TRACEABLE HUMAN-TO-HUMANOID SIGN LANGUAGE BENCHMARKING

Anonymous authors
Paper under double-blind review

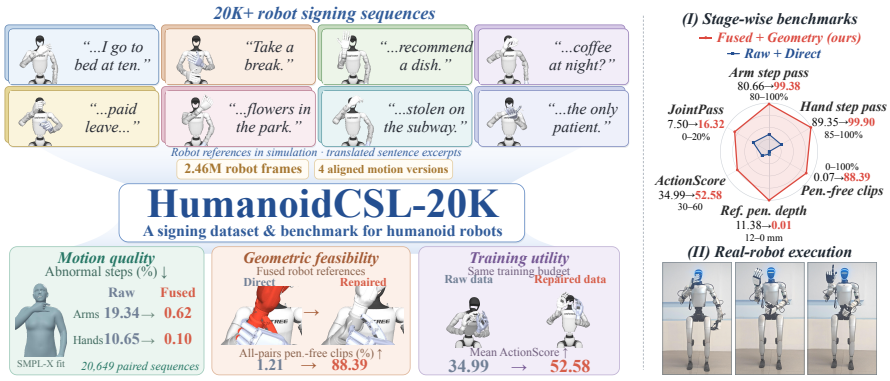

Figure 1: HumanoidCSL-20K links 20,648 signing sequences across four aligned versions of human and robot motion. Simulated examples and benchmark results show fewer abrupt joint changes and robot self-collisions after repair, plus higher control scores under the same training budget.

ABSTRACT

Sign data collection is costly, and teleoperation scales poorly, motivating reuse of large video corpora. Humanoid signing requires converting video-derived human motion into robot trajectories while preserving linguistic motion cues. Errors from fitting, human-motion repair, retargeting, robot geometry repair, and control are hard to separate from the final trajectory alone. We introduce HumanoidCSL-20K, a dataset and benchmark of 20,648 sentence-level Chinese Sign Language sequences, each with four aligned versions: the source, the repaired human motion, the direct robot reference, and the geometry-repaired robot reference. Observation-supported local human-motion repair, full-robot geometry repair, and cross-representation provenance make each transformation traceable. Paired evaluations measure human-motion continuity and content preservation, robot-reference feasibility, and physical execution. A sign-specific kinematic-reference protocol scores handshape, location, palm orientation, and inter-hand relation over the full planned motion. Full-corpus results show fewer abnormal arm / hand steps and less inter-hand and hand-body penetration after repair. Control experiments separate reference learnability from curriculum effects, while component scores expose remaining execution errors.

1 INTRODUCTION

Signing humanoid robots could enable face-to-face communication in public-service settings (Qiao et al., 2025; Khan et al., 2026). The meaning of a sign is carried by continuous and temporally structured motion, in which handshape, location, movement, palm orientation, and bimanual coordination collectively shape it. Reaching a correct final pose alone is therefore insufficient. A signing robot must follow a reference trajectory throughout the sign.

Motion capture provides accurate 3D signing data (Jedlička et al., 2020; 2022), but every new sentence or signer requires a separate session with calibrated equipment, fluent signers, and post-processing. Finger motion is particularly costly to record: DexAvatar used nine cameras with data

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

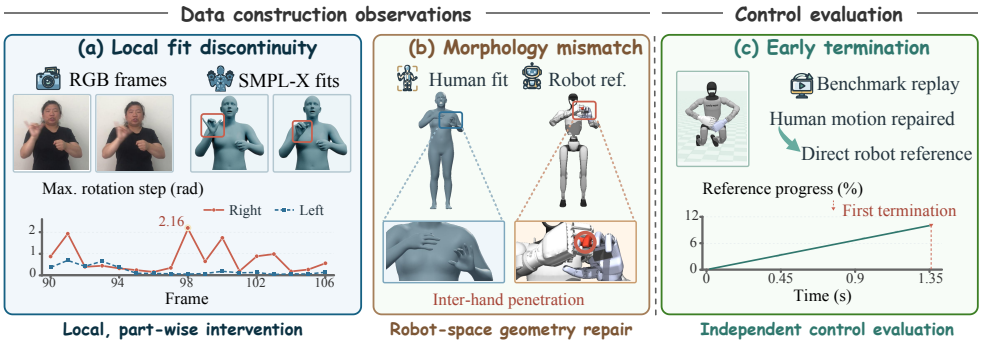

Figure 2: Why fitted human motion is not directly robot-ready. (a) Video fits contain local jumps and frozen hands. (b) Plausible human motion penetrates the robot's own body after retargeting. (c) A collision-free robot reference can still fail in execution. Replay stops at the first failure, and unreached frames score zero.

gloves and still had to filter implausible hand poses (Kundu et al., 2026). In contrast, sign video corpora offer far broader linguistic coverage. CSL-Daily contains 20,654 continuous Chinese Sign Language videos from ten signers with translation and gloss annotations (Zhou et al., 2021), and SOKE provides SMPL-X fits for these videos (Zuo et al., 2025). These resources are a practical starting point for robot signing.

However, fitted human motion cannot drive a robot directly (Figure 2). First, monocular fits are locally unreliable under fast articulation, occlusion, and blur; in the SOKE fits, 19.3% of arm frames and 10.7% of hand frames contain implausibly large rotation steps. Second, the robot's body differs from the signer's; direct retargeting of repaired motion still leaves 50.18% of frames with inter-hand, hand-body, or intra-hand penetration. Third, a collision-free reference remains hard to execute, as the controller must balance the whole body while tracking twelve finger joints; a general-purpose AMP (Peng et al., 2021) tracker stops within about ten frames in every replay.

These errors accumulate along the pipeline, making it difficult to trace a robot failure to its source when only the final motion is evaluated. Existing signing systems are evaluated end to end and provide no corpus-scale benchmark of aligned intermediate motions. We present HumanoidCSL-20K, a dataset and benchmark that retains four aligned versions of each sentence: the source fit, the repaired human motion, the direct robot reference, and the geometry-repaired robot reference. Each stage can thus be evaluated separately, and its effect isolated by replacing it while holding the other stages fixed. The main contributions of this paper are summarized as follows:

- We present HumanoidCSL-20K, the first publicly released large-scale sign-language dataset for humanoid robots. It contains 20,648 sentence-level sequences, each with the source, the aligned motion, and the optimized humanoid data.
- A local human-motion repair is developed, which cuts abnormal rotation steps to 0.62% for arms and 0.10% for hands, with no detectable recognition-BLEU drop against the source fits.
- We introduce a geometry repair for robot references that reduces penetrating frames to 0.29% and certifies 88.21% of sequences as penetration-free, while changing finger joints by at most 20°.
- We establish a stage-wise benchmark with execution metrics and a curriculum controller that learns body/wrist before finger tracking and completes every replay on repaired references.

2 Related Work

Sign-language motion data. 3D signing data come from motion capture or from reconstruction of sign-language videos (Yu et al., 2024). Sign-language production methods generate human-level signing from glosses or text (Tang et al., 2025c; Zuo et al., 2025), and reference-free metrics evaluate produced signing without ground-truth trajectories (Cory et al., 2026). These datasets and production methods stop at human motion. HumanoidCSL-20K extends the SOKE fits of CSL-Daily to

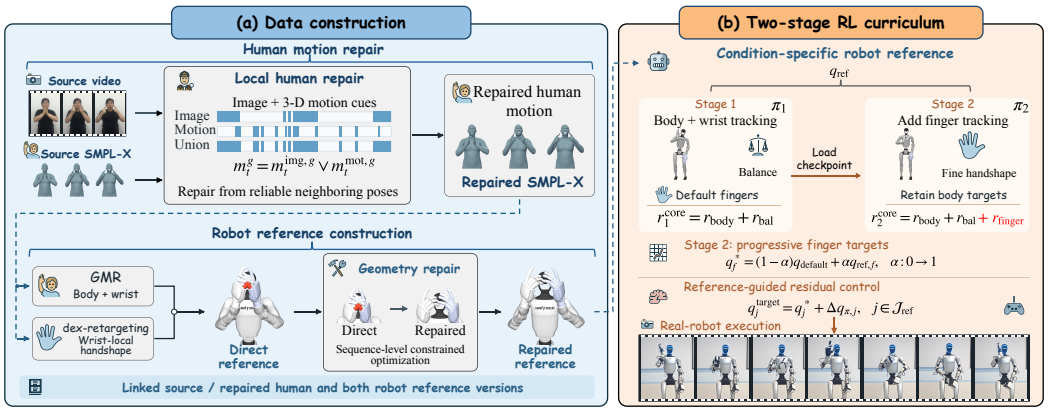

Figure 3: Pipeline overview: (a) linked human and robot motion construction; (b) body-wrist tracking followed by fine handshape tracking.

robot references and records every repair applied to them; our metrics compare each manual component against a known reference trajectory.

Human-to-robot motion transfer. General Motion Retargeting maps body motion to humanoids (Araujo et al., 2025), DexPilot, AnyTeleop, and dex-retargeting map human hands to dexterous robot hands (Handa et al., 2020; Qin et al., 2023; 2022), physics-informed retargeting scales across diverse embodiments (Pan et al., 2025), and kinodynamic retargeting enforces dynamic and contact feasibility (Zhang et al., 2026; Chen et al., 2025). Combining separately retargeted body and hands can make the robot penetrate itself (Section 5.3), which our geometry repair removes over the whole sequence. For robot signing, production methods generate signing motion at the human level (Tang et al., 2025b), SignBot combines retargeting, learned control, and conversational interaction, and Khan et al. refine retargeted trajectories with vision-language guidance and collision mitigation. Both complete systems are evaluated end-to-end, whereas we release the intermediate motions for separate evaluation.

Humanoid motion tracking. Imitation and motion-prior methods (Peng et al., 2018; 2022) and humanoid tracking systems (He et al., 2024b;a; Cheng et al., 2024; Fu et al., 2024; Liao et al., 2026) train policies to follow retargeted human motion, and constrained or residual formulations bound the deviation from the reference (Yan et al., 2024; Sun et al., 2026). Versatile controllers consolidate diverse control modes through motion imitation (He et al., 2025) or masked motion inpainting (Tessler et al., 2024), and universal motion representations provide reusable motor skills across large clip libraries (Luo et al., 2024). Our controller adopts the per-reference tracking of BeyondMimic with residual actions and ramps in finger targets after body and wrist tracking is learned. Imperfect-demonstration methods estimate demonstration quality to reweight supervision (Wu et al., 2019; Zhang et al., 2021; Beliaev et al., 2022; Cui et al., 2019) or bridge discrete observations to continuous trajectories (Tang et al., 2025a); our repair records instead state how each segment was produced, so tracking errors can be grouped by the origin of each reference segment (Section 5.4).

3 Dataset Construction

3.1 Source Motion and Its Errors

The source human motion $\mathcal{H}^{\text{raw}}$ consists of the SOKE SMPL-X fits of CSL-Daily videos for 20,652 sequences. Local human-motion repair yields $\mathcal{H}^{\text{fused}}$ and its direct robot reference $\mathcal{G}^{\text{direct}}$ for 20,649 sequences, and geometry repair yields $\mathcal{G}^{\text{repair}}$ for 20,648 (Figure 3). All versions share sequence identifiers, source-video timestamps, and a human-to-robot body-part mapping, so every frame traces back to the video and its annotation (Appendix A).

Figure 4 shows two error patterns in $\mathcal{H}^{\text{raw}}$. On average, 19.3% of arm frames and 10.7% of hand frames contain a rotation step above 0.30 and 0.70 rad, respectively (panels a, b). Most of these errors are brief, as 97.8% of candidate segments last at most three frames (panel c). The second pattern is

long static hand spans, which last 45 frames at the median and occur in 36.75% of sequences (panel d). Brief jumps can be bridged from neighboring frames, whereas long spans need a held pose. Both patterns are local in time and body part, so the repair is local rather than sequence-wide.

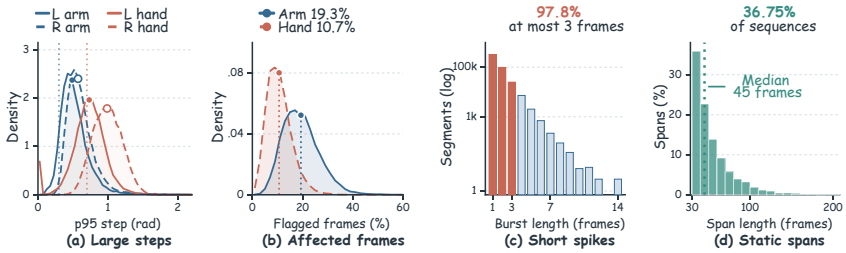

Figure 4: Source-motion diagnostics over 20,652 sequences: (a) rotation-step distributions, (b) affected-frame rates, (c) short motion bursts, and (d) long static-hand spans.

3.2 Local human-motion repair

We repair the left and right arms and hands as four separate part groups. A frame is flagged for a group when either of two cues fires. The image cue uses RTMPose-l WholeBody keypoints (Jiang et al., 2023) and marks low keypoint confidence or visibility, blur, and abrupt keypoint or bounding-box changes. The motion cue marks large fitted rotation steps. For part group g with joints $\mathcal{J}_g$ and threshold θ_g (0.30 rad for arms, 0.70 rad for hands), the step cue and the repair mask are

$$m_t^{\mathrm{mot},g} = \mathbf{1}\Big[\max_{j\in\mathcal{J}_g} d_{SO(3)}\big(R_{t-1,j}, R_{t,j}\big) > \theta_g\Big], \qquad m_t^{g} = m_t^{\mathrm{img},g} \vee m_t^{\mathrm{mot},g}, \tag{1}$$

where $m_t^{\mathrm{img},g}$ is the image cue; the motion cue also flags short oscillations (Appendix B.1). We call the repaired motion Fused because its mask is the union of the two cues. Each flagged segment is rebuilt from reliable neighboring frames. A short gap between reliable anchor frames $a < b$ is filled by spherical interpolation,

$$R_{t,j} = \mathrm{Slerp}\Big(R_{a,j}, R_{b,j}, \tfrac{t-a}{b-a}\Big), \qquad a < t < b, \tag{2}$$

and a longer gap holds $R_{a,j}$ and then blends into the next reliable observation. Remaining flagged arm windows receive light smoothing, while wrists and fingers are not smoothed. For every frame and part group, the output records which operation was applied, with a fixed observation-support score for how directly that operation is supported by the video; shared timestamps carry these provenance records to the robot references and policy rollouts (Appendices B.1 and A.2).

The abnormal-step metric in Section 4 uses the same 0.30 and 0.70 rad thresholds as the motion cue, so for Fused it measures how completely the detected jumps are removed. We therefore also report two results that do not depend on these thresholds: a frozen sign recognizer shows no BLEU loss after repair (Section 5.2), and single-stage policies trained on Fused references complete 58 of 60 replays, compared with 34 of 60 for Raw references under the same budget (Section 5.4).

3.3 Robot-reference construction and geometry repair

General Motion Retargeting maps $\mathcal{H}^{\mathrm{fused}}$ to the G1 body and wrist poses, and dex-retargeting maps each human hand to an Inspire hand. Their combination is the direct reference $\mathcal{G}^{\mathrm{direct}}$. The body and hands are retargeted separately, so neither solver checks the assembled robot for collisions. Since the robot's links are also bulkier than human limbs, signing poses that bring the hands close to each other or to the body become penetrations. Geometry repair removes these penetrations in two steps: an arm stage followed by a finger stage.

The arm stage optimizes the 14 shoulder, elbow, and wrist joints $\boldsymbol{x}_t$ over the whole sequence, with the root, waist, legs, and fingers fixed:

$$\begin{aligned} &\min_{\boldsymbol{x}_{1:N}} \sum_{t=1}^{N} \Big(\mathcal{E}_{\mathrm{track}}(\boldsymbol{x}_t) + \lambda_r\, \mathcal{E}_{\mathrm{rel}}(\boldsymbol{x}_t)\Big) + \mathcal{E}_{\mathrm{smooth}}(\boldsymbol{x}_{1:N}) \\ &\text{s.t. } d_k\big(\boldsymbol{x}(\tau)\big) \geq \gamma, \quad k \in \mathcal{K}_{\mathrm{body}},\ \tau \in \Omega_M. \end{aligned} \tag{3}$$

Here $\mathcal{E}_{\text{track}}$ keeps the shoulder, elbow, and wrist frames $b \in \mathcal{B}$ close to their scaled human targets $T^*_{b,t}$, and $\mathcal{E}_{\text{smooth}}$ penalizes temporal differences:

$$\mathcal{E}_{\text{track}}(\boldsymbol{x}_t) = \sum_{b\in\mathcal{B}} \big\| W_b \operatorname{Log}\big(T_b(\boldsymbol{x}_t)^{-1} T^*_{b,t}\big)^{\vee} \big\|_2^2,$$
$$\mathcal{E}_{\text{smooth}}(\boldsymbol{x}) = \lambda_v \|D_1 \boldsymbol{x}\|_2^2 + \lambda_a \|D_2 \boldsymbol{x}\|_2^2 + \lambda_j \|D_3 \boldsymbol{x}\|_2^2, \tag{4}$$

where $\operatorname{Log}(\cdot)^{\vee}$ maps a relative pose to its six-dimensional twist, W_b weights its position and rotation parts, and D_d is the d-th temporal difference (Appendix B.3). The inter-hand relation term

$$\mathcal{E}_{\text{rel}}(\boldsymbol{x}_t) = \big\| (\boldsymbol{p}_{R,t} - \boldsymbol{p}_{L,t}) - (\boldsymbol{p}^h_{R,t} - \boldsymbol{p}^h_{L,t}) \big\|_2^2 \tag{5}$$

keeps the vector between the robot wrists $\boldsymbol{p}_{L,t}, \boldsymbol{p}_{R,t}$ close to that between the scaled human wrists $\boldsymbol{p}^h_{L,t}, \boldsymbol{p}^h_{R,t}$. Every inter-hand and hand-body pair $k \in \mathcal{K}_{\text{body}}$ must keep a clearance $\gamma = 3\,\text{mm}$, at frames and at $M = 8$ instants between consecutive frames (Ω_M), so the arms cannot cross between frames. Joint, speed, acceleration, and jerk limits also apply. It is solved by sequential quadratic programming with linearized distances and penalized slack (Appendix B.3).

The finger stage fixes the arms and adjusts the 12 active finger joints, six per hand, with passive joints following the mechanical coupling. With $\Delta\boldsymbol{q}^f_t$ the change from the direct handshape and $\mathcal{K}_f$ the pairs involving a finger segment, it solves

$$\min_{\Delta\boldsymbol{q}^f, \boldsymbol{s}} \sum_{t=1}^{N} \|\Delta\boldsymbol{q}^f_t\|_2^2 + \mathcal{E}_{\text{smooth}}(\Delta\boldsymbol{q}^f) + \rho^2 \|\boldsymbol{s}\|_2^2$$
$$\text{s.t. } |\Delta q^f_{t,j}| \le \theta_{\max}, \quad d_k\big(\boldsymbol{q}(\tau)\big) + s_{k,\tau} \ge \gamma_f, \quad s_{k,\tau} \ge 0, \quad k \in \mathcal{K}_f,\ \tau \in \Omega_M, \tag{6}$$

where $\theta_{\max} = 20^\circ$ is a hard budget and the clearance $\gamma_f = 0.2\,\text{mm}$ is enforced through the penalized slack $\boldsymbol{s}$; where the two conflict, the handshape is kept and slight contact remains.

Geometry repair produces $\mathcal{G}^{\text{repair}}$ for 20,648 sequences that pass joint-range, speed, finger-coupling, and numerical checks. Let $\delta_{n,u} = [-\min_{k\in\mathcal{K}_{\text{eval}}} d_k(\boldsymbol{q}_{n,u})]_+$ be the worst-pair penetration depth of sequence n at sample u over all audited pairs $\mathcal{K}_{\text{eval}} = \mathcal{K}_{\text{body}} \cup \mathcal{K}_{\text{intra}}$, including intra-hand pairs. A sequence is *certified* when

$$c_n = \prod_{u\in\mathcal{U}^{30}_n \cup \mathcal{U}^{50}_n} \mathbf{1}\big[\delta_{n,u} \le \epsilon\big] = 1, \qquad \epsilon = 1\,\mu\text{m}, \tag{7}$$

where $\mathcal{U}^{30}_n$ contains the 30 Hz frames with 16 interpolated instants per interval and $\mathcal{U}^{50}_n$ the 50 Hz control grid with three interior samples per interval. The 18,214 certified sequences (88.21%) are penetration-free; the other 2,434 retain contact on 2.2% of their samples, classified in Appendix C.2. All statistics use the full $\mathcal{G}^{\text{repair}}$, and the released flag lets users select the certified subset.

