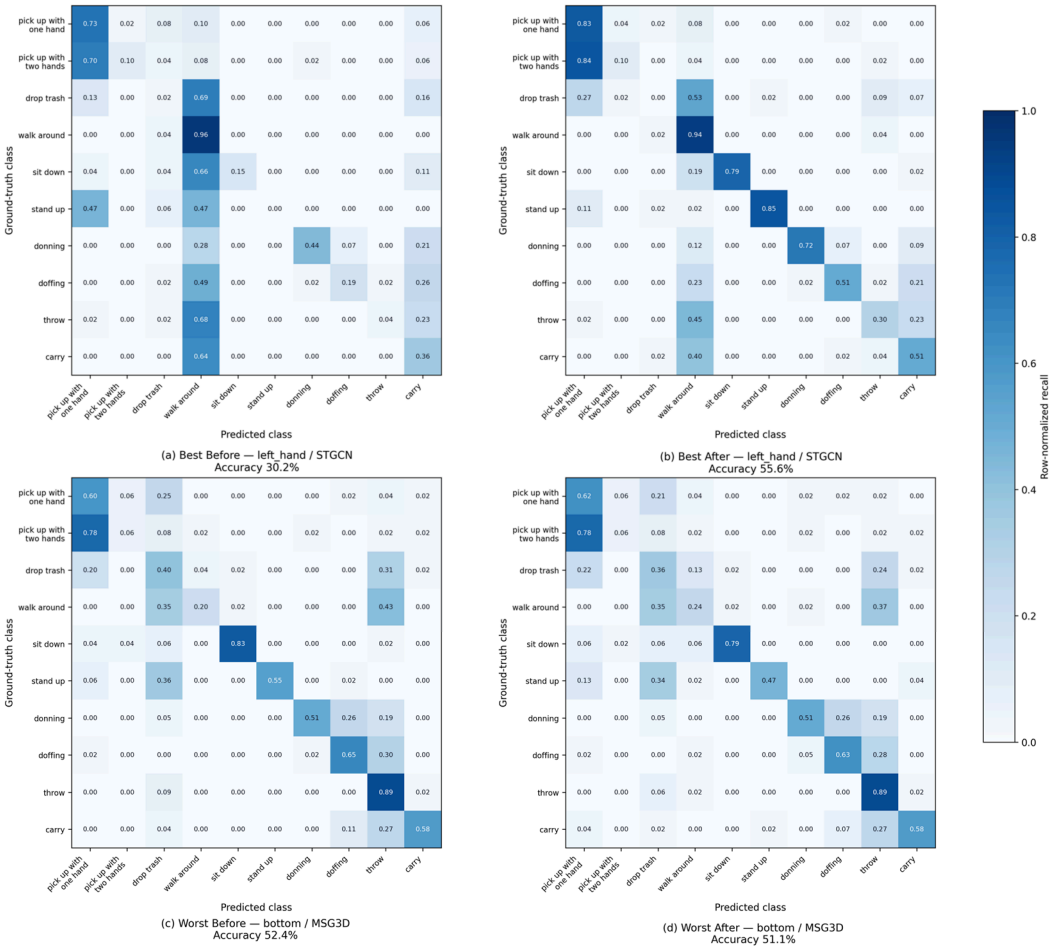


Figure 4. Row-normalized confusion matrices for the model–occlusion combinations with the best and worst accuracy changes on NW-UCLA.

The improvement was not uniform across all classes. The recall of move decreased from 0.17 to zero, and half of these samples were classified as table work after imputation. This suggests that generated hand motion can occasionally introduce manipulation-like characteristics that shift an otherwise correct feature representation toward another activity. The proposed method therefore improves the overall decision boundary without guaranteeing improvement for every action class.

The largest observed negative change for the collected dataset occurred for CTR-GCN under temporal interruption, where accuracy decreased from 78.8% to 74.1%. The main change was observed for ladder work, with its recall falling from 0.93 to 0.71, while confusion with climb increased from 0.07 to 0.29. At the same time, the recall of climb improved from 0.91 to 0.96. These results suggest that the generated middle segment introduced stronger climbing-like motion characteristics, reducing the distinction between climbing and remaining on the ladder while working. This is consistent with the difficulty of reconstructing the correct motion phase when an entire continuous interval is unavailable.

For NW-UCLA, ST-GCN under left-hand occlusion showed a substantially larger improvement, with accuracy increasing from 30.2% to 55.6%. Particularly clear recall improvements were observed for sit down, stand up, donning, doffing, throw, and carry. Before imputation, many of these actions were classified as walk around, suggesting that ST-GCN relied heavily on the remaining bottom-body motion when hand-related joints were unavailable. Completing the missing motion restored upper-body coordination and reduced this concentration of errors.

The largest observed negative change for NW-UCLA occurred for MSG3D under bottom-body occlusion. Its accuracy changed only slightly from 52.4% to 51.1%, and the confusion matrices show no single dominant failure. Small improvements in several classes were offset by reductions in sit down, stand up, drop trash, and doffing. This pattern indicates a local redistribution of classification errors rather than a broad degradation of the representation. Many NW-UCLA actions are primarily characterized by upper-body interaction, so plausible reconstruction of the bottom body may provide limited additional discriminative information and can slightly disturb the motion relationships modeled by MSG3D.