4 Benchmark Tasks and Evaluation

4.1 Stage-wise benchmark tasks

The benchmark defines one task per transformation in Figure 3. Each task fixes its input, output, and evaluation set, so a new method can replace one step and be compared with the released versions.

Human-motion repair. The input is $\mathcal{H}^{\text{raw}}$ and the output is a repaired human motion. The released baselines are Raw, Global smoothing (Gaussian smoothing of the whole sequence; Appendix B.1), and Fused. Continuity is measured by the abnormal-step rate of each part group g, the share of steps whose largest within-group rotation exceeds θ_g (Equation 1),

$$A^g = \frac{100}{N} \sum_{n=1}^{N} \frac{1}{2|\mathcal{T}_n|} \sum_{s\in\{L,R\}} \sum_{t\in\mathcal{T}_n} \mathbf{1}\Big[\max_{j\in\mathcal{J}_{g,s}} d_{SO(3)}\big(R_{t-1,j}, R_{t,j}\big) > \theta_g\Big], \tag{8}$$

where s is the body side, $\mathcal{T}_n$ the steps of sequence n, and $N = 20{,}649$. At 30 Hz these steps correspond to 9 and 21 rad/s, far above the arm (4–5.4 rad/s) and finger (4.8 rad/s) speed limits of the robot references (Appendix B.3), so a remaining abnormal step cannot be executed as is. Because the motion cue of Fused uses the same thresholds (Section 3.2), content preservation is measured separately: a frozen pose-only Uni-Sign recognizer (Li et al., 2025) translates each version of 75 held-out sequences (38 development and 37 test examples not used in its training), scored by BLEU.

Robot-reference construction. The input is a human motion and the output is a reference for the G1 robot with Inspire hands. The released baselines are direct retargeting of Raw, Global, and Fused motion, and $\mathcal{G}^{\text{repair}}$. Feasibility is audited over all contact pairs of the assembled robot, $\mathcal{K}_{\text{eval}}$: inter-hand, hand-body, and intra-hand pairs. Each reference is evaluated on the 50 Hz grid samples $\mathcal{U}_n^{\text{grid}}$ and the three interior samples per interval $\mathcal{U}_n^{\text{int}}$, whose union is $\mathcal{U}_n^{50}$. With the depth $\delta_{n,u}$ of Equation 7, the penetration rates and the penetration-free share are

$$P^{\star}_{\text{pen}} = \frac{100}{N}\sum_{n=1}^{N}\frac{1}{|\mathcal{U}_n^{\star}|}\sum_{u\in\mathcal{U}_n^{\star}}\mathbf{1}[\delta_{n,u}>\epsilon], \qquad P_{\text{free}} = \frac{100}{N}\sum_{n=1}^{N}\prod_{u\in\mathcal{U}_n^{50}}\mathbf{1}[\delta_{n,u}\le\epsilon], \tag{9}$$

where $\star \in \{\text{grid}, \text{int}\}$ and $N = 20{,}648$; the mean depth averages $\delta_{n,u}$ in the same way. Computing each quantity per sequence keeps long clips from dominating (Appendix C.2). Certification (Section 3.3) also checks the 30 Hz grid and is therefore stricter than being penetration-free. Two further measures show how much repair changes the motion: the RMS change of the right-minus-left wrist vector and the RMS change of the active finger joints.

Robot execution. The input is a robot reference and the output is a simulated controller rollout that tracks the whole reference with fixed timestamps. The evaluation set is 20 randomly sampled sentence-level motions, each trained with three seeds (Appendix A.3). Rollouts are scored by the sign-specific metrics below and by penetration over the same pair set $\mathcal{K}_{\text{eval}}$ as the reference audit.

4.2 Sign-specific execution metrics

A whole-body joint average can hide the few errors that decide a sign, such as a wrong handshape or crossed wrists, so we score the manual components separately. Let B be a fixed torso link and $W \in \{L, R\}$ a hand. Its torso-relative wrist position and palm orientation are $\widetilde{\boldsymbol{p}}_W = R_B^{\top}(\boldsymbol{p}_W - \boldsymbol{p}_B)$ and $\widetilde{R}_W = R_B^{\top}R_W$. At reference frame t, the executed and reference states give four errors for handshape, hand location, palm orientation, and inter-hand relation:

$$\begin{aligned} e_t^{\text{hand}} &= \max_{j\in\mathcal{J}_{\text{hand}}} |q_{t,j}^{\text{exec}} - q_{t,j}^{\text{ref}}|, & e_t^{\text{pos}} &= \max_{W\in\{L,R\}} \|\widetilde{\boldsymbol{p}}_{t,W}^{\text{exec}} - \widetilde{\boldsymbol{p}}_{t,W}^{\text{ref}}\|_2, \\ e_t^{\text{ori}} &= \max_{W\in\{L,R\}} d_{SO(3)}(\widetilde{R}_{t,W}^{\text{exec}}, \widetilde{R}_{t,W}^{\text{ref}}), & e_t^{\text{rel}} &= \|(\widetilde{\boldsymbol{p}}_R - \widetilde{\boldsymbol{p}}_L)_t^{\text{exec}} - (\widetilde{\boldsymbol{p}}_R - \widetilde{\boldsymbol{p}}_L)_t^{\text{ref}}\|_2. \end{aligned} \tag{10}$$

Here $\mathcal{J}_{\text{hand}}$ contains the twelve active finger joints, and the maximum keeps a local failure from being averaged away. Movement and timing, the remaining cues of a sign, enter through frame-by-frame comparison without temporal warping, which penalizes delayed or desynchronized hands.

Each component $x \in \{\text{hand}, \text{pos}, \text{ori}, \text{rel}\}$ has a nominal scale σ_x of 0.30 rad, 0.10 m, 0.50 rad, and 0.10 m, which also serves as its pass threshold. For handshape, 0.30 rad is about one fifth of the 1.44 rad flexion range of the four Inspire fingers; for palm orientation, 0.50 rad (29°) is below the 45° midpoint between canonical palm directions; for location and relation, 0.10 m is a declared tolerance, coarser than the spacing of some facial locations (Appendix D.2). Scaling all thresholds together changes some rankings, but Fused + Geometry stays above the other references (Appendix D.2).

ActionScore maps each error to a fidelity $\psi_x(e) = 2^{-(e/\sigma_x)^2}$, which is 1 at zero error and 0.5 at one nominal scale, and averages it over the planned motion:

$$C_{i,s}^{x} = \frac{100}{T_i}\sum_{t=0}^{T_i-1} A_{i,s,t}\,\psi_x(e_t^x), \qquad \text{AS}_{i,s} = \frac{1}{4}\sum_{x} C_{i,s}^{x}. \tag{11}$$

Here T_i is the planned length of motion i, and $A_{i,s,t} = 1$ if frame t is reached before failure under seed s, so unreached frames score zero. The four components are weighted equally, and a score of 100 requires exact execution throughout. JointPass is the share of planned frames that are reached, pass all four thresholds, and satisfy the physical constraints:

$$\text{JointPass}_{i,s} = \frac{100}{T_i}\sum_{t=0}^{T_i-1} A_{i,s,t}\,G_{i,s,t}\prod_{x}\mathbf{1}[e_t^x \le \sigma_x], \tag{12}$$

where $G_{i,s,t} = 1$ if every pair in $\mathcal{K}_{\text{eval}}$ penetrates by at most ϵ and every joint stays within its range up to 10^{-3} rad. ActionScore credits partial accuracy continuously, whereas JointPass requires all components and constraints at once. Per-component pass rates are reported in Appendix C.3.

4.3 REPLAY PROTOCOL

Two replay regimes answer different questions. In self-target replay, each policy is trained and evaluated on its own processed reference, which measures how learnable that reference is. In fixed-reference replay, all controllers track the same $\mathcal{G}^{\text{repair}}$ trajectory, which isolates controller differences such as the curriculum. Paired conditions share the robot model, initial state, timestamps, evaluation grid, and failure rules. Each replay tracks the full finger target from the first frame and stops at the first fall, task termination, or numerical failure. Unreached frames stay in all denominators, so stopping early cannot avoid difficult segments. Tracking errors on reached frames are reported together with coverage, and penetration uses every recorded state. For a metric $\mathcal{M}$ and two conditions A and B evaluated on motions $\mathcal{I}$ and seeds $\mathcal{S}$, the reported paired difference averages seeds within each motion and weights motions equally,

$$\bar{\Delta}_{A,B} = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \big(\mathcal{M}^{A}_{i,s} - \mathcal{M}^{B}_{i,s}\big), \tag{13}$$

and its 95% interval resamples motions (Appendices C.3 and C.4).

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Controller. Our controller follows the per-reference tracking of BeyondMimic: a PPO policy is trained for each reference at 50 Hz, with 200 Hz physics, and outputs residuals around the reference joint targets. Signing couples whole-body balance with fine finger motion, so the curriculum first learns balance and wrist placement, which carry hand location and palm orientation, and adds handshape afterwards. The first 1,500 of 3,000 iterations track the body and wrists with the fingers at a default pose. The remaining 1,500 add finger tracking, and a ramp moves the finger targets from the default pose to the reference. The no-ramp variant applies finger targets at once, and the single-stage variant tracks everything for all 3,000 iterations. All variants share the network, PPO settings, and interaction budget (Appendix B.4 lists all settings).

Evaluation. Human-motion and geometry results use the full corpus (Section 4.1). Control results use 20 randomly sampled motions with seeds 42, 1234, and 777, giving 60 replays per condition; paired differences and their 95% intervals follow Equation 13. As an external baseline, AMP is trained natively as one ProtoMotions policy over all 20 motions with the same interaction budget and uses the same scoring (Appendix B.4).

5.2 HUMAN-MOTION REPAIR

Table 1 shows that Global smoothing and Fused both remove most abnormal steps but differ in what they preserve. Global changes every frame and lowers BLEU-1 and BLEU-2 relative to Raw. Fused edits only flagged segments: outside the repair mask it leaves every hand pose unchanged and moves arm joints by 2.4° on average (Appendix C.1). Its BLEU stays within the 95% interval of Raw for every order (BLEU-4: -0.29, $[-2.01, 1.34]$), although the recognizer was trained on raw fits and would penalize any content removed by repair. Fused exceeds Global by 5.45 BLEU-1 and 3.84 BLEU-2 points (BLEU-2 interval $[0.97, 6.80]$), and both gains hold under a Bonferroni adjustment over the four BLEU orders. In Figure 5(b), each frame looks plausible, but the handshape closes and reopens across frames 196–198; Fused replaces the event and keeps the anchor at frame 191.

5.3 ROBOT-REFERENCE CONSTRUCTION

Table 2 shows that human-side processing alone does not make references feasible. With direct retargeting, about half of all frames penetrate for every human input, and at most 1.21% of sequences are penetration-free. Geometry repair (Figure 6) lowers penetrating frames from 50.18% to 0.29% and mean depth from 8.28 to 0.01 mm, and it certifies 88.21% of sequences (Section 3.3). Interior samples follow the grid, so collisions are removed rather than moved between frames.

The changes are concentrated where the robot's body requires them. Its forearm and wrist links have a 4–5 cm collision radius, so forearm crossings that nearly touch on a human signer must be opened

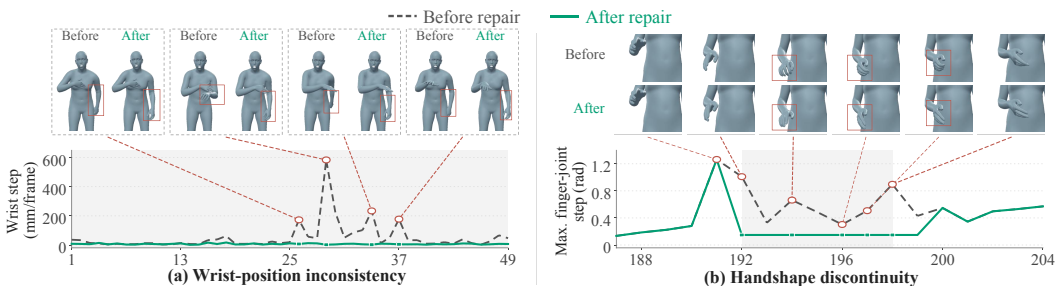

Figure 5: Representative local repair of isolated (a) wrist and (b) handshape discontinuities while retaining neighboring anchors. Boxes mark corrections, and leaders connect poses to curves.

Table 1: Human-motion continuity ($n = 20{,}649$) and content preservation ($n = 75$).

Method	Arm abn. (%) ↓	Hand abn. (%) ↓	BLEU-2 ↑	BLEU-4 ↑
Raw	19.34	10.65	**41.46**	**26.59**
Global smoothing	1.76	0.24	37.33	24.50
Fused	**0.62**	**0.10**	41.17	26.29

Table 2: Robot-reference geometry over all audited pairs at 50 Hz ($n = 20{,}648$).

Method	Pen. (%) ↓	Inter-fr. (%) ↓	Depth (mm) ↓	Pen.-free (%) ↑
Raw + Direct	53.11	53.16	11.38	0.07
Global + Direct	52.47	52.50	11.00	0.21
Fused + Direct	50.18	50.20	8.28	1.21
Fused + Geometry ($\mathcal{G}^{\text{repair}}$)	**0.29**	**0.30**	**0.01**	**88.39**

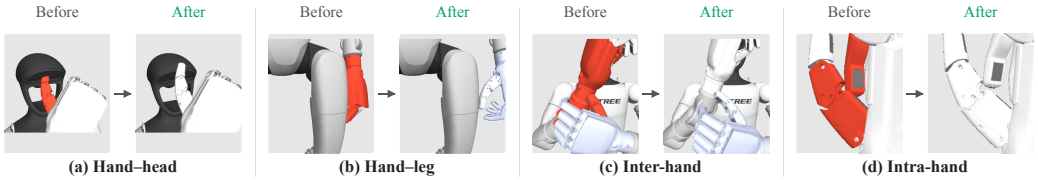

Figure 6: Geometry repair examples for hand-head, hand-leg, inter-hand, and intra-hand penetration; red marks collisions.

on the robot. The repair widens the inter-wrist vector by 7.96 cm RMS for this purpose, while the inter-hand relation term in Equation 3 keeps the rest of the relative hand placement close to the human reference. Handshapes change by only $1.40°$ RMS over the active finger joints (Appendix C.2).

5.4 Robot Execution

Reference processing. Down each column group of Table 3, every policy is trained on its own reference (self-target replay, Section 4.3). Repairing the source motion makes references much easier to learn: with single-stage training, Fused raises completion from 56.7% to 96.7% and ActionScore from 34.99 to 49.15 (paired gain 14.16, 95% interval $[7.23, 21.43]$). Geometry repair adds a further gain on top of Fused in both curricula. It raises ActionScore to 52.58 and 54.10 and JointPass to 16.32% and 17.98%, and under the full curriculum it halves rollout penetration from 26.08% to 13.03% while reducing depth from 0.663 to 0.112 mm (Table 24). Rollouts still penetrate more often than the reference (13.03% versus 0.29%), so execution is evaluated as a separate task.

Curriculum. The last row of Table 3 fixes the Fused + Geometry reference. All our curricula complete every replay, and the full curriculum gives the highest ActionScore and JointPass. Its ActionScore exceeds single-stage and no-ramp training by 1.52 and 0.66 points and is higher for every training seed. The clearest gains are in hand placement: wrist RMS falls from 13.97 to 13.42 cm (95% paired interval $[-1.04, -0.04]$) and palm RMS from $40.97°$ to $39.19°$ ($[-3.24, -0.30]$). Per-component pass rates also favor the full curriculum in the three-seed mean (Appendix C.3).

General-purpose tracker. AMP stops within about ten frames on every reference (Table 3), during the transition from the initial posture, and its ActionScore stays below 3. Repairing the reference does not change this outcome, whereas our controller completes every replay on the repaired ref-

Table 3: Execution on 20 motions and three seeds. Reading down, reference repair improves both of our controllers; reading across, our controllers far exceed AMP, and the full curriculum is best on the repaired reference.

	AMP			Ours, single-stage			Ours, two-stage + ramp		
Reference	AS ↑	Fin. ↑	Pen. ↓	AS ↑	Fin. ↑	Pen. ↓	AS ↑	Fin. ↑	Pen. ↓
Raw + Direct	2.73	0.0	32.24	34.99	56.7	36.15	35.51	60.0	35.53
Fused + Direct	2.68	0.0	26.39	49.15	96.7	25.97	49.29	98.3	26.08
Fused + Geometry	2.79	0.0	23.50	52.58	100.0	16.09	**54.10**	**100.0**	**13.03**

AS: ActionScore; Fin.: finished replays (%); Pen.: penetrating rollout frames (%) over all audited pairs, including intra-hand pairs. Unreached frames score zero. Without the ramp, two-stage training on Fused + Geometry reaches 53.44 / 100.0 / 13.55. Table 24 in Appendix D.1 adds JointPass, depth, and Global smoothing.

erence. Under this protocol, a general-purpose motion prior does not transfer directly to dexterous signing, whereas reference-conditioned tracking designed for signing executes the full motion.

Using the repair records. Because the repair records follow each reference into its rollouts, execution errors can be grouped by how each reference segment was produced. Right-arm tracking error is 0.162 rad RMS on observed frames and 0.240 rad on long held spans (Appendix D.3), so a controller developer can separate errors on held segments from errors on observed motion. The records can also guide training: weighting supervision by the records gives lower high-support finger RMS than removing low-support motions (0.059 versus 0.066 rad).

Hardware replay. Figure 7 shows a physical Unitree G1 with two Inspire hands replaying a signing sequence alongside the source video frames; Appendix E.2 describes the setup and replays.

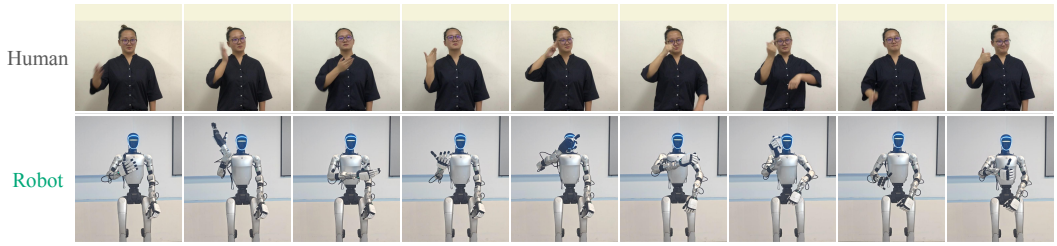

Figure 7: Replay on a physical Unitree G1 humanoid: source-signer frames (top) and the corresponding robot replay (bottom).

6 Conclusion and Limitations

HumanoidCSL-20K provides 20,648 sentence-level Chinese Sign Language sequences, each with four aligned versions: the source, the repaired human motion, the direct robot reference, and the geometry-repaired robot reference. Observation-supported local human-motion repair removes nearly all abnormal steps without a detectable loss in frozen-recognizer BLEU. Full-robot geometry repair lowers penetrating frames from 50.18% to 0.29% and certifies 88.21% of sequences, while changing handshapes by only 1.40° RMS. In execution, repaired references are easier to learn and collide less in rollouts, and our curriculum controller completes every replay on them. Because every version is kept and aligned, each step of the pipeline can be evaluated and replaced on its own.

Limitations and future work. The benchmark covers one sign language, one humanoid, and only manual cues. Recognition BLEU and simulated execution do not establish Deaf comprehension. Execution is tested on only 20 motions in simulation with qualitative hardware replay. Future work will add non-manual cues, other robots and sign languages, dialogue-level interaction, and comprehension studies with Deaf signers.

AI USE STATEMENT

Large Language Models (LLMs) were used in this work as a general-purpose writing and coding assistance tool—for example, to improve grammar, clarify phrasing, check technical terminology in the manuscript, and to assist with implementation of data-processing scripts and evaluation utilities. LLMs did not contribute to the conception of the research idea, theoretical analysis, experimental design, or interpretation of results. All scientific content, including equations, algorithms, and claims, was developed and verified by the authors. No LLM was used to generate novel technical content or to draft substantial portions of the paper. As required by ICLR policy, we confirm that LLMs are not listed as authors, and we take full responsibility for all content under our names.

ETHICS STATEMENT

This work studies motion construction and evaluation for humanoid signing, with the long-term aim of supporting accessibility for Deaf and hard-of-hearing communities. It uses CSL-Daily videos and their SOKE motion fits; source recordings and derived artifacts remain subject to the original access and redistribution conditions. Recorded signers may be identifiable, and conversion to motion representations does not by itself establish anonymity. The reported evaluation measures motion fidelity and physical feasibility and includes no user study of sign comprehension. Evaluation with fluent signers and Deaf community participation is needed before treating the system as a communication aid. Robot-generated signing should support human communication without displacing access to qualified interpreters. Qualitative hardware replay does not establish safety for unsupervised interaction.

REPRODUCIBILITY STATEMENT

Appendix A documents the data representations, source-time alignment, provenance, and evaluation subsets. Appendix B specifies human-motion repair, geometry repair, and policy training, including network architecture, PPO hyperparameters, and curriculum schedules. Appendix C defines the metrics, thresholds, termination rules, aggregation, and failure accounting. The author-controller comparison includes all 20 motions for seeds 42, 1234, and 777 in each configuration. The external AMP baseline adds nine trained policies and 180 scored replays over the same motions, seeds, and three reference tracks. Appendix D.1 reports per-seed results and detailed diagnostics. The versioned evaluation outputs retain trial, reference, model, feature-cache, and implementation hashes to connect aggregate results to their underlying artifacts. Appendix E.2 separately describes the scope and alignment of the qualitative hardware demonstrations. Qualitative demonstrations are available on the anonymous project page. We will publicly release the code, evaluation scripts, benchmark metadata, and derived motion artifacts permitted by the source licenses upon publication.

REFERENCES

Joao Pedro Araujo, Yanjie Ze, Pei Xu, Jiajun Wu, and C Karen Liu. Retargeting matters: General motion retargeting for humanoid motion tracking. *arXiv preprint arXiv:2510.02252*, 2025.

Mark Beliaev, Andy Shih, Stefano Ermon, Dorsa Sadigh, and Ramtin Pedarsani. Imitation learning by estimating expertise of demonstrators. In *International Conference on Machine Learning*, pp. 1732–1748. PMLR, 2022.

Xingyu Chen, Hanyu Wu, Sikai Wu, Mingliang Zhou, Diyun Xiang, Haodong Zhang, Yangchen Zhou, Yukang Gao, Yi Gu, and Renjing Xu. A scalable whole-body motion transfer via implicit kinodynamic motion retargeting. *arXiv preprint arXiv:2509.15443*, 2025. URL https://arxiv.org/abs/2509.15443.

Xuxin Cheng, Yandong Ji, Junming Chen, Ruihan Yang, Ge Yang, and Xiaolong Wang. Expressive whole-body control for humanoid robots. *arXiv preprint arXiv:2402.16796*, 2024.

Oliver Cory, Maksym Ivashechkin, Karahan Sahin, Oline Ranum, Jianhe Low, Edward Fish, Anton Pelykh, Ozge Mercanoglu Sincan, and Richard Bowden. BackTranslation2.0 – a linguistically motivated metric to assess sign language production. *arXiv preprint arXiv:2606.28673*, 2026.

Yuchen Cui, David Isele, Scott Niekum, and Kikuo Fujimura. Uncertainty-aware data aggregation for deep imitation learning. In *2019 International Conference on Robotics and Automation (ICRA)*, pp. 761–767. IEEE, 2019.

Zipeng Fu, Qingqing Zhao, Qi Wu, Gordon Wetzstein, and Chelsea Finn. HumanPlus: Humanoid shadowing and imitation from humans. In *8th Annual Conference on Robot Learning*, 2024. URL https://openreview.net/forum?id=WnSl42M9Z4.

Ankur Handa, Karl Van Wyk, Wei Yang, Jacky Liang, Yu-Wei Chao, Qian Wan, Stan Birchfield, Nathan Ratliff, and Dieter Fox. DexPilot: Vision-based teleoperation of dexterous robotic hand-arm system. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9164–9170. IEEE, 2020.

Tairan He, Zhengyi Luo, Xialin He, Wenli Xiao, Chong Zhang, Weinan Zhang, Kris M. Kitani, Changliu Liu, and Guanya Shi. OmniH2O: Universal and dexterous human-to-humanoid whole-body teleoperation and learning. In *8th Annual Conference on Robot Learning*, 2024a. URL https://openreview.net/forum?id=oL1WEZQal8.

Tairan He, Zhengyi Luo, Wenli Xiao, Chong Zhang, Kris Kitani, Changliu Liu, and Guanya Shi. Learning human-to-humanoid real-time whole-body teleoperation. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 8944–8951. IEEE, 2024b.

Tairan He, Wenli Xiao, Toru Lin, Zhengyi Luo, Zhenjia Xu, Zhenyu Jiang, Jan Kautz, Changliu Liu, Guanya Shi, Xiaolong Wang, Linxi Fan, and Yuke Zhu. HOVER: Versatile neural whole-body controller for humanoid robots. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9989–9996. IEEE, 2025.

Pavel Jedlička, Zdeněk Krňoul, Jakub Kanis, and Miloš Železný. Sign language motion capture dataset for data-driven synthesis. In *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*, pp. 101–106, 2020.

Pavel Jedlička, Zdeněk Krňoul, Miloš Železný, and Luděk Müller. MC-TRISLAN: A large 3D motion capture sign language data-set. In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, pp. 88–93, 2022.

Tao Jiang, Peng Lu, Li Zhang, Ningsheng Ma, Rui Han, Chengqi Lyu, Yining Li, and Kai Chen. RTMPose: Real-time multi-person pose estimation based on MMPose. *arXiv preprint arXiv:2303.07399*, 2023.

Nabeela Khan, Bowen Wu, Runwu Shi, Benjamin Yen, Takeshi Ashizawa, Carlos Toshinori Ishi, Takashi Minato, and Kazuhiro Nakadai. From sign language generation to humanoid execution: Vision-language guided retargeting with collision mitigation. *arXiv preprint arXiv:2607.17769*, 2026.

Kaustubh Kundu, Hrishav Bakul Barua, Lucy Robertson-Bell, Zhixi Cai, and Kalin Stefanov. DexAvatar: 3D sign language reconstruction with hand and body pose priors. In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5842–5852. IEEE, 2026.

Zecheng Li, Wengang Zhou, Weichao Zhao, Kepeng Wu, Hezhen Hu, and Houqiang Li. Uni-Sign: Toward unified sign language understanding at scale. In *International Conference on Learning Representations*, volume 2025, pp. 14611–14630, 2025.

Qiayuan Liao, Takara E Truong, Xiaoyu Huang, Yuman Gao, Guy Tevet, Koushil Sreenath, and C Karen Liu. BeyondMimic: From motion tracking to versatile humanoid control via guided diffusion. *Science Robotics*, 11(117):eadx8924, 2026.

Zhengyi Luo, Jinkun Cao, Josh Merel, Alexander Winkler, Jing Huang, Kris M. Kitani, and Weipeng Xu. Universal humanoid motion representations for physics-based control. In *International Conference on Learning Representations*, 2024.

Chaoyi Pan, Changhao Wang, Haozhi Qi, Zixi Liu, Homanga Bharadhwaj, Akash Sharma, Tingfan Wu, Guanya Shi, Jitendra Malik, and Francois Hogan. SPIDER: Scalable physics-informed dexterous retargeting. *arXiv preprint arXiv:2511.09484*, 2025.

Xue Bin Peng, Pieter Abbeel, Sergey Levine, and Michiel Van de Panne. DeepMimic: Example-guided deep reinforcement learning of physics-based character skills. *ACM Transactions on Graphics*, 37(4):1–14, 2018.

Xue Bin Peng, Ze Ma, Pieter Abbeel, Sergey Levine, and Angjoo Kanazawa. AMP: Adversarial motion priors for stylized physics-based character control. *ACM Transactions on Graphics*, 40 (4):1–20, 2021.

Xue Bin Peng, Yunrong Guo, Lina Halper, Sergey Levine, and Sanja Fidler. ASE: Large-scale reusable adversarial skill embeddings for physically simulated characters. *ACM Transactions on Graphics*, 41(4):1–17, 2022.

Guanren Qiao, Sixu Lin, Ronglai Zuo, Zhizheng Wu, Kui Jia, and Guiliang Liu. SignBot: Learning human-to-humanoid sign language interaction. *arXiv preprint arXiv:2505.24266*, 2025.

Yuzhe Qin, Hao Su, and Xiaolong Wang. From one hand to multiple hands: Imitation learning for dexterous manipulation from single-camera teleoperation. *IEEE Robotics and Automation Letters*, 7(4):10873–10881, 2022.

Yuzhe Qin, Wei Yang, Binghao Huang, Karl Van Wyk, Hao Su, Xiaolong Wang, Yu-Wei Chao, and Dieter Fox. AnyTeleop: A general vision-based dexterous robot arm-hand teleoperation system. *arXiv preprint arXiv:2307.04577*, 2023.

Zhenguo Sun, Yibo Peng, Yuan Meng, Xukun Li, Yu Sun, Haojun Jiang, Bo-Sheng Huang, Zhenshan Bing, Xinlong Wang, and Alois Knoll. RobotDancing: Residual-action reinforcement learning enables robust long-horizon humanoid motion tracking. *IEEE Robotics and Automation Letters*, 2026.

Shengeng Tang, Jiayi He, Lechao Cheng, Jingjing Wu, Dan Guo, and Richang Hong. Discrete to continuous: Generating smooth transition poses from sign language observations. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3481–3491. IEEE, 2025a.

Shengeng Tang, Jiayi He, Dan Guo, Yanyan Wei, Feng Li, and Richang Hong. Sign-IDD: Iconicity disentangled diffusion for sign language production. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7266–7274, 2025b.

Shengeng Tang, Feng Xue, Jingjing Wu, Shuo Wang, and Richang Hong. Gloss-driven conditional diffusion models for sign language production. *ACM Transactions on Multimedia Computing, Communications and Applications*, 21(4):1–17, 2025c.

Chen Tessler, Yunrong Guo, Ofir Nabati, Gal Chechik, and Xue Bin Peng. MaskedMimic: Unified physics-based character control through masked motion inpainting. *ACM Transactions on Graphics*, 43(6):1–21, 2024.

Yueh-Hua Wu, Nontawat Charoenphakdee, Han Bao, Voot Tangkaratt, and Masashi Sugiyama. Imitation learning from imperfect demonstration. In *International Conference on Machine Learning*, pp. 6818–6827. PMLR, 2019.

Yashuai Yan, Esteve Valls Mascaro, Tobias Egle, and Dongheui Lee. I-CTRL: Imitation to control humanoid robots through constrained reinforcement learning. *arXiv preprint arXiv:2405.08726*, 2024.

Zhengdi Yu, Shaoli Huang, Yongkang Cheng, and Tolga Birdal. SignAvatars: A large-scale 3D sign language holistic motion dataset and benchmark. In *European Conference on Computer Vision*, pp. 1–19. Springer, 2024.

Songyuan Zhang, Zhangjie Cao, Dorsa Sadigh, and Yanan Sui. Confidence-aware imitation learning from demonstrations with varying optimality. In *Advances in Neural Information Processing Systems*, volume 34, pp. 12340–12350, 2021.

Xiaoyu Zhang, Steven Haener, Varun Madabushi, and Maegan Tucker. Kinodynamic motion retargeting for humanoid locomotion via multi-contact whole-body trajectory optimization. *arXiv preprint arXiv:2603.09956*, 2026.

Hao Zhou, Wengang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. Improving sign language translation with monolingual data by sign back-translation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1316–1325. IEEE, 2021.

Ronglai Zuo, Rolandos Alexandros Potamias, Evangelos Ververas, Jiankang Deng, and Stefanos Zafeiriou. Signs as tokens: A retrieval-enhanced multilingual sign language generator. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 23806–23816. IEEE, 2025.

Appendix

Table of Contents

A Data artifacts and provenance **14**
A.1 Stored versions and time alignment 14
A.2 Repair provenance 15
A.3 Construction coverage and evaluation subsets 15

B Construction and control implementation **16**
B.1 Human-motion repair 16
B.2 Robot platform and collision model 17
B.3 Robot retargeting and geometry repair 18
B.4 Reference-tracking policy and curriculum 20

C Metric definitions and evaluation protocols **22**
C.1 Human-motion continuity and content preservation 22
C.2 Geometry metrics and certification 24
C.3 Execution metrics and replay protocol 26
C.4 Aggregation and failure accounting 27

D Supplementary results **28**
D.1 Complete control results 28
D.2 Tolerance sensitivity 29
D.3 Provenance-stratified tracking error 32

E Qualitative results **33**
E.1 Geometry-repair cases 33
E.2 Hardware replay 34
E.3 Human-motion repair casebook 35

A Data artifacts and provenance

The appendix follows the order of the pipeline. Appendix A describes the stored motion versions, their time alignment, the repair records, and the evaluation subsets. Appendix B gives the construction and control settings, and Appendix C defines the metrics and their aggregation. Appendix D reports the complete control results, the threshold sensitivity, and the provenance-stratified analysis. Appendix E collects geometry-repair cases, hardware replays, and human-repair cases.

A.1 Stored versions and time alignment

Each sequence is identified by its CSL-Daily video and sentence annotation and by the corresponding SOKE SMPL-X fit. Table 4 lists what is stored for each version. Human versions store SMPL-X arrays at the native 30 FPS. Robot versions store the floating base and 53 joint coordinates: 29 body coordinates and 24 hand coordinates, of which six per hand are actuated and the rest follow the finger coupling. Every robot version records its robot model, joint ordering, collision model, and coupling definitions. The Global smoothing output $\mathcal{H}^{\mathrm{global}}$ is stored as a processing baseline. Each version also records the configuration that produced it and whether generation succeeded.

All versions share the source time axis. Missing observations, cropped intervals, and synthetic transitions are marked by masks rather than removed, so paired methods keep the same timestamps. Human metrics and geometry repair use the native 30 Hz samples. Geometry evaluation, policy training, and replay use a common 50 Hz grid with $t = j/50$ strictly below the final source timestamp $(T-1)/30$. On this grid, scalar joint coordinates are interpolated linearly and the floating-base orientation by SLERP; passive finger coordinates are then reconstructed from the coupling, and forward kinematics is recomputed. The human-to-robot part mapping carries the repair records of the left and right arm and hand to the corresponding robot joints. Synthetic transitions between clips carry a separate validity flag instead of a repair record.

Table 4: Stage-wise motion representations and linked records.

State	Motion payload	Linked records
$\mathcal{H}^{\text{raw}}$	SMPL-X root, translation, shape, body, and hand arrays	Video and sentence IDs, source frames and timestamps, coordinate convention
$\mathcal{H}^{\text{fused}}$	Same human-motion schema	Part masks, repair intervals, anchors, operations, observation-support scores
$\mathcal{G}^{\text{direct}}$	Floating base and 53-coordinate robot motion	Robot model, joint ordering, collision model, coupling definitions
$\mathcal{G}^{\text{repair}}$	Same robot-motion schema	Deformation, residual penetration, constraint checks, certification flag

A.2 Repair provenance

For every frame and part group, human repair stores the operation that produced the pose and the fixed observation-support score in Table 5.

Table 5: Operation labels and fixed scores for human-motion repair provenance.

Provenance category	Score
Retained source fit	1.00
Exit transition from a frozen segment	0.70
Two-anchor interpolation	0.50
Single-anchor freezing	0.25
Long-segment freezing	0.10

The score states how directly an operation is supported by the video. It is neither an error probability nor a measure of sign quality: natural stillness and weak visual evidence can both lead to holding, so the frequency of an operation is not an error rate. Geometry deformation, residual penetration, and controller tracking are stored as separate fields. The high-support subset used in Appendix D.3 keeps frames with a score above 0.5, that is, retained fits and exit transitions.

A.3 Construction coverage and evaluation subsets

Every source sequence stays in the manifest with a status for each stage: human-fit validity, human repair, direct retargeting, geometry repair, production validation, conversion to the 50 Hz grid, and control readiness. A sequence that fails at one stage keeps its failure reason, so output, validation, and control coverage have separate denominators.

Of the 20,652 source sequences, 20,649 have repaired human motion (2,461,106 frames at 30 FPS) and direct robot references for the Raw, Global, and Fused inputs. The three excluded sequences, given by their CSL-Daily identifiers, are `S000005_P0004_T00`, `S000007_P0004_T00`, and `S001818_P0000_T00`. Geometry repair yields 20,648 references that pass the production checks (2,460,979 frames). The remaining sequence, `S005379_P0009_T00`, fails the residual check of the quadratic program and has no numerical result; it counts as a failure in coverage and is excluded from paired deformation statistics. Human and Direct coverage is 99.985%, and repaired coverage is 99.981%. On the 50 Hz grid, the geometry evaluation contains 4,074,112 frames and 12,160,392 interior samples per condition. Table 6 lists the evaluation sets, which were selected before the experiment results were inspected.

The 20 motions of $\mathcal{D}_{\text{control}}$ were sampled at random from CSL-Daily and are fixed across all control conditions and seeds. Table 7 lists their identifiers with English translations of the official sentence annotations.

Table 6: Data partitions used by the construction and evaluation pipeline.

Set	Size	Source	Role
$\mathcal{D}_{\text{all}}$	20,652	Full source manifest	Production coverage, corpus statistics, and paired human and geometry evaluation
$\mathcal{D}_{\text{diag}}$	600	Fixed stratified subset	Repair-cue diagnostics (Figure 8)
$\mathcal{D}_{\text{content}}$	75	Recognizer-held-out dev/test sequences	Content preservation (frozen recognizer)
$\mathcal{D}_{\text{control}}$	20	Fixed random sample	Self-target and fixed-reference replay

Table 7: The 20 control motions with English translations of their CSL-Daily sentence annotations.

Sequence ID	Sentence-level meaning (English translation)
S000374_P0004_T00	I go to bed early too, at ten every night. Going to bed early is good for your health.
S000443_P0000_T00	Take a break.
S000748_P0004_T00	Take the bus to the train station from across the road.
S000919_P0008_T00	What tasty dishes does your restaurant have? Please recommend some.
S001518_P0004_T00	Why do you drink coffee at night?
S001836_P0000_T00	Only by delighting the eye and mind can works of art move and educate people.
S001883_P0000_T00	The enemy always overestimates its own strength.
S002075_P0000_T00	Maintain broad contact with the public; do not isolate yourself.
S002219_P0000_T00	This week, the company will announce my proposed paid-leave plan for low-income mothers and fathers.
S002297_P0000_T00	Strengthen composition teaching to improve students' written and oral expression.
S003335_P0000_T00	How could you forget so quickly what the teacher just explained?
S003637_P0001_T00	I saw many colorful flowers in the park, which were worthy of admiration.
S003742_P0001_T00	He has plenty of time to deal with this matter.
S003956_P0002_T00	The teacher severely criticized me for making a mistake.
S004575_P0008_T00	Feel proud of being able to help others.
S004729_P0003_T00	My phone was stolen on the subway.
S004989_P0007_T00	Do your own work well and ignore other people's ridicule.
S005258_P0006_T00	He was the only patient in the emergency room, so many doctors came.
S005432_P0009_T00	Sichuan opera, as its name suggests, is a local opera popular in Sichuan.
S006044_P0000_T00	Your PowerPoint presentation has no tables at all; it is all text.

B Construction and control implementation

B.1 Human-motion repair

The image cue uses the RTMPose keypoints of Section 3.2: hand-keypoint confidence, wrist, elbow, and shoulder visibility, hand bounding boxes, blur, and temporal changes in keypoints and boxes. The motion cue uses changes and oscillations of the fitted rotations, with thresholds of 0.30 rad for arms and 0.70 rad for hands between adjacent 30 FPS samples. Masks are formed separately for the left arm, left hand, right arm, and right hand and expanded by one frame before repair.

Anchors are selected from unflagged frames of the same part group. A gap of at most 10 frames with an anchor on each side is filled by spherical interpolation. If the resulting per-frame step would exceed 90% of the group threshold, the right anchor moves one unflagged frame further at a time, up to twice the gap length, and the extension is recorded. A gap of at most 10 frames with a single anchor holds the available pose. A longer interior gap holds the preceding reliable pose and blends into the next reliable observation over its last four frames. A long gap at the start of a sequence holds the medoid pose of the flagged segment and blends into the first reliable frame in the same way, and a long gap at the end holds the last reliable pose. Gaps without any anchor are

left unrepaired and marked as such. After repair, arm windows that still contain large steps receive light smoothing. A window is detected where the largest collar, shoulder, or elbow step exceeds the larger of 0.08 rad and the sequence median plus three scaled median absolute deviations; it is expanded by five frames, smoothed in rotation space with a Gaussian of $\sigma = 4$ frames, and blended into the unsmoothed motion with $\sigma = 2$ frames. Wrist and finger rotations are not smoothed after repair.

The Global baseline applies Gaussian smoothing in rotation space to the whole sequence, with $\sigma = 1.75$ frames for the body and arms and $\sigma = 0.8$ frames for the hands. Root, translation, and timing are unchanged, and local repair is disabled, so Global represents sequence-wide smoothing.