Figures 5 and 6 show joint t-SNE visualizations for the same representative combinations. For CTR-GCN under left-hand occlusion on the collected dataset, the feature distribution after imputation shows a more compact table-work region, while the principal climb and ladder-work regions remain identifiable. Carry and move samples continue to overlap with neighboring classes. The improvement therefore appears to result mainly from local correction of ambiguous samples rather than a complete reorganization of the learned feature space.

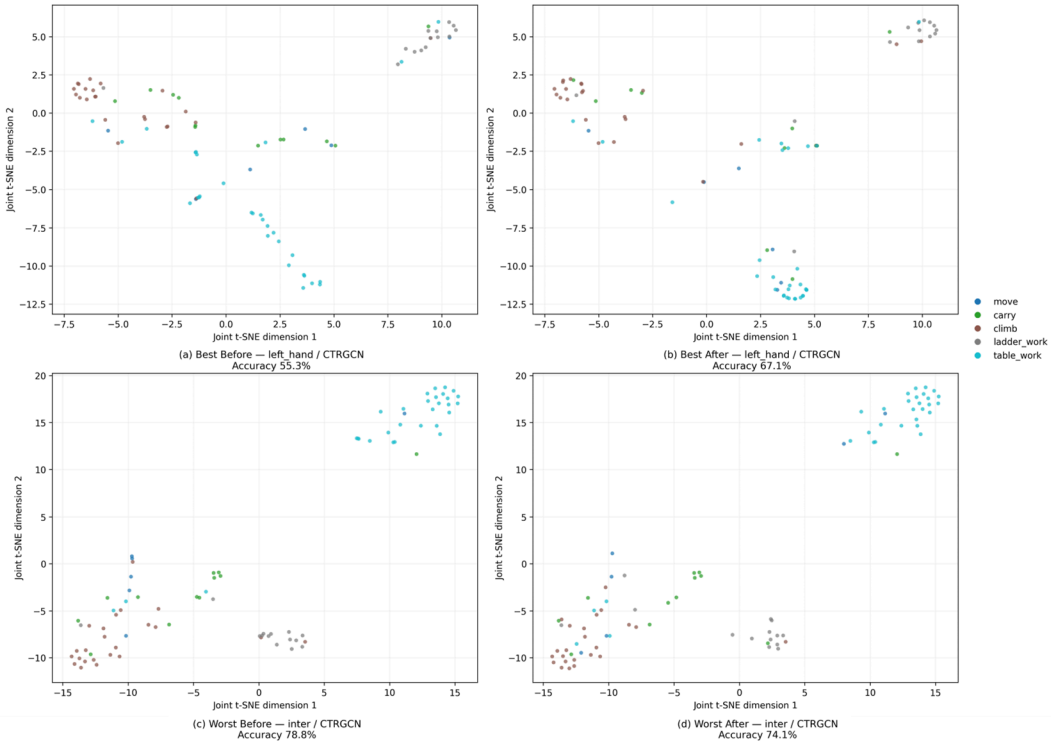


Figure 5. Joint t-SNE feature visualizations before and after imputation for the best positive and worst negative accuracy changes on the collected dataset.

Under temporal interruption, the collected dataset t-SNE distributions remain globally similar before and after imputation. The major class regions occupy comparable areas, with changes mainly occurring near overlapping regions. This agrees with the confusion matrices, which show that the decrease is concentrated in the relationship between ladder work and climb rather than caused by a collapse of all class representations.

For NW-UCLA, the left-hand condition shows clearer local grouping after imputation for several classes, particularly sit down, stand up, and donning, which is consistent with the corresponding recall improvements. However, several object-related actions remain partially mixed, which agrees with the residual confusion observed in the matrices. In comparison, the bottom-body MSG3D result shows little global change in the t-SNE

distribution, consistent with its small accuracy decrease. The t-SNE results are therefore used only as qualitative supporting evidence and not as an independent quantitative measure of class separability.

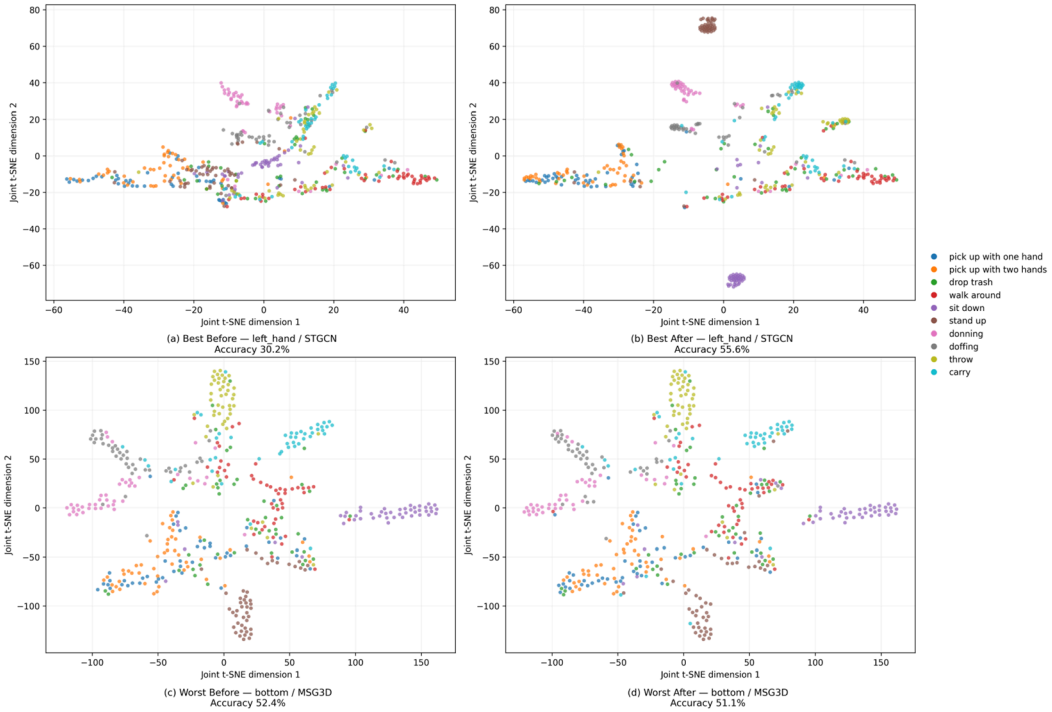


Figure 6. Joint t-SNE feature visualizations before and after imputation for the best positive and worst negative accuracy changes on NW-UCLA.

Figure 7 further illustrates the skeleton-completion process using examples from the main DDIM50 experiment on the collected dataset. In the successful example, the selected candidate initially differs from the occluded input in scale and global position. Frame-wise alignment brings the generated skeleton into the observed coordinate system, after which the final integration preserves the original visible joints and inserts generated coordinates only into the missing region. The resulting skeleton becomes closer to the displayed ground truth and supports the correction of the recognition result. This example demonstrates why spatial alignment and missing-region-only integration are complementary components of the proposed pipeline.

The temporal interruption example illustrates a remaining limitation. During the fully missing interval, no visible joints are available as spatial anchors. Alignment therefore provides little constraint, and the integrated motion is determined mainly by the generated candidate. Although the generated pose remains anatomically plausible, the visualization shows that it does not necessarily reproduce the original joint positions or the actual motion phase. This indicates that generated motion may improve recognition by restoring useful action cues without accurately reconstructing the missing skeleton itself. When the generated motion is inconsistent with the original spatial trajectory or temporal progression, it may also introduce misleading information into the second-stage recognition.

Overall, the quantitative and qualitative analyses show that generative imputation is most effective when the missing region contains discriminative motion and sufficient visible context remains available. The main failure modes arise when the trigger misses confidently incorrect samples or when continuous missing intervals provide too few spatial and temporal constraints.

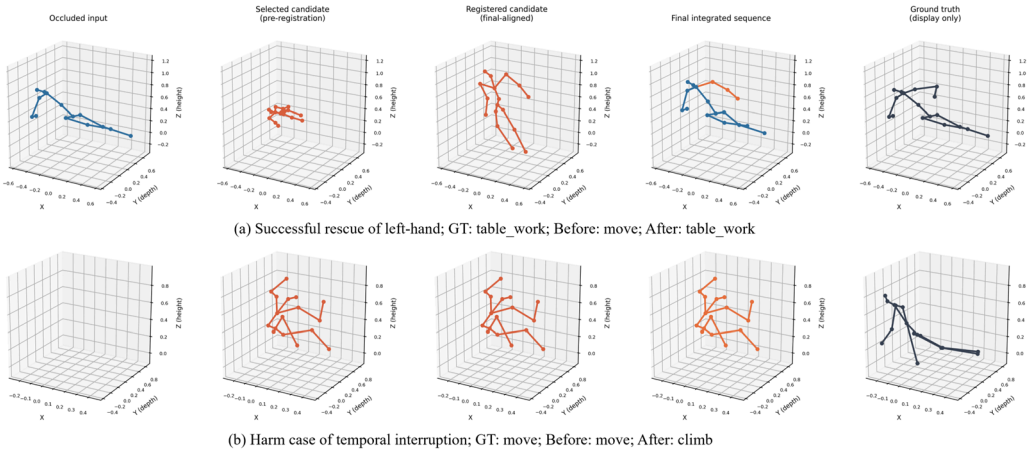


Figure 7. Skeleton-completion process visualizations on the collected dataset. (**a**) A successful rescue in which an initial move prediction was corrected to the ground-truth table-work class after imputation. (**b**) A harm case under temporal interruption, in which an initially correct move prediction changed to climb after imputation. Blue represents the originally observed joints retained from the occluded input, whereas orange represents the generated joints. The dark skeleton shows the ground truth for visual comparison only.

4.3. Ablation, Computational Cost, and Comparison with Skeleton-in-Context

The ablation experiments first examined whether selective generation was preferable to applying generative imputation to every input. Table 5 compares the proposed validation-calibrated confidence-margin trigger with force generation and also summarizes the DDIM50 and DDIM100 sampling settings. All configurations in Table 5 shared the same first-stage accuracy of 61.01% on the collected dataset.

Table 5. Generation trigger and sampling-step analysis on the collected dataset.