Figure 8 shows that the two cues are complementary. On the fixed 600-sequence diagnostic set, it compares image-cue rates on frames flagged by a diagnostic motion rule, which uses the larger of the fixed threshold and the sequence median plus three scaled median absolute deviations, with the rates on unflagged frames. Out-of-frame observations are no more frequent on flagged arm frames (19.5% in both) and less frequent on flagged hand frames (14.9% versus 19.7%); missing or low-confidence keypoints behave similarly. Blur, bounding-box instability, and keypoint jumps are more frequent on flagged frames. No image cue fires on about 53% of flagged arm frames and 67% of flagged hand frames, so neither cue covers the frames found by the other. The comparison shows co-occurrence, not cause.

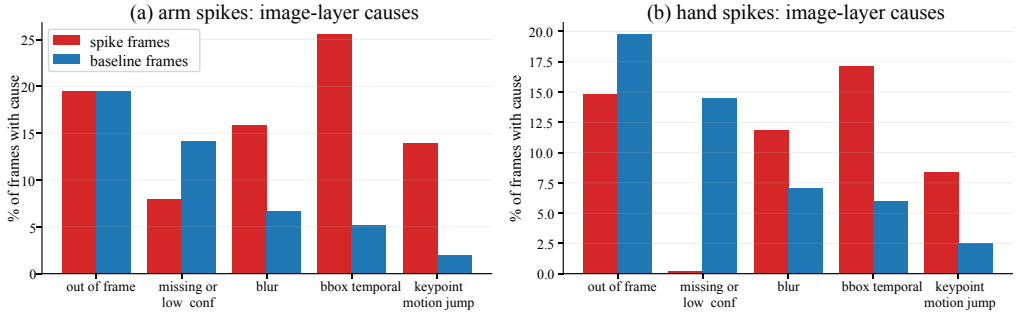

Figure 8: Image-cue rates on frames flagged by the diagnostic motion rule (red) and on unflagged frames (blue) in the fixed 600-sequence diagnostic set.

B.2 Robot platform and collision model

All robot references target a Unitree G1 humanoid with two Inspire RH56DFTP hands (Figure 9). The G1 body has 29 revolute degrees of freedom: 12 in the legs, 3 in the waist, and 7 in each arm (shoulder pitch, roll, and yaw; elbow; wrist roll, pitch, and yaw). Each hand has 6 actuated joints and 6 passive joints driven by linear couplings, so the assembled model has 53 revolute joints, 41 actuated coordinates, and 12 coupling constraints (Table 8). The same model is used for retargeting, geometry repair, the collision audit, and simulated control.

Table 8: Kinematic structure of the G1 + RH56DFTP platform. Hand ranges are given per hand.

Group	DoF	Joints and ranges
Legs	12	Hip pitch, roll, yaw; knee; ankle pitch, roll (per leg)
Waist	3	Yaw, roll, pitch
Arms	14	Shoulder pitch, roll, yaw; elbow; wrist roll, pitch, yaw (per arm)
Hands, actuated	12	Thumb rotation $[0, 1.164]$ rad; thumb flexion $[0, 0.586]$ rad; index, middle, ring, and little flexion $[0, 1.438]$ rad
Hands, passive	12	Distal thumb and finger joints coupled to the actuated joints
Total	53	41 actuated coordinates, 12 mimic constraints

Collision proxies. Signed distances between collision proxies are computed with the MuJoCo distance query. Proxies are grouped into left and right hand, arm, and leg, torso, and head. Body

links use their collision meshes, with cylinders at the shoulders and four 5 mm contact spheres per foot. Each hand has 17 finer proxies: three wrist-housing meshes, a palm box, a palm-sensor box, and twelve capsules (radius 7–8 mm, half-length 12.5–25 mm) that bound the finger segments. The coarse proxies capture contact between the hands and the bulky arm and torso casings, and the capsules capture the narrow gaps between fingers.

Audited pair set. $\mathcal{K}_{\text{body}}$ pairs every hand proxy with every proxy of the other hand and of the arms, legs, torso, and head; $\mathcal{K}_{\text{intra}}$ pairs the proxies within each hand. Three rules remove pairs whose contact is structural rather than a motion error: pairs on the same rigid assembly, pairs on adjacent articulated assemblies, and the two overlapping wrist-joint housings. The audit contains 1,519 pairs (Table 9).

Table 9: Audited collision pairs of the G1 + RH56DFTP model.

Set	Category	Pairs
$\mathcal{K}_{\text{body}}$	Inter-hand	289
	Hand-contralateral arm	136
	Hand-same-side arm	134
	Hand-torso	102
	Hand-head	34
	Hand-leg	612
$\mathcal{K}_{\text{intra}}$	Within each hand	212
$\mathcal{K}_{\text{eval}}$	Total audited	1,519
Excluded	Adjacent assemblies (54), same rigid assembly (6), wrist housings (2)	62

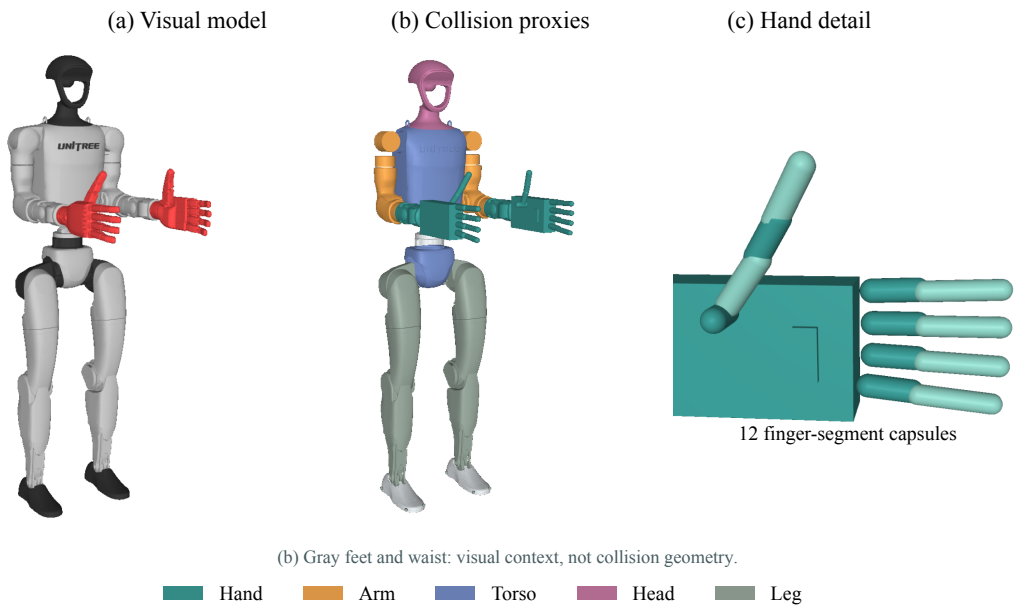

Figure 9: Unitree G1 with Inspire RH56DFTP hands. (a) Visual model. (b) Collision proxies colored by group. (c) Hand detail with the twelve finger-segment capsules; alternating shades distinguish adjacent segments. All views are rendered from the production MuJoCo model.

B.3 Robot retargeting and geometry repair

General Motion Retargeting supplies the body and global wrist motion, and dex-retargeting supplies the wrist-local handshape. Their outputs are combined with common joint clamping and the finger coupling to obtain $\mathcal{G}^{\text{direct}}$. Geometry repair takes this trajectory as input. The arm stage fixes the root, legs, waist, and fingers and optimizes the 14 shoulder, elbow, and wrist coordinates to remove

inter-hand and hand-body penetration. The finger stage then fixes everything except the 12 active finger joints and removes intra-hand penetration within a per-joint budget.

Arm stage. Let $\boldsymbol{x}_t \in \mathbb{R}^{14}$ be the arm coordinates, $\boldsymbol{x}^{(0)}$ their projected initialization, and $\mathcal{E}_{\text{track}}$ and $\mathcal{E}_{\text{rel}}$ the terms of Equations 4 and 5, with $W_b = \text{diag}(10I_3, 5I_3)$ for shoulders and elbows and $\text{diag}(10I_3, 20I_3)$ for wrists (position, then rotation). Let $\boldsymbol{r}_t$ stack the two wrist pose residuals and $W_c = \text{diag}(100I_3, 20I_3, 100I_3, 20I_3)$. The wrist-residual continuity term is

$$\mathcal{E}_{\text{cont}}(\boldsymbol{x}) = \sum_{t=0}^{N-2} \|W_c(\boldsymbol{r}_{t+1} - \boldsymbol{r}_t)\|_2^2 + \sum_{t=0}^{N-3} \|2W_c(\boldsymbol{r}_{t+2} - 2\boldsymbol{r}_{t+1} + \boldsymbol{r}_t)\|_2^2, \tag{14}$$

and the full arm-stage problem summarized by Equation 3 is

$$\begin{aligned} \min_{\boldsymbol{x},\boldsymbol{s}} \quad & \sum_t (\mathcal{E}_{\text{track}}(\boldsymbol{x}_t) + \lambda_r \mathcal{E}_{\text{rel}}(\boldsymbol{x}_t)) + \lambda_v \|D_1\boldsymbol{x}\|_2^2 + \lambda_a \|D_2\boldsymbol{x}\|_2^2 + \lambda_j \|D_3\boldsymbol{x}\|_2^2 \\ & + \lambda_c \mathcal{E}_{\text{cont}}(\boldsymbol{x}) + \lambda_0 \|\boldsymbol{x} - \boldsymbol{x}^{(0)}\|_2^2 + \lambda_s \|\boldsymbol{s}\|_2^2 \\ \text{s.t.} \quad & \boldsymbol{x}_{\min} \le \boldsymbol{x}_t \le \boldsymbol{x}_{\max}, \quad |D_1\boldsymbol{x}| \le \boldsymbol{v}\Delta t, \quad |D_2\boldsymbol{x}| \le \boldsymbol{a}\Delta t^2, \quad |D_3\boldsymbol{x}| \le \boldsymbol{j}\Delta t^3, \\ & \widehat{d}_{k,\tau}^{(\ell)}(\boldsymbol{x}) + s_{k,\tau} \ge \gamma, \quad s_{k,\tau} \ge 0, \quad (k,\tau) \in \mathcal{A}^{(\ell)} \subseteq \mathcal{K}_{\text{body}} \times \Omega_M. \end{aligned} \tag{15}$$

The trajectory is optimized at 30 Hz with $\Delta t = 1/30$ s. For $\tau = i + m/M \in \Omega_M$ with $i \in \{0, \dots, N-2\}$ and $m \in \{0, \dots, M\}$, $\boldsymbol{x}(\tau) = (1 - m/M)\boldsymbol{x}_i + (m/M)\boldsymbol{x}_{i+1}$. At iteration ℓ, the active set $\mathcal{A}^{(\ell)}$ contains the pairs closer than 8 mm, and $\widehat{d}_{k,\tau}^{(\ell)} = d_k(\boldsymbol{q}_\tau^{(\ell)}) + (\boldsymbol{g}_{k,\tau}^{(\ell)})^\top(\boldsymbol{x}(\tau) - \boldsymbol{x}^{(\ell)}(\tau))$ linearizes the signed distance with its gradient $\boldsymbol{g}_{k,\tau}^{(\ell)}$ with respect to the arm coordinates. The clearance is $\gamma = 3$ mm for every pair. Table 10 lists the coefficients; because the terms keep their physical units, the coefficients are scalings rather than dimensionless trade-offs. The speed, acceleration, and jerk limits are 4 rad/s, 45 rad/s^2, and 300 rad/s^3 for shoulders and elbows and 5.4 rad/s, 70 rad/s^2, and 500 rad/s^3 for wrists.

Table 10: Geometry-repair objective and metric scalings. Squared implementation weights are shown as their effective coefficients in Equation 15.

Quantity	Value	Implementation value	Role
λ_r	144	$w_r = 12$	Inter-wrist displacement residual
λ_v	0.3	0.3	First joint difference
λ_a	8	8	Second joint difference
λ_j	100	100	Third joint difference
λ_c	1	1	Wrist-residual continuity (Equation 14)
λ_0	0.02	0.02	Deviation from the projected initialization
λ_s	2.25×10^8	$w_s = 15000$	Collision slack and nonlinear hinge violation
W_c position/rotation	100/20	100/20	Wrist-residual continuity; second difference uses $2W_c$
Tracking position/rotation	10/5	10/5	Shoulder and elbow pose residuals
Wrist position/rotation	10/20	$10/(5\times4)$	Wrist pose residuals

Each iteration solves a quadratic program in which a $10^{-4}I$ diagonal term stabilizes each per-frame block, and a step is accepted only after the nonlinear objective and all pairs of $\mathcal{K}_{\text{body}}$ are evaluated. An adaptive trust region and separated seeds handle locally trapped configurations. The first attempt starts from the projected direct reference; later attempts may start from the preceding solution or a separated feasible pose. The default solve uses $M = 8$, a failed candidate is retried with $M = 16$ or 32, and the validator uses $\max(16, 2M)$ subdivisions. The validator checks inter-hand and hand-body geometry, joint ranges, speed, acceleration, jerk, finger coupling, quaternion norms, and finite values. The relation term is part of the objective, not an acceptance test. Candidates that fail validation stay in the output record but do not enter $\mathcal{G}^{\text{repair}}$.

Finger stage. The finger stage optimizes the 12 active finger joints $\boldsymbol{q}^f$ (thumb rotation, thumb flexion, and index, middle, ring, and little flexion on each hand) and holds all other coordinates fixed.

The 24 finger coordinates follow the linear coupling $\Phi \in \mathbb{R}^{24\times 12}$ as $\Phi \boldsymbol{q}^f$. With $\Delta \boldsymbol{q}^f = \boldsymbol{q}^f - \boldsymbol{q}^{f,(0)}$ the change from the direct handshape, the full form of Equation 6 is

$$\begin{aligned}\min_{\Delta \boldsymbol{q}^f, \boldsymbol{s}} \quad & \sum_{t=0}^{N-1} \|\Delta \boldsymbol{q}_t^f\|_2^2 + \beta_1 \|D_1 \Delta \boldsymbol{q}^f\|_2^2 + \beta_2 \|D_2 \Delta \boldsymbol{q}^f\|_2^2 + \beta_3 \|D_3 \Delta \boldsymbol{q}^f\|_2^2 + \rho^2 \sum_{(k,\tau)\in\mathcal{A}^{(\ell)}} s_{k,\tau}^2 \\ \text{s.t.} \quad & |\Delta q_{t,j}^f| \le \theta_{\max}, \quad \boldsymbol{q}_{\min}^f \le \boldsymbol{q}_t^f \le \boldsymbol{q}_{\max}^f, \quad |D_d \boldsymbol{q}^f| \le \boldsymbol{u}_d, \\ & \widehat{d}_{k,\tau}^{(\ell)}(\boldsymbol{q}^f) + s_{k,\tau} \ge \gamma_f, \quad s_{k,\tau} \ge 0, \quad (k,\tau) \in \mathcal{A}^{(\ell)} \subseteq \mathcal{K}_f \times \Omega_M.\end{aligned} \tag{16}$$

$\mathcal{K}_f$ contains every audited pair whose distance depends on $\boldsymbol{q}^f$, namely intra-hand pairs and inter-hand or hand-body pairs with a finger segment. A pair enters $\mathcal{A}^{(\ell)}$ when its distance is below 3 mm, and its linearization is $\widehat{d}_{k,\tau}^{(\ell)}(\boldsymbol{q}^f) = d_k(\boldsymbol{q}_\tau^{f,(\ell)}) + (\boldsymbol{g}_{k,\tau}^{(\ell)})^\top \Phi(\boldsymbol{q}^f(\tau) - \boldsymbol{q}^{f,(\ell)}(\tau))$, where $\boldsymbol{g}_{k,\tau}^{(\ell)}$ is the gradient of d_k with respect to the 24 finger coordinates. The weights $\beta_1, \beta_2, \beta_3$ take the place of $\lambda_v, \lambda_a, \lambda_j$ in the $\mathcal{E}_{\text{smooth}}$ term of Equation 6. The derivative bounds are $\boldsymbol{u}_d = \max(u_{\max}^d \Delta t^d, |D_d \boldsymbol{q}^{f,(0)}|) + 10^{-10}$ for $d = 1, 2, 3$, with $u_{\max}^1 = v_{\max}^f$, $u_{\max}^2 = a_{\max}^f$, and $u_{\max}^3 = j_{\max}^f$ from Table 11, so derivatives already present in the direct reference are not amplified. Once all slack is zero, the solver continues to minimize $\sum_t \|\Delta \boldsymbol{q}_t^f\|_2^2$, which returns the fingers toward the direct handshape. Table 11 lists the parameters.

Table 11: Implementation parameters for the finger stage (Equation 16).

Parameter	Value	Unit / Dimension	Role
$\theta_{\max}$	20.0° (0.34907)	rad	Hard per-joint budget on active finger joints
γ_f	0.2	mm (2×10^{-4} m)	Target geometric clearance margin
β_1	0.5	dimensionless	First difference regularizer (D_1)
β_2	5.0	dimensionless	Second difference regularizer (D_2)
β_3	25.0	dimensionless	Third difference regularizer (D_3)
ρ	10,000	$\rho^2 = 10^8$	Quadratic penalty on penetration slack $s_{k,\tau}$
$v_{\max}^f$	4.8	rad/s	Active finger angular velocity upper bound
$a_{\max}^f$	100.0	rad/s^2	Active finger angular acceleration upper bound
$j_{\max}^f$	2000.0	rad/s^3	Active finger angular jerk upper bound
Trust region step	0.15	rad	Maximum coordinate step per SQP iteration
Subdivision schedule	$\{4, 8, 16\}$	intervals/frame	Progressive continuous collision temporal refinement
Iteration limits	$(24, 32, 48)$	iterations	Max SQP iterations for subdivision rounds $M \in \{4, 8, 16\}$

A sequence is certified when it satisfies Equation 7 on both the 30 Hz grid with 16 subdivisions and the 50 Hz grid with three interior samples. The 50 Hz check is needed because resampling at a non-integer ratio exposes poses between the 30 Hz samples that the native check does not cover. All 20,648 sequences of $\mathcal{G}^{\text{repair}}$ also pass the production checks, and the finger stage leaves the root, waist, and arm coordinates of the arm stage unchanged. Every uncertified sequence completed all three rounds of the subdivision schedule. For the 2,434 uncertified sequences, $\mathcal{G}^{\text{repair}}$ stores the final within-budget result together with per-pair residual records (Appendix C.2). The collision model, pair set, 1 μm tolerance, and 20° budget are the same for all sequences.

B.4 Reference-tracking policy and curriculum

Following BeyondMimic, each policy of our controller is trained on one reference motion. All our control conditions share the residual tracking policy, observations, physics, termination rules, and PPO settings. The actor and critic are ELU MLPs with hidden widths 512, 256, and 128. Training uses 1,024 parallel environments, 24 rollout steps per iteration, two epochs, four minibatches, a fixed learning rate of 3×10^{-5}, discount 0.99, GAE parameter 0.95, clip ratio 0.2, entropy coefficient 2.5×10^{-4}, value coefficient 0.5, maximum gradient norm 0.5, and initial action noise 0.20. Actor and critic inputs are normalized empirically. The simulator uses a 0.005 s step with four steps per 50 Hz control step.

The arm joints, and in the second stage the active finger joints, receive residual actions around their reference targets, clipped to ± 0.10 rad. Each action is scaled per joint by a quarter of the effort limit divided by the stiffness. Finger actions are disabled in the first stage and scaled by 0.02 in the second. Body-joint PD gains follow BeyondMimic, with stiffness $I_a\omega^2$ and damping $2\zeta I_a\omega$ for reflected armature I_a, $\omega = 2\pi \cdot 10$ rad/s, and $\zeta = 2$; the hand joints use a stiffness of 5 N m/rad and a damping of 0.2 N m s/rad. Each episode starts at a time drawn uniformly over the reference, with small perturbations of the root pose (at most 2 cm horizontally, 5 mm vertically, and 0.05 rad in orientation) and of the joint positions (at most 0.03 rad). Episodes last at most 10 s in the first stage and 14 s in the second and end early under the termination rules of the replay (Appendix C.3). Domain randomization, external pushes, and observation noise are disabled.

The reward per control step is

$$R_t = \Delta t_c \sum_{k \in \mathcal{R}} w_k r_k(t), \qquad \Delta t_c = 0.02 \text{ s}. \tag{17}$$

For a set $\mathcal{A}$ of tracked bodies, vector and rotation errors and the tracking kernel are

$$E_{\text{vec}} = \frac{1}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} \|\boldsymbol{y}_{a,t} - \boldsymbol{y}^{\text{ref}}_{a,t}\|_2^2, \qquad E_{\text{rot}} = \frac{1}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} d_{SO(3)}(R_{a,t}, R^{\text{ref}}_{a,t})^2,$$
$$r_{\text{trk}}(E;\sigma) = \exp(-E/\sigma^2). \tag{18}$$

Joint errors E_q use the same mean-squared form over joints, and anchor terms have $|\mathcal{A}| = 1$. Table 12 lists every nonzero term of the two training stages. Single-stage runs use the Stage-2 column throughout; two-stage runs use Stage 1 for the first 1,500 iterations and Stage 2 for the remaining 1,500. The indicator z_t marks frames of a preparation segment. The benchmark references contain only the signing motion, so $z_t = 0$ and the four ready terms are inactive in all reported runs. No torque or joint-acceleration penalty is used.

Table 12: Reward terms used by the reported Stage-1 and Stage-2 tasks. Tracking rows use r_{trk} from Equation 18. A dash denotes an inactive term. The contact term counts configured bodies whose peak force in the sensor history exceeds 1 N.

Term	Unweighted term r_k	σ	$w_k^{(1)}$	$w_k^{(2)}$
Anchor position	$r_{\text{trk}}(\|\boldsymbol{p} - \boldsymbol{p}^{\text{ref}}\|_2^2;\sigma)$	0.50 m	1.00	1.00
Anchor orientation	$r_{\text{trk}}(d^2_{SO(3)};\sigma)$	0.60 rad	0.75	0.75
Body position	$r_{\text{trk}}(E_{\text{vec}};\sigma)$, 14 bodies	0.35 m	3.00	3.00
Body orientation	$r_{\text{trk}}(E_{\text{rot}};\sigma)$, 14 bodies	0.60 rad	1.00	1.00
Body linear velocity	$r_{\text{trk}}(E_{\text{vec}};\sigma)$, 14 bodies	1.00 m/s	0.25	0.25
Body angular velocity	$r_{\text{trk}}(E_{\text{vec}};\sigma)$, 14 bodies	3.14 rad/s	0.25	0.25
Upper-body joint position	$r_{\text{trk}}(E_q;\sigma)$, 17 joints	0.40 rad	1.75	1.75
Leg pose prior	$r_{\text{trk}}(E_q;\sigma)$, 12 joints	0.60 rad	0.25	0.25
Default hand pose	$r_{\text{trk}}(E_q;\sigma)$, 24 joints	0.50 rad	0.25	–
Wrist orientation	$r_{\text{trk}}(E_{\text{rot}};\sigma)$, two wrists	0.45 rad	–	0.30
Active-finger position	$r_{\text{trk}}(E_q;\sigma)$, 12 joints	0.35 rad	–	0.40
Active-finger velocity	$r_{\text{trk}}(E_{\dot{q}};\sigma)$, 12 joints	6.00 rad/s	–	0.05
Action rate	$\|\boldsymbol{a}_t - \boldsymbol{a}_{t-1}\|_2^2$	–	-0.03	-0.03
Joint limit	$\min\Big(10, \sum_j [\ell_j - q_j]_+ + [q_j - u_j]_+\Big)$	–	-0.25	-0.25
Undesired contact	Number of force-threshold violations	–	-0.10	-0.10
Flat orientation	$\|\boldsymbol{g}^b_{xy}\|_2^2$	–	-2.00	-2.00
Base angular velocity	$\|\boldsymbol{\omega}^b_{xy}\|_2^2$	–	-0.50	-0.50
Base linear velocity	$\|\boldsymbol{v}^b_{xy}\|_2^2$	–	–	-0.20
Ready pose	$z_t r_{\text{trk}}(E_{q,29}; 0.35)$	0.35 rad	–	1.00
Ready joint velocity	$z_t \sum_{j \in \mathcal{J}_{29}} \dot{q}_j^2$	–	–	-0.10
Ready base linear velocity	$z_t \|\boldsymbol{v}^b\|_2^2$	–	–	-0.20
Ready base angular velocity	$z_t \|\boldsymbol{\omega}^b\|_2^2$	–	–	-0.20

The 14 tracked bodies are the pelvis, torso, and the left and right hip-roll, knee, ankle-roll, shoulder-roll, elbow, and wrist-yaw links. The upper-body joint set contains the three waist joints and the 14

arm joints, and the ready set adds the 12 active finger joints. The leg prior covers the 12 leg joints. The undesired-contact term excludes the ankle-roll, wrist, hand, base, and force-sensor links.

The actor receives $\boldsymbol{o}_t^{\pi} \in \mathbb{R}^{274}$ and the critic $\boldsymbol{o}_t^{V} \in \mathbb{R}^{406}$ (Table 13). All blocks come from the current control step, with no history, lookahead, phase clock, or observation noise. Orientations use the first two columns of the rotation matrix. The 12 passive finger coordinates keep their slots in the measured joint blocks but are set to zero, because their values follow from the active joints.

Table 13: Actor and critic observation spaces. A dash denotes a block absent from that input. Reference position and velocity are the current 53-coordinate targets on the 50 Hz control grid.

Observation block	Actor	Critic
Reference joint positions and velocities	106	106
Reference-anchor relative position	–	3
Reference-anchor relative orientation	6	6
Tracked-body positions (14×3)	–	42
Tracked-body orientations (14×6)	–	84
Base linear velocity	–	3
Base angular velocity	3	3
Joint position relative to default	53	53
Joint velocity relative to default	53	53
Previous action	53	53
Total	**274**	**406**

Two-stage training tracks the body, arms, and wrists with the default handshape for 1,500 iterations, then restores the policy and observation normalizer and adds finger tracking for another 1,500. Every condition, including single-stage training, rebuilds its optimizer at iteration 1,500, so all conditions use 73,728,000 environment steps per policy. In the second stage, the action reference of active finger joint j is ramped from the default pose q_j^0 to the reference:

$$\bar{q}_{n,j} = q_j^0 + \alpha(n)\big(q^{\mathrm{ref}}_{t(n),j} - q_j^0\big), \qquad \alpha(n) = \mathrm{clip}\left(\frac{n-h}{r}, 0, 1\right), \\ h = \max(1, \lfloor 0.035T \rfloor), \qquad r = \max(10, \lfloor 0.21T \rfloor). \tag{19}$$

Here n counts action updates since the episode reset, $t(n)$ is the corresponding reference frame, and T is the number of reference frames on the control grid. All training references have $61 \le T \le 481$, so $h + r < T$. The residual action is added to $\bar{q}_{n,j}$ before joint-limit clipping and passive-finger reconstruction, while observations and finger rewards use the full reference throughout. Because training episodes start at random reference times, the ramp restarts at every episode reset; evaluation applies the full finger reference from the first frame. The no-ramp condition shares the first-stage policy and applies the full finger reference at once.

External AMP baseline. We adapt the public ProtoMotions 3.1 implementation of AMP to the same robot model and to the Raw + Direct, Fused + Direct, and Fused + Geometry references. For each reference track and seed, one policy is trained on all 20 motions with 64 parallel environments, 32 rollout steps, batch size 2,048, a discriminator replay buffer of 25,000, discriminator batch size 2,048, and a full-body discriminator. Training uses the IsaacGym backend of this implementation, with 36,000 updates or 73,728,000 environment steps per run. As for our controller, the trained policies are scored in the common IsaacLab replay of Appendix C.3. The AMP action path clips coupled finger targets to the hard joint limits, whereas the scorer reconstructs the soft-limit coupling; this changes one passive right-thumb reference in two of the 20 motions by at most 0.0205 rad.

C Metric definitions and evaluation protocols

C.1 Human-motion continuity and content preservation

Human-motion metrics are computed per sequence at 30 Hz and then averaged over the 20,649 paired sequences. On each side, the arm group contains the collar, shoulder, elbow, and wrist, and

the hand group the fifteen finger joints. A step is abnormal when the largest within-group rotation exceeds 0.30 rad for arms or 0.70 rad for hands (Equation 8).

Table 14 gives the paired effects. The abnormal steps that remain after Fused repair lie mostly at the edges of the repair mask. For side s, the step from frame t to $t+1$ is a mask edge when $m_t^s \neq m_{t+1}^s$, using the mask recorded before repair. Over the 20,636 sequences with at least one edge (646,083 edges), 1.22% of Fused edges exceed a threshold. All 7,730 of them are arm steps, split evenly between entry and exit (3,865 each). When a short segment is bridged by spherical interpolation, a large rotation between its anchors is spread over the interpolated steps, and the steps into and out of the segment can still exceed the arm threshold.

Table 14: Paired continuity effects over 20,649 sequences. Entries are mean differences with 95% sequence-bootstrap intervals; pp denotes percentage points.

Metric	Fused−Raw	Fused−Global
Arm abnormal steps (pp)	-18.725 $[-18.823, -18.627]$	-1.143 $[-1.175, -1.111]$
Hand abnormal steps (pp)	-10.554 $[-10.619, -10.488]$	-0.140 $[-0.154, -0.126]$

Tables 15 and 16 report how much each method changes the motion. Each unit is one side-frame of the arm or hand group, and its source deviation is the largest within-group rotation distance from Raw; a unit counts as changed when this distance exceeds 0.01 rad (0.57°). The stored repair mask covers 67.3% of side-frames. Outside it, Fused leaves every hand unit unchanged and moves arm units by 0.042 rad (2.4°) on average, whereas Global moves nearly every unit, by 0.077 rad for arms and 0.059 rad for hands. The Fused arm changes outside the mask lie mostly within four frames of a flagged frame, where exit transitions blend back into the observed motion; 4.3% of these units exceed 0.2 rad. Inside the mask, Fused replaces the source pose with an interpolated or held pose, at mean deviations of 0.525 rad for arms and 0.555 rad for hands. Averaged over all units, Fused therefore changes fewer units than Global (79.75% versus 99.68% for arms and 61.34% versus 85.17% for hands) with larger means (0.3741 and 0.3807 rad), and 96.4% of its arm deviation and all of its hand deviation come from replaced flagged frames.

The path ratio divides the cumulative rotation-path length of each sequence by that of Raw, pooling joint steps over both sides; a pair with both lengths at most 10^{-9} rad has ratio 100%. For side s with largest joint step a_t^s, the step variation is $30^3 \operatorname{mean}_{t,s} |a_{t+2}^s - 2a_{t+1}^s + a_t^s|$, which measures changes in step size rather than angular jerk. Fused shortens rotation paths to 36.49% (arms) and 34.23% (hands) of Raw and lowers hand step variation from 7,495.1 to 1,371.9, below Global's 2,722.4. Because a shorter path could also reflect removed articulation, content preservation is tested separately with the frozen recognizer.

Table 15: Supplementary motion diagnostics on all 20,649 pairs. Path ratio is rotation-path length relative to Raw; source deviation is measured from Raw.

Group / condition	Changed units (%)	Path ratio (%)	Step variation	Source dev. (rad)
Arms / Raw	0.00	100.00	5618.9	0.0000
Arms / Global smoothing	99.68	47.21	585.4	0.1391
Arms / Fused repair	79.75	36.49	1037.3	0.3741
Hands / Raw	0.00	100.00	7495.1	0.0000
Hands / Global smoothing	85.17	58.95	2722.4	0.1134
Hands / Fused repair	61.34	34.23	1371.9	0.3807

The content-preservation evaluation applies one frozen Uni-Sign model to Raw, Global, and Fused on the same 75 sequences, 38 official development and 37 official test examples excluded from recognizer training. The three conditions share the pose adapter, alignment fitted on Raw, FP32 inference, beam size four, and character-level text. BLEU uses the standard 13a tokenizer. Table 17 gives cumulative corpus BLEU with brevity penalty for all four orders. Paired intervals use 10,000 sequence resamples with seed 20260916 and recompute corpus BLEU from the resampled statistics. Table 18 reports pointwise 95% intervals and 98.75% intervals with a Bonferroni adjustment over the four orders within each contrast. Under this adjustment, Fused exceeds Global on BLEU-1

Table 16: Source deviation from Raw split by the stored repair mask. Supported side-frames lie outside the mask (32.7% of side-frames) and flagged side-frames inside it (67.3%). Deviation means and changed shares pool side-frames over all 20,649 sequences; the last column is the share of the sequence-averaged source deviation in Table 15 that comes from flagged side-frames.

Group / condition	Supported dev. (rad)	Supported changed (%)	Flagged dev. (rad)	From flagged (%)
Arms / Global smoothing	0.0769	99.62	0.1632	82.5
Arms / Fused repair	0.0418	38.94	0.5246	96.4
Hands / Global smoothing	0.0594	99.57	0.1364	83.5
Hands / Fused repair	0.0000	0.00	0.5549	100.0

and BLEU-2, and Global stays below Raw on BLEU-1 and BLEU-2. Every Fused−Raw interval contains zero and lies within ± 2.6 BLEU points, so the recognizer, which was trained on raw motion, detects no content change from repair.

Table 17: All BLEU orders for Raw, Global smoothing, and Fused repair on the same 75-sequence content cohort.

Human motion	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Raw	54.39	41.46	32.88	26.59
Global smoothing	48.65	37.33	30.06	24.50
Fused repair	54.10	41.17	32.64	26.29

Table 18: Paired content effects across all BLEU orders. The final column adjusts for four orders within each contrast.

Contrast / metric	Difference [95% interval]	Four-order interval
Global−Raw / BLEU-1	-5.735 $[-8.869, -2.633]$	$[-9.766, -1.736]$
Global−Raw / BLEU-2	-4.129 $[-7.056, -1.249]$	$[-7.973, -0.441]$
Global−Raw / BLEU-3	-2.825 $[-5.592, -0.080]$	$[-6.460, 0.656]$
Global−Raw / BLEU-4	-2.092 $[-4.713, 0.532]$	$[-5.430, 1.167]$
Fused−Raw / BLEU-1	-0.290 $[-1.420, 0.911]$	$[-1.782, 1.240]$
Fused−Raw / BLEU-2	-0.290 $[-1.781, 1.199]$	$[-2.254, 1.704]$
Fused−Raw / BLEU-3	-0.248 $[-1.917, 1.399]$	$[-2.449, 1.900]$
Fused−Raw / BLEU-4	-0.294 $[-2.014, 1.341]$	$[-2.581, 1.825]$
Fused−Global / BLEU-1	5.445 $[2.361, 8.645]$	$[1.583, 9.440]$
Fused−Global / BLEU-2	3.839 $[0.968, 6.804]$	$[0.249, 7.690]$
Fused−Global / BLEU-3	2.576 $[-0.194, 5.493]$	$[-0.963, 6.289]$
Fused−Global / BLEU-4	1.798 $[-1.002, 4.727]$	$[-1.672, 5.517]$

C.2 Geometry metrics and certification

Geometry uses the frozen robot model and the MuJoCo 3.6.0 signed-distance query over all 1,519 pairs of $\mathcal{K}_{\text{eval}}$ (Appendix B.2). A sample penetrates when its worst-pair depth exceeds $\epsilon = 10^{-6}$ m; depth is reported in millimeters and is zero for separated geometry. Grid and interior samples keep separate denominators, and each quantity in Equation 9 is computed per sequence before averaging. Raw, Global, and Fused motions pass through the same retargeting implementation, joint postprocessing, and finger coupling, so the three Direct conditions differ only in their human input. Fused + Geometry applies both repair stages to the Fused + Direct trajectory. Table 19 gives paired differences from 10,000 sequence-bootstrap resamples (seed 20260916), and Table 20 breaks the Fused result down by pair category.

On the 50 Hz samples, 18,251 sequences (88.39%) are penetration-free. Certification additionally requires clearance on the 30 Hz grid and holds for 18,214 sequences (88.21% of $\mathcal{G}^{\text{repair}}$). Table 21 summarizes construction coverage, certification, deformation, and runtime. Direct fails the temporal

Table 19: Paired effects for the four-condition geometry comparison over all 1,519 pairs in $\mathcal{K}_{\text{eval}}$ and 20,648 matched sequences. Entries give the difference and 95% interval; pp denotes percentage points. Geometry denotes Fused + Geometry.

Metric	Global−Raw (Direct)	Geometry− Fused + Direct	Geometry− Global + Direct
Pen. frames (pp)	-0.637 $[-0.706, -0.570]$	-49.893 $[-50.204, -49.576]$	-52.182 $[-52.434, -51.925]$
Inter-frame pen. (pp)	-0.663 $[-0.730, -0.597]$	-49.902 $[-50.213, -49.585]$	-52.195 $[-52.448, -51.936]$
Mean pen. depth (mm)	-0.383 $[-0.419, -0.347]$	-8.270 $[-8.357, -8.184]$	-10.991 $[-11.081, -10.903]$
Pen.-free clips (pp)	0.140 $[0.082, 0.199]$	87.180 $[86.720, 87.631]$	88.178 $[87.742, 88.624]$

Table 20: Per-category penetration for Fused + Direct and Fused + Geometry ($\mathcal{G}^{\text{repair}}$) at 50 Hz across all 20,648 sequences. All entries are macro percentages of penetrating samples. Interior samples use three fixed points per resampled interval.

	Grid frames (%)		Interior samples (%)	
Pair scope	Direct	Geometry ($\mathcal{G}^{\text{repair}}$)	Direct	Geometry ($\mathcal{G}^{\text{repair}}$)
Inter-hand + hand-body	26.6152	0.0993	26.6520	0.1140
Intra-hand	31.4145	0.2841	31.4028	0.2848
All audited pairs	50.1835	0.2904	50.2044	0.3027

Note: Across all audited pairs, $\mathcal{G}^{\text{repair}}$ lowers penetrating samples by 99.4% relative to Direct on both the grid and interior samples. On the 18,214 certified sequences, no sample penetrates in any scope.

limits in every sequence. Joint deviation is the RMS change over the 14 arm coordinates and grid frames, and relation change is the RMS change of the right-minus-left wrist vector in the torso frame. The repaired means are 0.1886 rad (95% interval $[0.1869, 0.1904]$, about $10.8°$) and 7.955 cm ($[7.877, 8.036]$). The relation change follows from the robot's body: its forearm and wrist links have a 4–5 cm collision radius, so forearm crossings that nearly touch on a human signer must be opened, while $\mathcal{E}_{\text{rel}}$ keeps the wrist vector otherwise close to the human reference. The runtime is concurrent CPU wall-clock time per sequence, with medians of 2.25 and 179.97 s; geometry repair adds 364.48 s on average (95% interval $[358.42, 370.64]$).

Table 21: Construction outcome for the Fused production pair. Counts use all 20,652 planned inputs; deformation and runtime average over the 20,648 completed pairs. RMS measures the change from Direct, and runtime is concurrent CPU wall-clock time per sequence.

Robot reference	Output	Checks passed	Certified	Joint RMS (rad)	Relation RMS (cm)	Runtime (s)
Fused + Direct	20,649	0	0	0.0000	0.000	2.46
Fused + Geometry	20,648	20,648	18,214	0.1886	7.955	366.93

Sample-level clearance. Over all 16,234,504 grid and interior samples, 0.31% of $\mathcal{G}^{\text{repair}}$ samples penetrate, compared with 50.54% for Direct; the sequence averages in Table 20 are 0.29% and 50.18% on the grid. In the 18,214 certified sequences, no sample penetrates on either grid, and the largest recorded depth is 0.000997 mm. The finger stage changes the active joints by a frame-weighted RMS of $1.40°$.

Handshape deformation budget. The per-joint budget $\theta_{\max}$ bounds how far the finger stage may change a handshape. On sequence `S000374_P0004_T00` (Table 7), budgets of $5°$ and $10°$ leave intra-hand overlaps of 14.2 and 13.8 mm, whereas $20°$ clears the sequence with a largest active-joint change of $11.7°$. We use $20°$ for the full corpus. Across the 15,714 certified sequences that required finger changes, the largest active-joint change per sequence has a median of $8.6°$ and a

90th percentile of $16.8°$, and 461 sequences reach the budget. Because the objective minimizes displacement, the frame-weighted RMS stays at $1.40°$. The budget applies to the active joints, and the passive joints follow them through the mechanical coupling.

Uncertified sequences. Table 22 divides the 2,434 uncertified sequences by the check they fail; each sequence belongs to one group. (i) *Control-grid residuals* (762 sequences) pass every 30 Hz check and fail only on the 50 Hz grid, which the finger stage does not sample. (ii) *Intra-hand residuals* (615) keep intra-hand penetration at 30 Hz without contact between parts; 562 of them end at the $20°$ budget on at least one active joint (within 10^{-3} deg). (iii) *Cross-part residuals* (1,057) fail an inter-hand or hand-body check at 30 Hz, usually together with an intra-hand residual; 810 end at the budget. The 2,434 uncertified sequences retain contact on 2.2% of their samples, and 1,424 of them end at the budget. Their mean worst-pair depth, averaged over all their samples including separated ones, is 0.037 mm. Per-pair residual records are released with each sequence.