Configuration	After (%)	ΔAcc (pp)	Trigger Rate (%)	Rescue	Harm	Sampling Time per Triggered Sample (s)	Amortized Sampling Time per Input (s)
Validation margin (DDIM50)	65.00	3.99	32.29	257	135	86.60	28.00
Force generation (DDIM50)	63.30	2.29	100.00	507	437	—	—
Validation margin (DDIM100)	64.58	3.56	32.29	256	147	105.60	34.10

The validation-calibrated trigger achieved a final accuracy of 65.00% while forwarding only 32.29% of the recognition decisions to OmniControl. Force generation processed every sample but reached 63.30%. Although the direct accuracy difference between these two settings was not statistically significant, their error-transition patterns were substantially different. Force generation corrected 507 first-stage errors but also changed 437 initially correct predictions into errors, leaving a net correction of only 70 decisions. Selective generation produced fewer rescues, but its 257 rescues and 135 harms resulted in a larger net correction of 122 decisions.

The trigger therefore reduced generation calls by 67.7% while concentrating generation on a more productive subset of samples. Relative to the number of generated samples, the net corrected decisions corresponded to approximately 12.35% under the validation-calibrated setting and only 2.29% under force generation. These results indicate that the

main role of the trigger is not simply to increase the number of corrected predictions. Instead, it limits unnecessary modification of samples that are already recognizable and improves the allocation efficiency of the computationally expensive generation process. Nevertheless, the confidence-margin rule remains an uncertainty-based selector and can miss confidently incorrect predictions.

The influence of the diffusion sampling steps was examined by comparing DDIM50 and DDIM100 under matched conditions. The same first-stage predictions, triggered samples, prompts, candidate number, and downstream processing were used, with only the number of effective denoising steps changed. DDIM50 reached 65.00%, while DDIM100 obtained 64.58%. The difference was only 0.42 percentage points in favor of DDIM50, and the confidence interval for the paired difference included zero. Increasing the sampling steps from 50 to 100 therefore provided no observable recognition advantage.

In contrast, the additional diffusion steps produced a clear computational cost. The sampling time increased from 86.6 to 105.6 s per triggered sample, corresponding to approximately 21.9% additional processing time. After accounting for the trigger rate, the average sampling cost per input decision rose from 28.0 to 34.1 s. These results suggest that a longer denoising trajectory does not necessarily generate motion that is more useful for recognition. The proposed framework does not require the generated candidate to reproduce the ground-truth motion exactly, but requires a plausible completion that restores discriminative information while remaining compatible with the observed skeleton. Additional denoising steps may therefore improve aspects of motion generation without producing a corresponding improvement in the final recognition decision. DDIM50 was consequently adopted as the main setting because it provided a more favorable balance between recognition performance and inference cost.

The runtime analysis further confirmed that motion generation is the dominant computational bottleneck. The complete numerical pipeline for the main collected dataset experiment required approximately 25.10 h when optional visualization was excluded. Of this time, motion generation accounted for 24.02 h, representing more than 95% of the total processing time. First-stage and second-stage recognition required approximately 20.40 and 40.60 min, respectively, while trigger calibration, input preparation, and integration contributed only a small fraction of the total runtime. Selective triggering and timestep-respacing therefore address the same practical limitation from two complementary directions: the trigger reduces how often generation is performed, while DDIM50 reduces the cost of each individual generation process.

The proposed method was also compared with Skeleton-in-Context using the available motion-completion-only checkpoint. Table 6 summarizes the results. Because the SiC setting represents only its motion-completion function, the comparison should be interpreted as an external skeleton-completion baseline rather than a reproduction of the complete multi-task SiC framework.

Table 6. Comparison between the proposed method and Skeleton-in-Context.

Dataset	Method	Before (%)	After (%)	ΔAcc (pp)	Completion Rate (%)
Collected	Proposed method	61.01	65.00	3.99	32.29
Collected	SiC MC-only	57.06	50.10	−6.96	100.00
NW-UCLA	Proposed method	42.63	50.31	7.67	28.24
NW-UCLA	SiC MC-only	42.63	49.63	7.00	100.00

For the collected dataset, the proposed method improved its own pipeline by 3.99 percentage points, whereas SiC MC-only showed a decrease of 6.96 percentage points.

However, the two experiments were executed as independent pipelines and differed in 239 of the 3060 first-stage recognition decisions. Their final accuracies should therefore not be treated as a strict paired comparison; the more appropriate observation is that the proposed pipeline produced a positive before-and-after change for the collected data, whereas the SiC completion pipeline did not transfer consistently to these construction-related actions. This difference may reflect a mismatch between the motion prior learned by SiC and the action distribution of the collected dataset.

A direct comparison was possible for NW-UCLA because the two methods shared the same first-stage predictions. The proposed method reached 50.31%, while SiC MC-only obtained 49.63%. The resulting difference of 0.68 percentage points favored the proposed method, but the confidence interval included zero and therefore does not support a significant accuracy advantage. More importantly, the proposed method achieved this performance while processing only 28.24% of the recognition decisions, whereas SiC completed every input sequence. The proposed method also produced positive changes across a larger proportion of the model–occlusion combinations. These observations support the use of recognition-guided selective completion, though they do not indicate that the proposed motion generator is universally more accurate than SiC.