Table 22: Partition of the 2,434 uncertified sequences by failed validation check. Groups are mutually exclusive. Penetrating samples pool the 50 Hz grid and interior samples over all audited pairs.

Group	Sequences	At $20°$ cap	Penetrating samples (%)
(i) Control-grid only	762	52	0.18
(ii) Native intra-hand	615	562	3.03
(iii) Native cross-part	1,057	810	3.19
All uncertified	2,434	1,424	2.21

C.3 Execution metrics and replay protocol

Table 23 lists the two tolerance profiles. The nominal profile defines the pass thresholds and the ActionScore scales of Section 4.2; the alternative profile is used only for sensitivity analysis (Appendix D.2). A component passes when $e_t^x \le \sigma_x$. Handshape, location, and orientation use the worse hand or the largest active-joint error, and relation uses the right-minus-left wrist vector.

Table 23: Nominal and alternative tolerance profiles. The palm tolerances are about $28.6°$ and $45.8°$.

Component	Per-frame error	Nominal	Alternative
Handshape	Largest absolute error over 12 active joints	0.30 rad	0.50 rad
Location	Larger left/right wrist position error	10 cm	15 cm
Orientation	Larger left/right palm rotation error	0.50 rad	0.80 rad
Relation	Error in the right-minus-left wrist vector	10 cm	15 cm

Component attainment (CA) is the per-component pass rate over all planned frames,

$$\mathrm{CA}_{i,s} = \frac{100}{4T_i} \sum_{t=0}^{T_i-1} A_{i,s,t} \sum_{x \in \{\mathrm{hand,pos,ori,rel}\}} \mathbf{1}[e_t^x \le \sigma_x], \tag{20}$$

so unreached frames count as failures. Table 29 reports the pass rate of each component.

Worked example. Consider 100 planned frames, of which 80 are reached, and normalized errors of 0, 0.5, 1, and 2 for the four components on every reached frame. The per-frame fidelities are 1, 0.8409, 0.5, and 0.0625, giving component scores of 80, 67.27, 40, and 5 and an ActionScore of 48.07. Three components pass on each reached frame, so CA is 60%; JointPass is zero because relation fails throughout. Coverage is 80%, and the trial is unfinished.

Replay. Each replay tracks the complete reference, including the full finger target, from the first frame on the 50 Hz grid, with one recorded state per reference frame and no temporal alignment. The references contain only the signing motion, so every scored frame is a signing frame. Scores use the measured simulation state rather than the commanded target. The torso frame B is the

torso link; wrist positions are taken at the wrist-yaw links and palm orientations at the palm links. The active coordinates of each hand are the two thumb joints and the proximal joints of the index, middle, ring, and little fingers. A trial ends at its first failure: anchor-height error above 0.25 m, a projected-gravity z discrepancy above 0.8, an ankle-height error above 0.25 m, or a numerical failure. Ordinary episode timeouts are disabled, so the reference length determines completion. Tracking metrics exclude the failing frame and all later frames, whereas penetration uses every recorded state, including the failing one. Motion after a reset is never appended to a terminated rollout.

For motion i with frame index $\mathcal{I}_i$, coverage and the reached-frame RMS of component x are

$$\mathrm{Coverage}_{i,s} = \frac{\sum_{t\in\mathcal{I}_i} A_{i,s,t}}{|\mathcal{I}_i|}, \qquad \mathrm{RMS}^{x}_{\mathrm{reached}} = \sqrt{\frac{\sum_{t\in\mathcal{I}_i : A_{i,s,t}=1} (e^x_t)^2}{\sum_{t\in\mathcal{I}_i} A_{i,s,t}}}. \tag{21}$$

An empty reached set gives no value. Hand RMS averages the squared error over the twelve active joints, whereas the handshape pass test uses their largest error. The physical indicator $G_{i,s,t}$ of Equation 12 uses the same 10^{-6} m tolerance over $\mathcal{K}_{\mathrm{eval}}$ as rollout penetration. Rollout penetration counts a recorded frame once if any audited pair penetrates, and unexecuted frames are not counted as collision-free. We also report the share of recorded frames deeper than 1 mm and the mean and peak worst-pair depth, with zero for separated frames. The coupling error (Table 27) is the RMS deviation of the executed passive finger joints from the coupling applied to the executed active joints, with each coupled value clipped to its soft joint range.

For active finger joint j mapped to human hand group $g(j)$, the high-support subset of finger targets is

$$\mathcal{T}_{\mathrm{prov}} = \{(t, j) \in \mathcal{I}_i \times \mathcal{J}_{\mathrm{hand}} \mid c_t^{g(j)} > 0.5\}, \tag{22}$$

where c_t^g is the score of Table 5. It is used only for the diagnostics in Appendix D.3; all benchmark scores use the full index.

Temporal tracking. Velocity and acceleration errors compare executed and reference changes on the 50 Hz grid. For the active hand joints, the torso-relative wrists, and the inter-wrist vector,

$$\delta v_t = \frac{(z^{\mathrm{exec}}_{t+1} - z^{\mathrm{exec}}_t) - (z^{\mathrm{ref}}_{t+1} - z^{\mathrm{ref}}_t)}{\Delta t}, \qquad \delta a_t = \frac{\delta v_{t+1} - \delta v_t}{\Delta t}, \qquad \Delta t = 0.02\,\mathrm{s}. \tag{23}$$

For palms, the angular velocity is $\omega_t = \mathrm{Log}(\widetilde{R}_{t+1}\widetilde{R}_t^{\top})^{\vee}/\Delta t$, and its residual is differenced to give the angular-acceleration error. No smoothing or temporal warping is applied. A velocity window requires both frames to be reached and an acceleration window all three; window coverage uses the planned counts $T_i - 1$ and $T_i - 2$. A constant position offset has zero velocity error, so these metrics complement the position errors.

C.4 Aggregation and failure accounting

The sequence or motion is the unit of inference. Human and geometry quantities are summarized per sequence. For control, our controller has 480 self-target replays (20 motions, seeds 42, 1234, and 777, eight configurations), and AMP has 180 (three reference tracks). Each motion receives equal weight after averaging its seeds. Paired intervals resample the 20 motions as blocks, keeping all conditions and seeds of a motion together, so they describe variation across motions given the three observed seeds. AMP reference-track effects use the same procedure. AMP and our controller are trained differently, so their comparison is reported without intervals; every AMP replay stops early, whereas our controller completes every replay on the repaired reference. Human and geometry intervals resample sequences.

Every result distinguishes planned, input-valid, generated, trained, and evaluated counts. When a task reference exists, a failure of reference generation, training, or evaluation receives zero ActionScore and counts as unfinished; its conditional errors are reported as missing. Interrupted runs resume under the same run identifier. Numerical instability, repeated falls, and non-convergence are kept as outcomes and are not replaced by another seed or checkpoint.

D SUPPLEMENTARY RESULTS

D.1 COMPLETE CONTROL RESULTS

Table 24 gives the complete results behind Table 3, adding Global smoothing, the no-ramp curriculum, JointPass, and penetration depth. Table 25 reports each seed. Tables 26–30 decompose the same trials into coverage, reached-frame errors, physical diagnostics, component scores, pass rates, and temporal errors; they are views of the same replays, not independent trials. All intervals below use 10,000 paired bootstrap resamples of the 20 motions with the three seeds kept within each motion, and all differences are computed from unrounded values.

Reference processing. With single-stage training, Fused raises finished replays from 34/60 for Raw to 58/60, a paired gain of 40.00 percentage points (95% interval $[23.33, 56.67]$). Raw policies stop earlier, so their reached-frame errors cover shorter intervals (Table 26). Global smoothing also makes references learnable (ActionScore 50.17), but it lowers BLEU-1 and BLEU-2 relative to Raw (Section 5.2), whereas Fused keeps the recognized content; adding geometry repair to Fused gives the highest scores. Under the two-stage curriculum, geometry repair lowers the share of frames deeper than 1 mm from 10.44% to 4.31% (paired change -6.13 percentage points, $[-12.56, -1.29]$) and hand RMS from 0.0878 to 0.0761 rad.

Curriculum. On the fixed Fused + Geometry reference, all three curricula finish all 60 replays. The full curriculum raises ActionScore by 1.52 over single-stage training and by 0.66 over no-ramp training, and it is higher for each of the three seeds (Table 25). Relative to single-stage training, it raises the pass rate of all four components (Table 29). Relative to no-ramp training, it lowers palm RMS by 0.76° ($[-1.49, -0.06]$) and raises all four continuous component scores (Table 28). Both two-stage curricula also lower the hand, wrist, and palm velocity and acceleration errors relative to single-stage training (Table 30); with and without the ramp, these temporal errors are similar.

AMP. Every AMP replay stops after 10–12 recorded frames, a coverage of 5.27–5.28% on all three reference tracks. The paired ActionScore effects are -0.058 for Fused + Direct minus Raw + Direct ($[-0.346, 0.185]$), $+0.117$ for Fused + Geometry minus Fused + Direct ($[-0.014, 0.255]$), and $+0.059$ for Fused + Geometry minus Raw + Direct ($[-0.180, 0.274]$).

Table 24: Complete controlled execution results on 20 motions and three seeds. Within each group, rows differ only in the reference; the Fused + Geometry rows of (a)–(c) differ only in the curriculum. Shading marks our full pipeline, and bold marks the best value among our controllers.

Controller / reference	ActionScore ↑	JointPass (%) ↑	Finished (%) ↑	Pen. (%) ↓	Depth (mm) ↓
(a) *Ours, single-stage*					
Raw + Direct	34.99	7.50	56.7	36.15	0.548
Global + Direct	50.17	12.46	100.0	27.67	0.309
Fused + Direct	49.15	13.56	96.7	25.97	0.865
Fused + Geometry	52.58	16.32	100.0	16.09	0.138
(b) *Ours, two-stage + ramp*					
Raw + Direct	35.51	7.03	60.0	35.53	0.539
Fused + Direct	49.29	12.25	98.3	26.08	0.663
Fused + Geometry	**54.10**	**17.98**	**100.0**	**13.03**	0.112
(c) *Ours, two-stage without ramp*					
Fused + Geometry	53.44	17.10	100.0	13.55	**0.104**
(d) *AMP, native multi-motion training*					
Raw + Direct	2.73	0.71	0.0	32.24	1.604
Fused + Direct	2.68	0.49	0.0	26.39	3.038
Fused + Geometry	2.79	0.96	0.0	23.50	0.475

Unreached frames score zero. Pen. and Depth use recorded states and all audited pairs, including intra-hand pairs.

Table 25: Per-seed control results on the same 20 motions. The main table averages seeds within each motion before averaging motions. Finished measures replay completion; ActionScore also accounts for reached-frame fidelity.

	ActionScore ↑			Finished (%) ↑		
Reference	42	1234	777	42	1234	777
(a) Single-stage training						
Raw + Direct	30.00	36.93	38.05	45	60	65
Global + Direct	49.27	50.30	50.92	100	100	100
Fused + Direct	50.13	49.46	47.85	100	100	90
Fused + Geometry	52.20	52.03	53.51	100	100	100
(b) Two-stage training						
Raw + Direct	34.82	36.00	35.70	65	60	55
Fused + Direct	49.36	49.84	48.68	95	100	100
Fused + Geometry, no ramp	53.46	53.27	53.60	100	100	100
Fused + Geometry	54.05	53.74	54.51	100	100	100

Table 26: Reached-frame tracking errors and deeper penetration for the same trials as Table 24. Coverage exposes the interval on which conditional errors are measured. The fixed Fused + Geometry curriculum comparison has complete coverage in every condition.

	Coverage	Reached-frame RMS ↓			Pen. > 1 mm
Reference	(%) ↑	Hand (rad)	Wrist (cm)	Palm (°)	(%) ↓
(a) Single-stage training: reference processing					
Raw + Direct	68.8	0.0896	13.02	48.50	15.69
Global + Direct	100.0	0.0821	14.45	43.65	13.08
Fused + Direct	98.4	0.0926	14.55	43.81	10.51
Fused + Geometry	100.0	0.0792	13.97	40.97	5.51
(b) Two-stage training: reference processing and ramp ablation					
Raw + Direct	72.1	0.0859	13.47	48.20	15.74
Fused + Direct	98.7	0.0878	14.44	43.62	10.44
Fused + Geometry, no ramp	100.0	0.0770	13.59	39.96	3.94
Fused + Geometry	100.0	0.0761	13.42	39.19	4.31

Hand RMS is the square root of the mean squared error over the twelve active finger joints and reached frames. Wrist and palm RMS use the larger bilateral error per frame. All RMS errors use reached frames before the first failure; no unexecuted suffix is imputed. The > 1 mm rate uses all recorded physical frames. These conditional diagnostics supplement the planned-interval action scores in the main table.

D.2 Tolerance sensitivity

Under the alternative profile of Table 23, the full method reaches an ActionScore of 73.11 and a JointPass of 53.67%, compared with 54.10 and 17.98% under the nominal profile. The ordering of our control conditions is the same under the two profiles (Tables 28 and 29).

Common threshold scaling. To test orderings between the two profiles and beyond, we scale all four nominal thresholds together by $s \in [0.25, 3]$ and recompute CA and JointPass from the same recorded states, reach masks, planned-frame denominators, collision test, and joint-limit allowance. The sweep uses all 480 replays of our controller and all 180 AMP replays, averaging seeds within each motion and weighting motions equally. For a percentage-valued score $M(s)$, the normalized area is $\int_{0.25}^{3} M(s)\, ds/2.75$ on a 0–100 scale. We integrate the empirical step functions exactly at their breakpoints and check every pairwise ordering at the union of breakpoints and both endpoints.

Fused + Geometry stays above the other references at every scale in panels (a) and (b), for both CA and JointPass (Figure 10, Table 31). In panel (c), the full curriculum has the largest CA and JointPass areas. Conditions with close scores exchange order at some scales: Global + Direct and

Table 27: Replay stability and physical diagnostics for the same trials as Table 24. Means average seeds within each motion and give equal weight to each motion. Depth statistics use the positive depth of the worst audited pair at each executed frame, with zero for separated pairs.

Reference	Replay		Additional diagnostics ↓			Depth (mm) ↓	
	Finished (%) ↑	Coverage (%) ↑	Relation (cm)	Limit (%)	Coupling (rad)	Mean	Peak
(a) Single-stage training: reference processing							
Raw + Direct	56.7	68.8	13.37	0.476	0.0957	0.548	9.21
Global + Direct	100.0	100.0	12.32	0.274	0.0629	0.309	6.33
Fused + Direct	96.7	98.4	11.37	0.307	0.0687	0.865	10.65
Fused + Geometry	100.0	100.0	10.38	0.498	0.0614	0.138	3.29
(b) Two-stage training: reference processing and ramp ablation							
Raw + Direct	60.0	72.1	13.65	0.383	0.0936	0.539	9.00
Fused + Direct	98.3	98.7	11.05	0.323	0.0667	0.663	10.62
Fused + Geometry, no ramp	100.0	100.0	10.11	0.348	0.0613	0.104	3.30
Fused + Geometry	100.0	100.0	10.05	0.188	0.0612	0.112	3.36

Finished is the percentage of complete replays without early termination; it does not test action accuracy. Coverage is reached/planned frames. Nominal JointPass is reported in Table 24. Limit violations use a 10^{-3} rad allowance over recorded physical frames. Relation and coupling are RMS errors over reached frames. Mean depth includes all recorded frames, rather than only penetrating frames. Peak depth is the macro-average of per-trial maximum depths across seeds and motions. Appendix C.3 defines the common protocol.

Table 28: ActionScore decomposition and tolerance sensitivity. The first four columns report component scores under the nominal scales using the complete planned-frame denominator; their mean gives the nominal ActionScore. Alternative-scale ActionScore applies the alternative tolerance profile to the same recorded trajectories. CA reports component-wise hard-threshold attainment under the nominal profile. All columns are higher-is-better.

Reference	Continuous component score				ActionScore		CA
	Hand-shape	Location	Orient.	Relation	Nominal	Alt.	CA nom. (%)
(a) Single-stage training: reference processing							
Raw + Direct	53.55	30.56	22.80	33.05	34.99	47.84	35.94
Global + Direct	81.24	36.11	35.30	48.03	50.17	69.32	50.06
Fused + Direct	77.30	34.50	33.02	51.77	49.15	68.05	48.44
Fused + Geometry	81.34	36.37	36.48	56.13	52.58	71.75	51.78
(b) Two-stage training: reference processing and ramp ablation							
Raw + Direct	57.01	29.97	22.03	33.02	35.51	49.24	35.72
Fused + Direct	78.47	33.83	32.77	52.09	49.29	68.50	48.21
Fused + Geometry, no ramp	82.09	37.13	37.29	57.25	53.44	72.52	52.55
Fused + Geometry	82.51	37.86	38.44	57.59	54.10	73.11	53.11

Fused + Direct in panel (a), the three curricula in JointPass in panel (c), and the three AMP reference tracks.

Articulation-site spacing. On the neutral SMPL-X template, we measure distances between facial landmarks: each eye center averages its six eyelid points, the nose is the nose tip, the mouth center averages the inner-lip midpoints, and the chin is the central contour point. The eye-to-nose distances are 5.8 cm on both sides, the inter-eye distance is 6.3 cm, nose-to-mouth is 4.4 cm, and mouth-to-chin is 4.9 cm. The 10 cm location tolerance therefore exceeds the spacing of some neighboring facial sites; it is a declared operating point rather than a bound derived from anatomical spacing.

Table 29: Hard-threshold execution metrics under the nominal and alternative tolerance profiles. The first four columns report component-wise attainment over all planned frames under the nominal profile, and their average gives nominal CA. Alternative CA and JointPass use the alternative profile. These quantities are percentages and are distinct from the continuous ActionScore reported in Table 24.

Reference	Hand-shape	Location	Orient.	Relation	CA nominal	CA alternative	JointPass alternative
(a) Single-stage training: reference processing							
Raw + Direct	61.26	29.67	18.30	34.53	35.94	54.12	23.47
Global + Direct	92.62	29.89	29.66	48.07	50.06	76.50	37.67
Fused + Direct	86.40	27.77	25.63	53.94	48.44	75.47	41.26
Fused + Geometry	91.05	28.89	28.96	58.21	51.78	80.35	49.95
(b) Two-stage training: reference processing and ramp ablation							
Raw + Direct	65.17	27.64	16.48	33.60	35.72	55.03	23.16
Fused + Direct	87.89	26.65	24.91	53.37	48.21	75.97	40.75
Fused + Geometry, no ramp	91.12	28.84	29.72	60.52	52.55	81.82	52.89
Fused + Geometry	91.68	29.88	31.21	59.68	53.11	82.88	53.67

Table 30: Velocity and acceleration tracking errors relative to the reference. RMS is computed on reached windows per trial, averaged across seeds within each motion and then equally across motions. Window coverage uses the planned denominator. All error columns are lower-is-better.

Reference	Windows (%) ↑	Hand	Wrist	Palm	Relation
Velocity error (hand and palm in rad/s; wrist and relation in m/s)					
(a) Single-stage training: reference processing					
Raw + Direct	68.6	1.042	1.212	11.546	1.324
Global + Direct	100.0	0.741	0.721	5.090	0.816
Fused + Direct	98.4	0.657	0.592	4.489	0.637
Fused + Geometry	100.0	0.517	0.556	4.072	0.616
(b) Two-stage training: reference processing and ramp ablation					
Raw + Direct	71.9	1.016	1.266	11.385	1.373
Fused + Direct	98.7	0.645	0.574	4.454	0.631
Fused + Geometry, no ramp	100.0	0.503	0.546	4.043	0.619
Fused + Geometry	100.0	0.504	0.546	4.033	0.621
Acceleration error (hand and palm in rad/s^2; wrist and relation in m/s^2)					
(a) Single-stage training: reference processing					
Raw + Direct	68.5	40.586	49.558	546.092	51.900
Global + Direct	100.0	20.675	10.513	82.978	11.697
Fused + Direct	98.4	17.866	9.832	85.577	10.135
Fused + Geometry	100.0	13.191	7.591	59.646	8.206
(b) Two-stage training: reference processing and ramp ablation					
Raw + Direct	71.8	39.310	52.192	537.963	54.236
Fused + Direct	98.7	17.293	9.605	85.060	9.982
Fused + Geometry, no ramp	100.0	12.848	7.398	58.572	8.198
Fused + Geometry	100.0	12.991	7.409	58.630	8.231

Table 31: Normalized threshold areas on $[0.25, 3]$, on a 0–100 scale. Each condition contains 20 motions and three seeds. The panel identifiers match Figure 10.

Panel	Condition	CA area	JointPass area
a	Raw + Direct	47.454	21.671
a	Global + Direct	69.305	37.678
a	Fused + Direct	68.044	39.974
a	Fused + Geometry	71.755	46.943
b	Raw + Direct	49.029	22.492
b	Fused + Direct	68.522	39.803
b	Fused + Geometry	73.045	49.898
c	Single-stage	71.755	46.943
c	Two-stage, no ramp	72.498	49.165
c	Two-stage + ramp	73.045	49.898
AMP	Raw + Direct	3.639	1.616
AMP	Fused + Direct	3.475	1.157
AMP	Fused + Geometry	3.583	1.717

D.3 Provenance-stratified tracking error

Because the repair records follow each reference into its rollouts, tracking errors can be grouped by the operation that produced each reference segment. This analysis trains one tracking policy with the architecture of Appendix B.4 on the 20 motions of $\mathcal{D}_{\text{control}}$ joined in a fixed order by synthetic transitions, with seeds 42, 1234, and 777, and is replayed over the whole joined reference. Table 32 reports the joint RMS error of this policy for each operation category and body side, averaged over seeds.

Table 32: Tracking error of the multi-motion policy grouped by reference provenance. Share is the percentage of frames of that side; RMS pools joint-frames and averages the three seeds.

		Right side			Left side		
Operation	Score	Share (%)	Arm RMS (rad)	Hand RMS (rad)	Share (%)	Arm RMS (rad)	Hand RMS (rad)
Retained source fit	1.00	65.43	0.162	0.062	36.08	0.154	0.057
Exit transition	0.70	0.12	0.129	0.054	2.00	0.173	0.059
Two-anchor interpolation	0.50	19.07	0.179	0.070	6.23	0.181	0.070
Single-anchor freezing	0.25	3.52	0.150	0.105	2.16	0.148	0.057
Long-segment freezing	0.10	0.50	0.240	0.061	42.17	0.105	0.065
Synthetic transition	–	11.36	0.153	0.044	11.36	0.113	0.044

The records separate two kinds of held segments. On the right side, which carries most of the signing, long held spans replace motion that the video does not support, and arm error rises from 0.162 rad RMS on retained fits to 0.240 rad. On the left side, long held spans cover 42.17% of frames, since in many control motions the left arm rests for long periods, and such a static pose is tracked easily (0.105 rad). Without the records, both cases would be pooled with observed motion; with them, a controller developer can separate error on held segments from error on observed motion.

The records can also guide training. Score weighting multiplies each per-joint and per-body squared error in the tracking rewards by the score of its part group, so held and interpolated segments contribute less supervision. Hard rejection instead removes the seven motions whose mean arm score is below 0.5 and trains the same policy on the remaining 13. On these 13 motions, the RMS error of the active finger joints on $\mathcal{T}_{\text{prov}}$ (Equation 22) is 0.059 rad with score weighting and 0.066 rad with rejection.

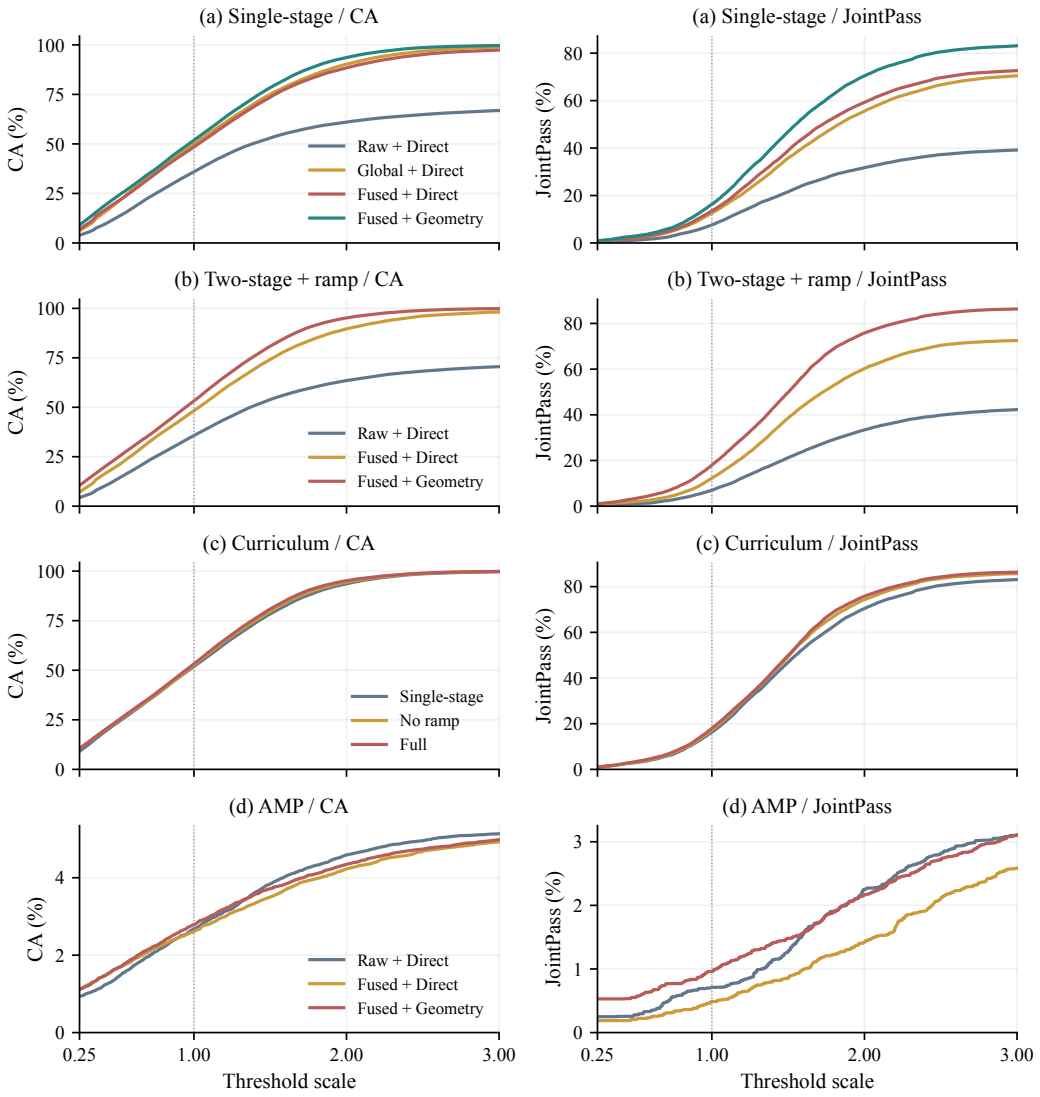

Figure 10: CA and JointPass under common scaling of all four action thresholds. Rows (a,b) compare references under single- and two-stage training, (c) compares the three curricula on Fused + Geometry, and (d) shows AMP (Table 24). Dotted lines mark the nominal scale. AMP uses a separate vertical scale. Curves are drawn on a dense display grid; areas and ordering checks use the exact empirical breakpoints.

E Qualitative results

E.1 Geometry-repair cases

Figure 11 shows the four cases of Figure 6 with their source human fits and full-body robot views: hand-head, hand-leg, inter-hand, and intra-hand penetration. Human fits use the nearest source time, and paired robot views share reference time, camera, and lighting. Corpus-level statistics are in Appendix C.2.

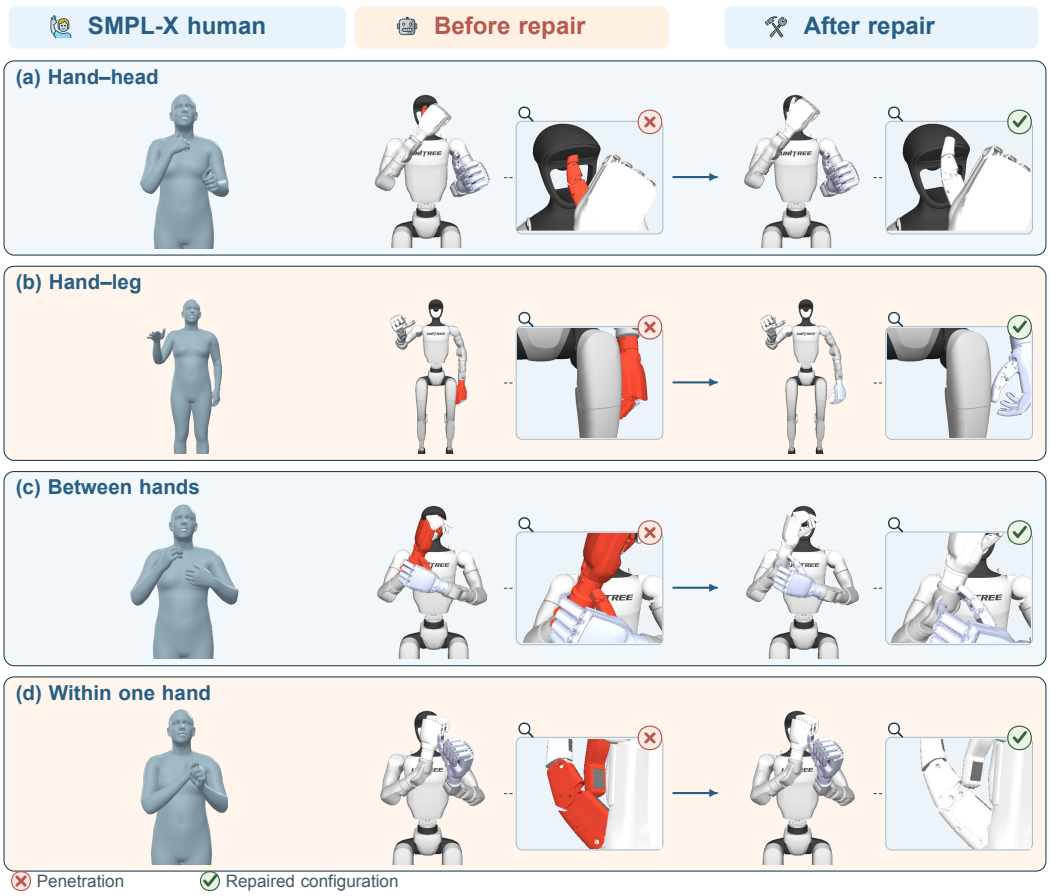

Figure 11: Human fits and robot poses before and after geometry repair. Red links and crosses mark penetration; checks indicate its removal.

E.2 Hardware replay

We recorded seven hardware replays of six motions from $\mathcal{D}_{\text{control}}$ on a Unitree G1 with two Inspire RH56DFTP hands; Figure 12 shows four of them. Each motion is driven onboard by a per-motion tracking policy trained as in Appendix B.4. For replay, the reference is extended with a standing hold and an entry transition, the signing interval is played at about half speed, and an exit transition and a final hold follow. The hardware replays are qualitative and are not scored.

For the figure and the videos, the source video is aligned to each recording by one global time scale and offset, fitted between the upper-body motion in the recording and the wrist and elbow speed of the reference. No local warping is applied, the robot recording keeps its original speed, and the first and last source frames are held during the added preparation and return phases.

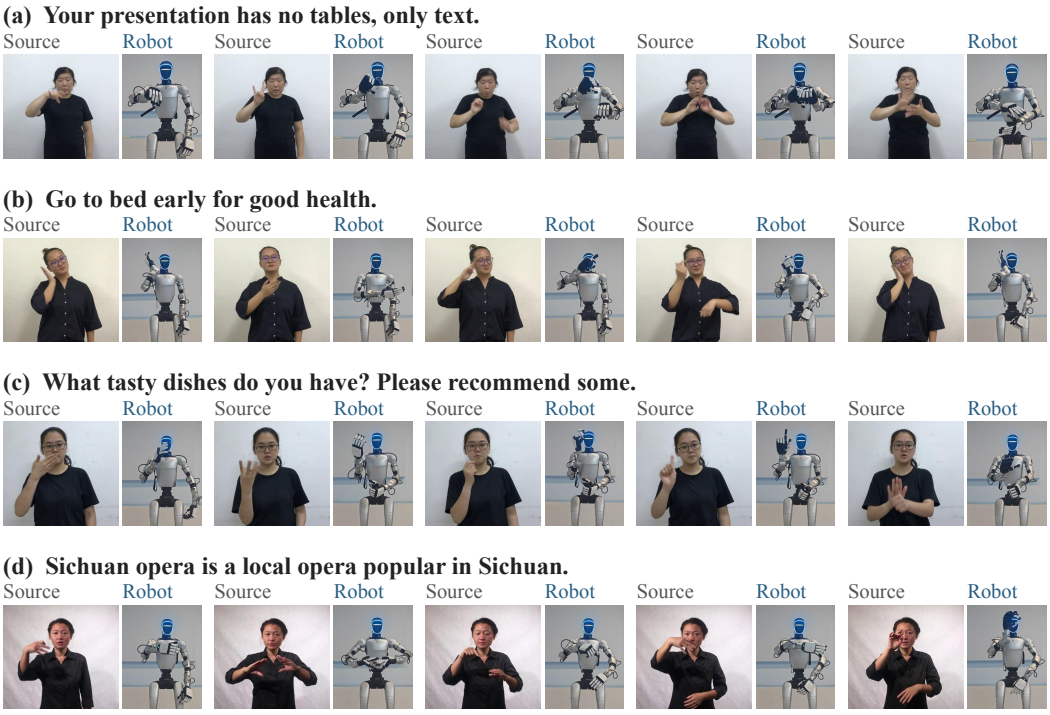

Figure 12: Source signing and hardware replay for four utterances. Each row contains five chronological frame pairs, with the source signer on the left and the real robot on the right. Row labels summarize the annotated utterance meanings. Pairs use the global motion-phase alignment described above; fixed crops retain the upper body and hands while excluding wall lettering.

E.3 Human-motion repair casebook

Figures 13–19 show seven repaired intervals from six source sentences; Figures 13 and 16 come from the same sentence. Each plate aligns source video frames with the SMPL-X fits before and after repair, enlarges the affected part, and plots a per-frame diagnostic at 30 Hz. Letters A–D mark matched instants, and paired meshes share camera and lighting. Gray shading marks repair intervals and sand shading marks exit blends. Figures 13–15 plot the largest hand-joint rotation step, including the steps into and out of the repaired span. Figures 16–19 plot the pelvis-relative wrist displacement per frame, with peaks taken over the displayed window. The Chinese sentences are the official CSL-Daily annotations of the complete utterances, with English translations.

The plates illustrate the repair operations and were produced with earlier settings of the same repair: Figures 13–15 repair only the hand groups, and Figures 16–19 repair the arm and hand groups with the same two cues, with light local smoothing added in Figure 16. The released repair uses the settings of Appendix B.1.

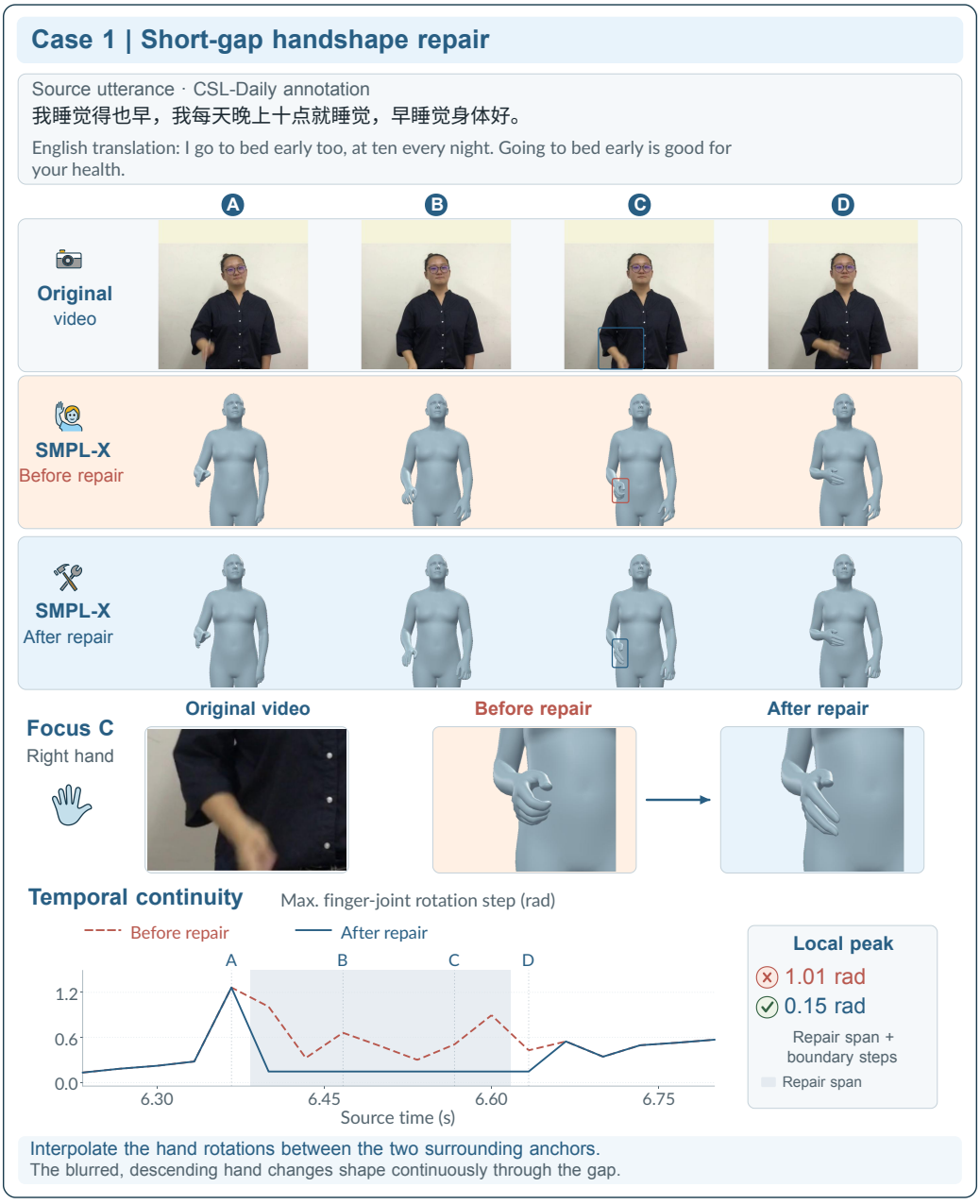

Figure 13: Short-gap interpolation reduces the local hand-rotation step peak from 1.01 to 0.15 rad.

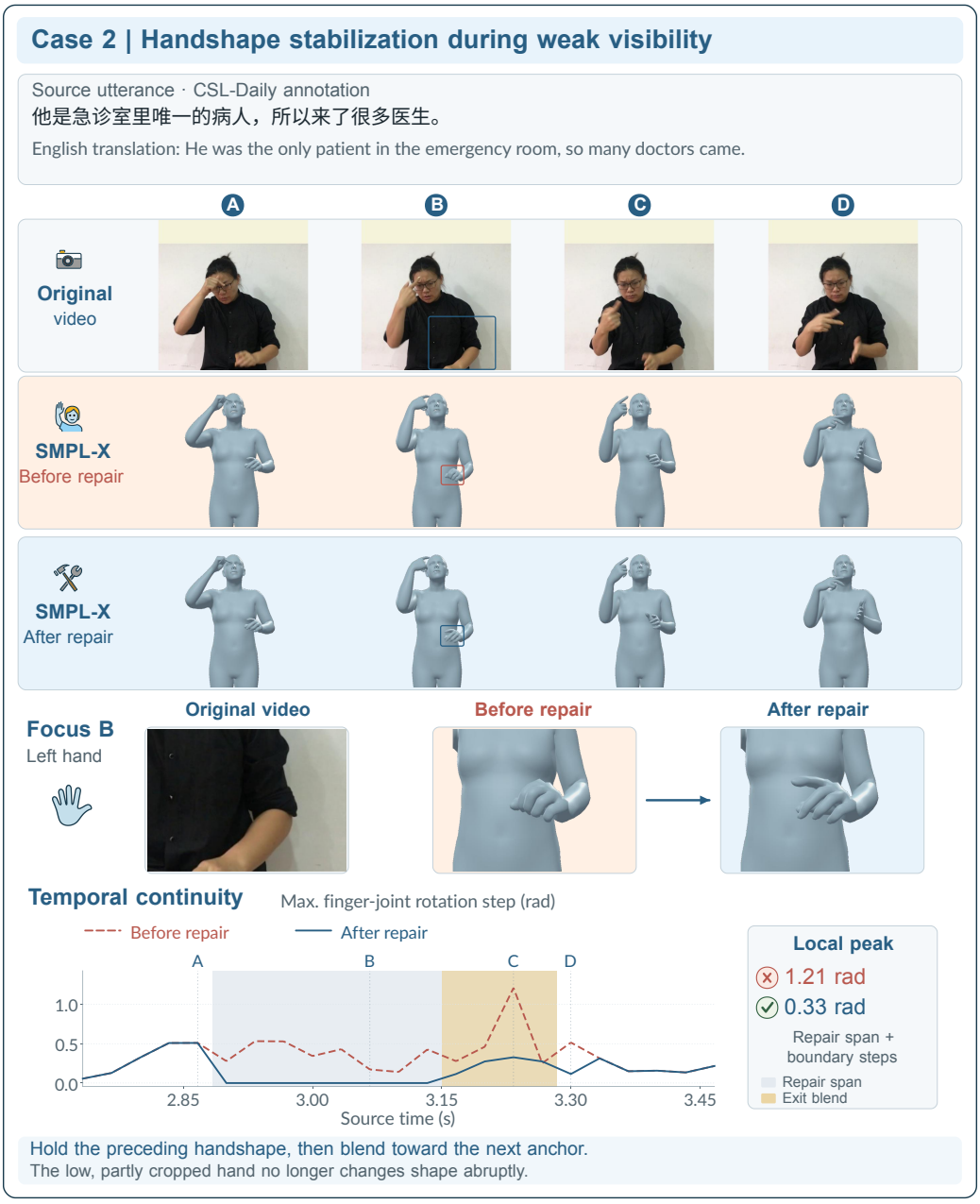

Figure 14: Holding and exit blending stabilize a partly out-of-frame hand, reducing the local rotation-step peak from 1.21 to 0.33 rad.