SiC required substantially less completion time in the recorded experiments, with observed processing times of approximately 0.108 s per clip on the collected dataset and 0.049 s per clip on NW-UCLA. The observed difference, together with the stage-level runtime analysis, indicates that diffusion-based multi-candidate generation remains computationally intensive. The proposed method adopts a multi-stage pipeline in which recognition, motion generation, candidate evaluation, spatial alignment, and integration are implemented as separate components. Consequently, the overall runtime depends on the computational efficiency of each introduced technique, particularly the motion generator. Compared with an end-to-end completion model, this separated design provides fewer opportunities for joint optimization, feature sharing, and reuse of intermediate representations, which may limit processing efficiency. Nevertheless, its modularity allows the completion mechanism to be incorporated into existing recognition models without occlusion-specific retraining.

Overall, the validation-calibrated trigger improves the efficiency of generation by focusing computation on uncertain samples and reducing harmful modifications of already recognizable sequences. DDIM50 further reduces the dominant generation cost without showing a measurable loss in recognition performance relative to DDIM100. The comparison with SiC demonstrates that skeleton completion is strongly dependent on the target action distribution and that a comparable final accuracy can be achieved through different completion strategies. Although the proposed method benefits from selective, recognition-guided use of a general motion prior, its multi-candidate diffusion process remains computationally expensive. Future improvements should therefore focus on more reliable recognition-error selection and more efficient motion generation.

5. Conclusions

In this study, we proposed a recognition-guided generative skeleton imputation method for improving skeleton-based action recognition under severe occlusion. Our method introduces a fixed pretrained controllable motion generator as an external motion prior at inference time, without requiring occlusion-specific retraining of the target recognition model. A validation-calibrated confidence-margin trigger selectively activates generation for uncertain samples. Multiple motion candidates are then generated, evaluated, spatially aligned, and integrated only into the missing regions while preserving the observed skeletal information.

The proposed method was evaluated using a collected dataset containing 565 indoor renovation action clips and the NW-UCLA dataset, with four graph-based recognition models and nine occlusion conditions. For the collected dataset, the average accuracy increased from 61.01% to 65.00%, corresponding to an improvement of 3.99 percentage points. External validation on NW-UCLA increased the average accuracy from 42.63% to 50.31%, yielding a gain of 7.67 percentage points. All four recognition models showed positive average changes. Left-hand occlusion produced the largest average improvement for both datasets, indicating that generative imputation is particularly effective when the missing region contains discriminative motion information and sufficient observed context remains available to constrain generation.

Despite these findings, our study has several limitations. The collected dataset contains one participant, and its original-sample-level partition does not provide subject- or recording-session-independent evaluation. The nine occlusion conditions are controlled artificial masks and do not fully reproduce the irregular, dynamic, and view-dependent occlusions encountered in actual construction environments. Although NW-UCLA provides external cross-view validation, it does not represent construction-specific activities. In addition, motion generation accounted for more than 95% of the measured numerical pipeline runtime, which currently limits real-time deployment.

Future work will evaluate this method using multi-participant and multi-site data collected under naturally occurring construction site occlusions, while also investigating improved confidence calibration and recognition-error detection, more reliable temporal integration, and more efficient generation using fewer candidates or lightweight motion models. These developments will help extend the method from controlled robustness evaluation to practical real-time construction monitoring.

Author Contributions: H.Y.: conceptualization, methodology, software, formal analysis, investigation, writing—original draft, visualization. Y.K.: supervision, project administration, writing—review and editing. N.K.: resources, data curation, review, technical guidance. K.I.: resources, data curation, review, technical guidance. K.K.: resources, data curation, review, technical guidance. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: This study and all data acquisition procedures were conducted in accordance with the Declaration of Helsinki and were approved by the Institutional Review Board of Akita University, Japan (approval code: 2022-013; approval date: 10 March 2023; updated date: 5 March 2025).

Informed Consent Statement: Written informed consent was obtained from the participant involved in the study.

Data Availability Statement: The collected participant video data and associated work site information were provided in collaboration with ADK Fuji System Co., Ltd., and are subject to confidentiality restrictions. Therefore, the raw data cannot be made publicly available. The NW-UCLA dataset is publicly available from its original source [59]. Derived data may be made available by the corresponding author upon reasonable request, subject to institutional and company approval.

Acknowledgments: The authors thank the Faculty of Informatics and Data Science, Akita University, for providing computational facilities and academic support during the research. The authors also thank ADK Fuji System Co., Ltd., for providing experimental data and technical support during the research. During the preparation of this manuscript, the authors used ChatGPT-5.6 Sol (OpenAI) for language editing and improving the readability of the manuscript. The authors reviewed and took full responsibility for the content of the publication.

Conflicts of Interest: The authors N.K., K.I., and K.K. are employed by the company ADK Fuji System Co., Ltd. Their contributions are disclosed in the Author Contributions Section. The remaining authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CTR-GCN	Channel-Wise Topology Refinement Graph Convolutional Network
DDIM	Denoising Diffusion Implicit Models
HAR	Human Action Recognition
H36M17	Human3.6M 17-Joint Representation
MC-only	Motion-Completion-Only
MSG3D	Multi-Scale Graph 3D Convolutional Network
NW-UCLA	Northwestern-UCLA Multiview Action 3D
SiC	Skeleton-in-Context
SMPL	Skinned Multi-Person Linear Model
SMPL22	22-Joint SMPL Representation
ST-GCN	Spatial-Temporal Graph Convolutional Network
ST-GCN++	Spatial-Temporal Graph Convolutional Network++
t-SNE	t-Distributed Stochastic Neighbor Embedding

References

1. Asao, K.; Seitani, H.; Stepanyan, A.; Xu, T. *The Impact of Aging and AI on Japan's Labor Market: Challenges and Opportunities*; IMF Working Papers; International Monetary Fund (IMF): Washington, DC, USA, 2025. [CrossRef]
2. Li, J.; Miao, Q.; Zou, Z.; Gao, H.; Zhang, L.; Li, Z.; Wang, N. A Review of Computer Vision-Based Monitoring Approaches for Construction Workers' Work-Related Behaviors. *IEEE Access* **2024**, *12*, 7134–7155. [CrossRef]
3. Liu, W.; Meng, Q.; Li, Z.; Hu, X. Applications of Computer Vision in Monitoring the Unsafe Behavior of Construction Workers: Current Status and Challenges. *Buildings* **2021**, *11*, 409. [CrossRef]
4. Badhan, S.J.; Samsami, R. Artificial Intelligence (AI) in Construction Safety: A Systematic Literature Review. *Buildings* **2025**, *15*, 4084. [CrossRef]
5. Adebowale, O.; Agumba, J. Artificial Intelligence Technology Applications in Building Construction Productivity: A Systematic Literature Review. *Acta Structilia* **2023**, *30*, 161–195. [CrossRef]
6. Ren, B.; Liu, M.; Ding, R.; Liu, H. A Survey on 3D Skeleton-Based Action Recognition Using Learning Method. *Cyborg Bionic Syst.* **2024**, *5*, 0100. [CrossRef] [PubMed]
7. Liu, M.; Liu, H.; Hu, Q.; Ren, B.; Yuan, J.; Lin, J.; Wen, J. 3D Skeleton-Based Action Recognition: A Review. *arXiv* **2025**, arXiv:2506.00915. [CrossRef]
8. Wang, C.; Yan, J. A Comprehensive Survey of RGB-Based and Skeleton-Based Human Action Recognition. *IEEE Access* **2023**, *11*, 53880–53898. [CrossRef]
9. Liu, Y.; Liu, R.; Hu, Y.; Wu, M.; Xin, W.; Miao, Q.; Wu, S.; Li, L. A Systematic Review of Skeleton-Based Action Recognition: Methods, Challenges, and Future Directions. *IEEE Trans. Neural Netw. Learn. Syst.* **2025**, *37*, 2046–2065. [CrossRef] [PubMed]
10. Zhang, J.; Lin, L.; Yang, S.; Liu, J. Self-Supervised Skeleton-Based Action Representation Learning: A Benchmark and Beyond. *arXiv* **2024**, arXiv:2406.02978. [CrossRef]
11. Duan, H.; Xu, M.; Shuai, B.; Modolo, D.; Tu, Z.; Tighe, J.; Bergamo, A. SkeleTR: Towards Skeleton-Based Action Recognition in the Wild. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023; pp. 13634–13644. [CrossRef]
12. Cormier, M.; Schmid, Y.; Beyerer, J. Enhancing Skeleton-Based Action Recognition in Real-World Scenarios Through Realistic Data Augmentation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW), Waikoloa, HI, USA, 1–6 January 2024; pp. 290–299. [CrossRef]
13. Tian, Y.; Chen, J.; Kim, J.I.; Kim, J. Lightweight Deep Learning Framework for Recognizing Construction Workers' Activities Based on Simplified Node Combinations. *Autom. Constr.* **2024**, *158*, 105236. [CrossRef]
14. Xiao, L.; Yang, X.; Peng, T.; Li, H.; Guo, R. Skeleton-Based Activity Recognition for Process-Based Quality Control of Concealed Work via Spatial–Temporal Graph Convolutional Networks. *Sensors* **2024**, *24*, 1220. [CrossRef] [PubMed]
15. Tian, Y.; Lin, S.; Xu, H.; Chen, G. A Spatial-Temporal Multi-Feature Network (STMF-Net) for Skeleton-Based Construction Worker Action Recognition. *Sensors* **2024**, *24*, 7455. [CrossRef] [PubMed]