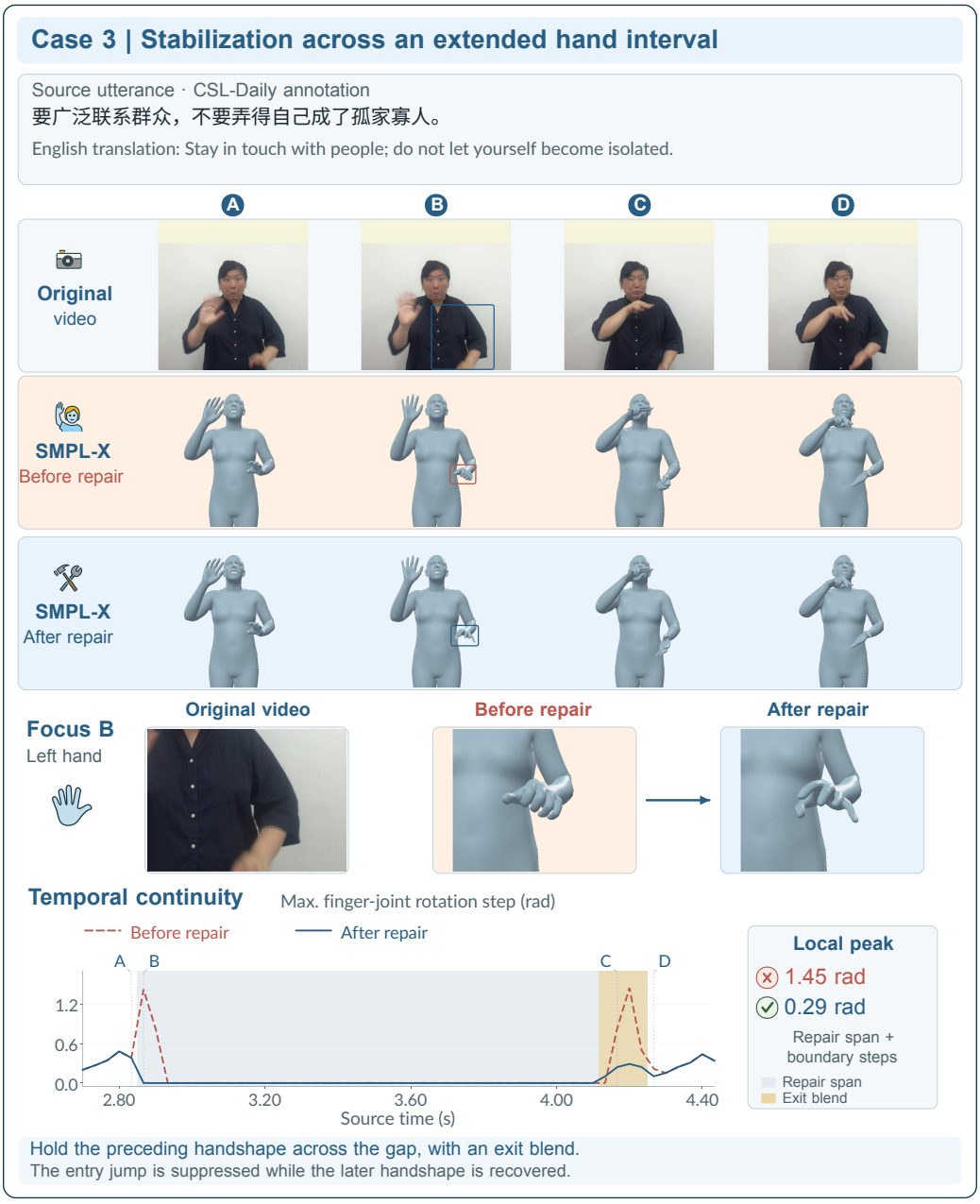

Figure 15: Handshape holding and exit blending bridge an extended interval, reducing the local rotation-step peak from 1.45 to 0.29 rad.

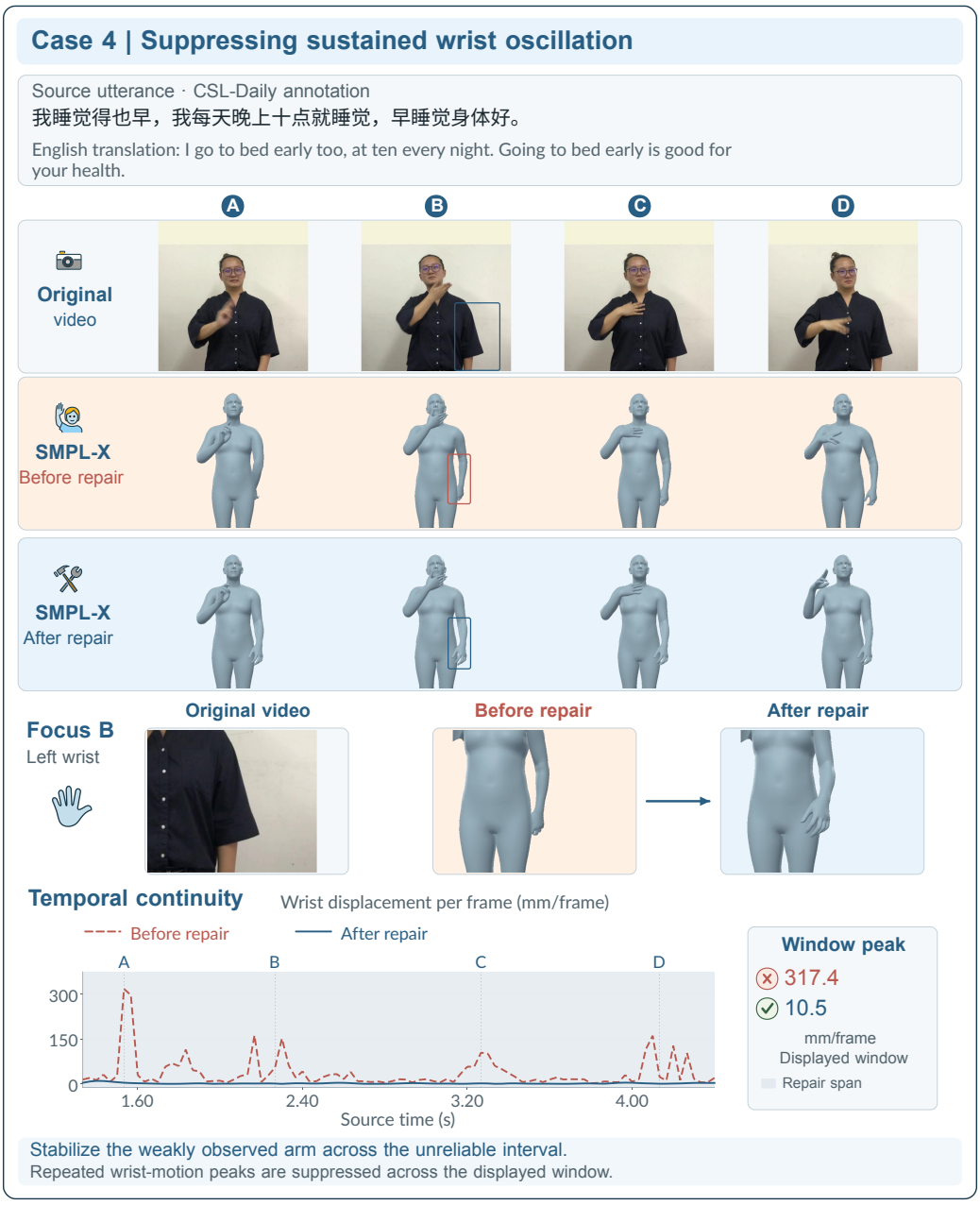

Figure 16: Repair and local smoothing suppress left-wrist jitter, reducing the displayed-window displacement peak from 317.4 to 10.5 mm/frame.

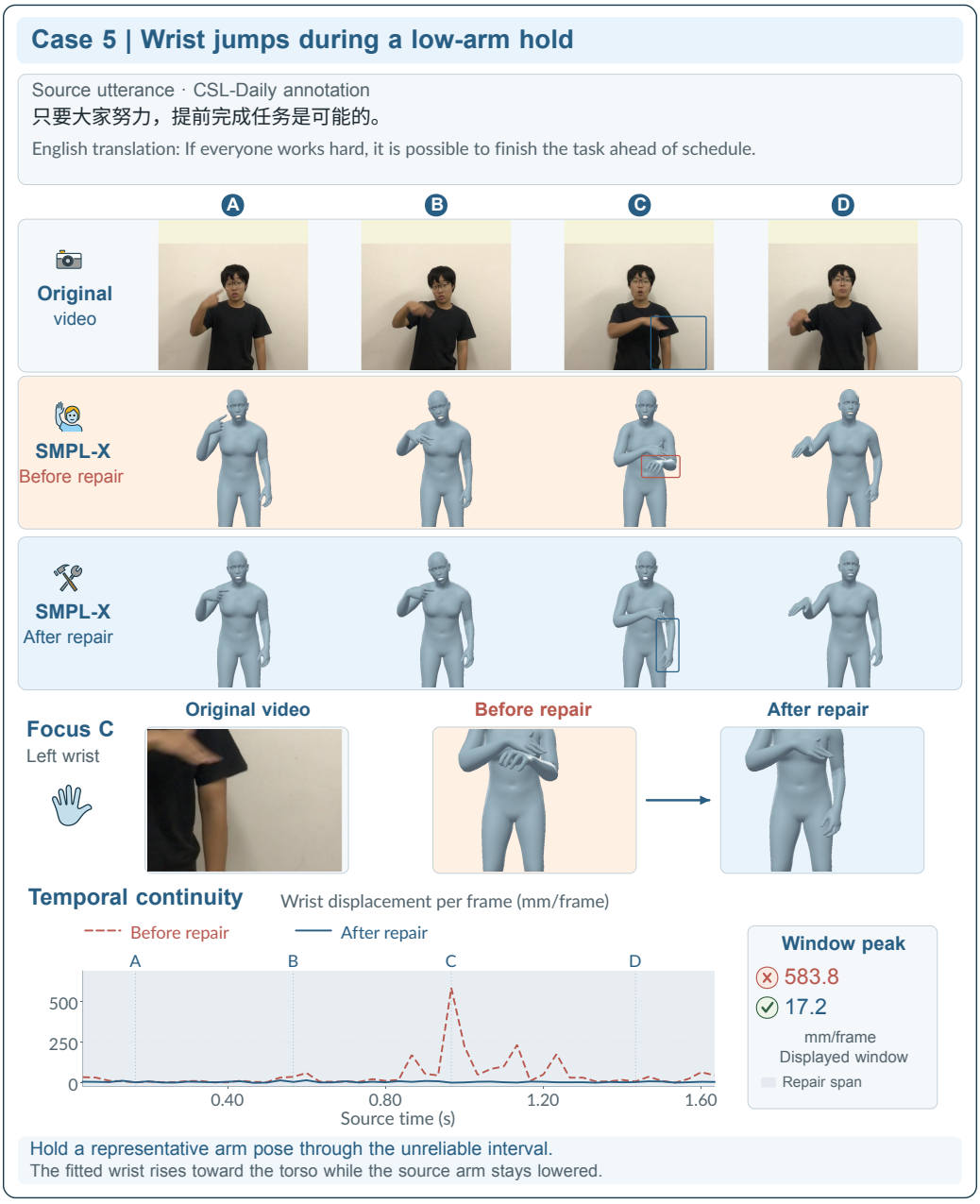

Figure 17: Holding a representative low-arm pose suppresses wrist excursions toward the torso, reducing the displacement peak from 583.8 to 17.2 mm/frame.

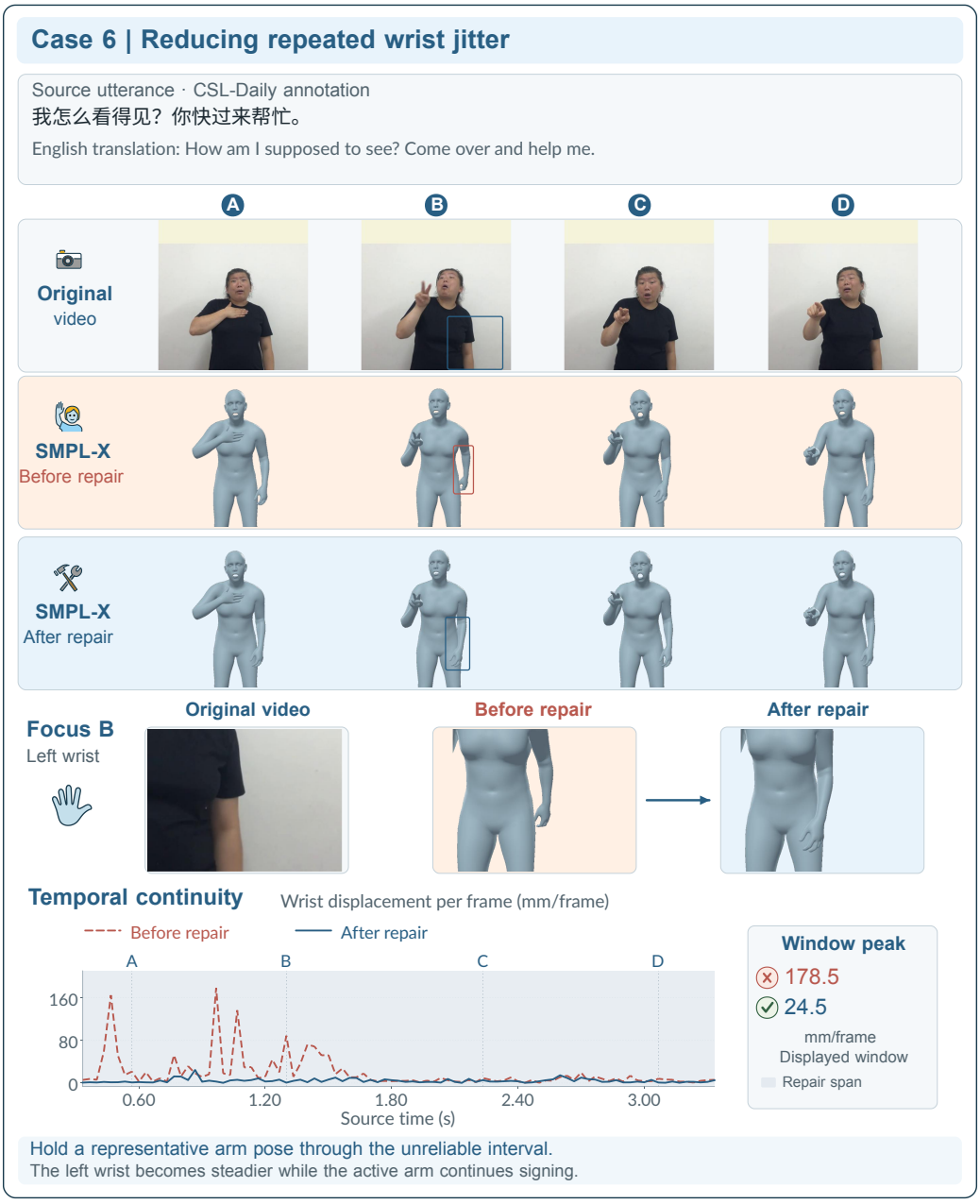

Figure 18: A left-arm hold suppresses wrist jitter while the right arm signs, reducing the displacement peak from 178.5 to 24.5 mm/frame.

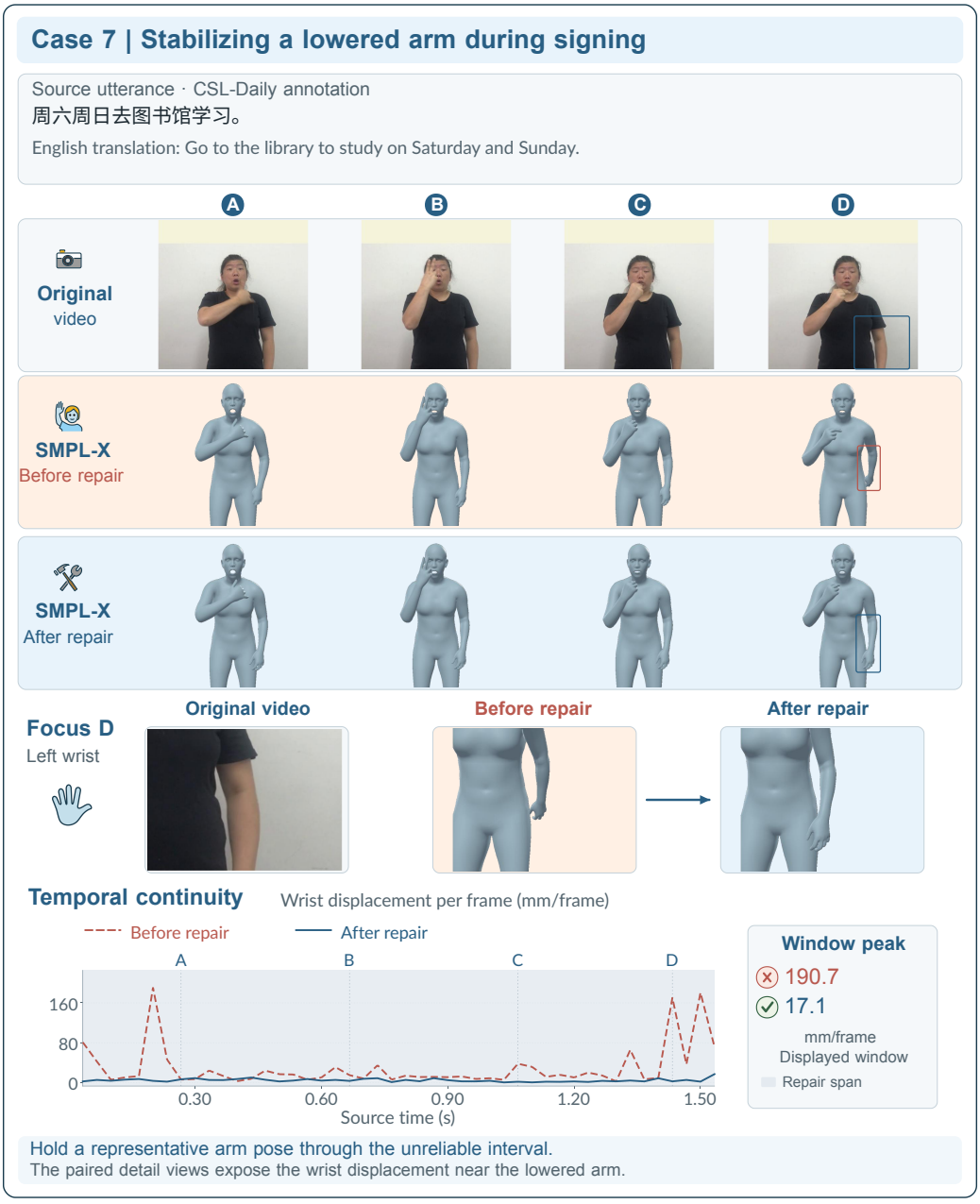

Figure 19: Holding a lowered arm reduces peak wrist displacement from 190.7 to 17.1 mm/frame.