16. Li, Z.; Li, D. Action Recognition of Construction Workers under Occlusion. *J. Build. Eng.* **2022**, *45*, 103352. [CrossRef]
17. Vernikos, I.; Spyropoulos, T.; Spyrou, E.; Mylonas, P. Human Activity Recognition in the Presence of Occlusion. *Sensors* **2023**, *23*, 4899. [CrossRef] [PubMed]
18. Chen, Z.; Wang, H.; Gui, J. Occluded Skeleton-Based Human Action Recognition with Dual Inhibition Training. In Proceedings of the 31st ACM International Conference on Multimedia, Ottawa, ON, Canada, 29 October–3 November 2023; pp. 2625–2634. [CrossRef]
19. Shi, W.; Li, D.; Wen, Y.; Yang, W. Occlusion-Aware Graph Neural Networks for Skeleton Action Recognition. *IEEE Trans. Ind. Inform.* **2023**, *19*, 10288–10298. [CrossRef]
20. Mathe, E.; Vernikos, I.; Spyrou, E.; Mylonas, P. Leveraging Artificial Occluded Samples for Data Augmentation in Human Activity Recognition. *Sensors* **2025**, *25*, 1163. [CrossRef] [PubMed]
21. Peng, K.; Roitberg, A.; Yang, K.; Zhang, J.; Stiefelhagen, R. Delving Deep into One-Shot Skeleton-Based Action Recognition with Diverse Occlusions. *IEEE Trans. Multimed.* **2023**, *25*, 1489–1504. [CrossRef]
22. Yun, H.; Nakamura, E.; Kageyama, Y.; Ishizawa, C.; Kato, N.; Igarashi, K.; Kawamoto, K. Action Recognition of Simulated Workplace with Occlusion based on Interpolated Skeleton Data using OpenPose. *Int. J. Innov. Comput. Inf. Control* **2024**, *20*, 231–247. [CrossRef]
23. Lee, W.; Park, S.; Kim, T. Denoising Graph Autoencoder for Missing Human Joints Reconstruction. *IEEE Access* **2024**, *12*, 57381–57389. [CrossRef]
24. Han, X.; Nishida, N.; Morita, M.; Sakai, T.; Jiang, Z. Compensation Method for Missing and Misidentified Skeletons in Nursing Care Action Assessment by Improving Spatio-temporal Graph Convolutional Networks. *Bioengineering* **2024**, *11*, 127. [CrossRef] [PubMed]
25. Guo, J.; Ji, Q.; Shan, G. Overcomplete Graph Convolutional Denoising Autoencoder for Noisy Skeleton Action Recognition. *IET Image Process.* **2023**, *18*, 233–246. [CrossRef]
26. Vernikos, I.; Spyrou, E. Skeleton Reconstruction Using Generative Adversarial Networks for Human Activity Recognition Under Occlusion. *Sensors* **2025**, *25*, 1567. [CrossRef] [PubMed]
27. Chen, L.; Zhang, J.; Li, Y.; Pang, Y.; Xia, X.; Liu, T. HumanMAC: Masked Motion Completion for Human Motion Prediction. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023; pp. 9544–9555. [CrossRef]
28. Valls Mascaro, E.; Ahn, H.; Lee, D. A Unified Masked Autoencoder with Patchified Skeletons for Motion Synthesis. *arXiv* **2023**, arXiv:2308.07301. [CrossRef]
29. Wang, X.; Fang, Z.; Li, X.; Li, X.; Chen, C.; Liu, M. Skeleton-in-Context: Unified Skeleton Sequence Modeling with In-Context Learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024; pp. 2436–2446. [CrossRef]
30. Shafir, Y.; Tevet, G.; Kapon, R.; Bermano, A.H. Human Motion Diffusion as a Generative Prior. *arXiv* **2023**, arXiv:2303.01418. [CrossRef]
31. Yun, H.; Kageyama, Y.; Ishizawa, C.; Kato, N.; Igarashi, K.; Kawamoto, K. Skeleton-Based Interpolation Approach with Motion Generation for Action Recognition on Diverse Occlusions. *Int. J. Innov. Comput. Inf. Control* **2024**, *20*, 1301–1317. [CrossRef]
32. Xie, Y.; Jampani, V.; Zhong, L.; Sun, D.; Jiang, H. OmniControl: Control Any Joint at Any Time for Human Motion Generation. In Proceedings of the International Conference on Learning Representations (ICLR), Vienna, Austria, 7–11 May 2024; pp. 28176–28194. [CrossRef]
33. Chen, Y.; Zhang, Z.; Yuan, C.; Li, B.; Deng, Y.; Hu, W. Channel-Wise Topology Refinement Graph Convolution for Skeleton-Based Action Recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 13359–13368. [CrossRef]
34. Song, Y.; Zhang, Z.; Shan, C.; Wang, L. Constructing Stronger and Faster Baselines for Skeleton-Based Action Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 1474–1488. [CrossRef] [PubMed]
35. Lee, J.; Lee, M.; Lee, D.; Lee, S. Hierarchically Decomposed Graph Convolutional Networks for Skeleton-Based Action Recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023; pp. 10444–10453. [CrossRef]
36. Duan, H.; Zhao, Y.; Chen, K.; Lin, D.; Dai, B. Revisiting Skeleton-Based Action Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 2969–2978. [CrossRef]
37. Chi, H.-G.; Ha, M.H.; Chi, S.; Lee, S.W.; Huang, Q.; Ramani, K. InfoGCN: Representation Learning for Human Skeleton-Based Action Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 20186–20196. [CrossRef]
38. Nanduri, S.K.; Delhi, V.S.K. Foundations for Construction 5.0: A Review-Based Taxonomy for Construction Worker Action Understanding. *J. Inf. Technol. Constr.* **2025**, *30*, 924–962. [CrossRef]

39. Tian, Y.; Li, H.; Cui, H.; Chen, J. Construction Motion Data Library: An Integrated Motion Dataset for On-Site Activity Recognition. *Sci. Data* **2022**, *9*, 726. [CrossRef] [PubMed]
40. Tian, Y.; Chen, J.; Kim, J.I.; Kwac, J. Multiple-Input Streams Attention Network for Skeleton-Based Construction Workers' Action Recognition Using Body-Segment Representation Strategies. *Autom. Constr.* **2023**, *156*, 105104. [CrossRef]
41. Bonyani, M.; Soleymani, M.; Wang, C. Construction Workers' Unsafe Behavior Detection through Adaptive Spatiotemporal Sampling and Optimized Attention-Based Video Monitoring. *Autom. Constr.* **2024**, *165*, 105508. [CrossRef]
42. Liu, W.; Shao, Y.; Zhai, S.; Yang, Z.; Chen, P. Computer Vision-Based Tracking of Workers in Construction Sites Based on MDNet. *IEICE Trans. Inf. Syst.* **2023**, *106*, 653–661. [CrossRef]
43. Khan, S.; Ali, H.; Ullah, Z.; Bulbul, M.F. An Intelligent Monitoring System of Vehicles on Highway Traffic. In Proceedings of the 12th International Conference on Open Source Systems and Technologies (ICOSST), Lahore, Pakistan, 19–21 December 2018; pp. 71–75. [CrossRef]
44. Chen, Y.; Peng, K.; Roitberg, A.; Schneider, D.; Zhang, J.; Zheng, J.; Chen, Y.; Liu, R.; Yang, K.; Stiefelhagen, R. Exploring Self-Supervised Skeleton-Based Human Action Recognition under Occlusions. In Proceedings of the International Joint Conference on Neural Networks (IJCNN), Rome, Italy, 30 June–5 July 2025; pp. 1–9. [CrossRef]
45. Singh, R.; Sharma, A. Occluded Skeleton-Based Multi-Stream Model Using Part-Aware Graph Convolutional Network. *Eng. Appl. Artif. Intell.* **2025**, *156*, 111183. [CrossRef]
46. Tehrani, A.; Yadollahzadeh-Tabari, M.; Zehtab-Salmasi, A.; Enayatifar, R. Wearable Sensor-Based Human Activity Recognition System Employing Bi-LSTM Algorithm. *Comput. J.* **2024**, *67*, 961–975. [CrossRef]
47. Kia, Z.; Yadollahzadeh-Tabari, M.; Motameni, H. Human Activity Recognition by Body-Worn Sensor Data Using Bi-Directional Generative Adversarial Networks and Frequency Analysis Techniques. *J. Supercomput.* **2025**, *81*, 342. [CrossRef]
48. Zhang, J.; Zhang, Y.; Cun, X.; Zhang, Y.; Zhao, H.; Lu, H.; Shen, X.; Ying, S. T2M-GPT: Generating Human Motion from Textual Descriptions with Discrete Representations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 14730–14740. [CrossRef]
49. Wang, Y.; Leng, Z.; Li, F.W.B.; Wu, S.; Liang, X. Fg-T2M: Fine-Grained Text-Driven Human Motion Generation via Diffusion Model. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023; pp. 22035–22044. [CrossRef]
50. Guo, C.; Mu, Y.; Javed, M.G.; Wang, S.; Cheng, L. MoMask: Generative Masked Modeling of 3D Human Motions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024; pp. 1900–1910. [CrossRef]
51. Dai, W.; Chen, L.-H.; Wang, J.; Liu, J.; Dai, B.; Tang, Y. MotionLCM: Real-Time Controllable Motion Generation via Latent Consistency Model. In Proceedings of the European Conference on Computer Vision (ECCV), Milan, Italy, 29 September–4 October 2024; pp. 390–408. [CrossRef]
52. Zhong, C.; Hu, L.; Zhang, Z.; Xia, S. AttT2M: Text-Driven Human Motion Generation with Multi-Perspective Attention Mechanism. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023; pp. 509–519. [CrossRef]
53. Ma, S.; Cao, Q.; Zhang, J.; Tao, D. Contact-Aware Human Motion Generation from Textual Description. *arXiv* **2024**, arXiv:2403.15709. [CrossRef]
54. Ultralytics. Ultralytics YOLOv8. Available online: https://github.com/ultralytics/ultralytics (accessed on 11 June 2026).
55. Zhao, Q.; Zheng, C.; Liu, M.; Wang, P.; Chen, C. PoseFormerV2: Exploring Frequency Domain for Efficient and Robust 3D Human Pose Estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 8877–8886. [CrossRef]
56. 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), New Orleans, LA, USA, 18–24 June 2022; pp. 5152–5161. [CrossRef]
57. Song, J.; Meng, C.; Ermon, S. Denoising Diffusion Implicit Models. *arXiv* **2020**, arXiv:2010.02502. [CrossRef]
58. Savitzky, A.; Golay, M.J.E. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Anal. Chem.* **1964**, *36*, 1627–1639. [CrossRef]
59. Wang, J.; Nie, X.; Xia, Y.; Wu, Y.; Zhu, S.-C. Cross-View Action Modeling, Learning and Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014; pp. 2649–2656. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.