Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

No Humans! Autonomous Development and Execution of a Forensic 3D Workflow by a General-Purpose AI Agent

Dominik Haenni

dominik.haenni@for-zh.ch

Zurich Forensic Science Institute https://orcid.org/0000-0002-2166-7109

Lars Christian Ebert

lars.ebert@for-zh.ch

Zurich Forensic Science Institute https://orcid.org/0000-0002-8215-7006

Constantin Lux

Zurich Institute of Forensic Medicine

Erika Dobler

Zurich Forensic Science Institute

Research Article

Keywords: forensic 3D reconstruction, stature estimation, agentic AI, large language model, Blender, photogrammetry, laser scanning, automation

Posted Date: October 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11216853/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare potential competing interests as follows: Lars Christian Ebert and Dominik Haenni previously participated in OpenAI's external red-teaming activities. This prior activity was unrelated to the present study. The study was conducted independently, and OpenAI had no role in its design, execution, analysis, interpretation, or reporting. The authors declare no other competing interests.

Abstract

Forensic three-dimensional (3D) scene reconstruction combines established scientific methods with substantial specialist software work. We tested whether a general-purpose AI agent could develop and conduct a complex forensic 3D workflow after task specification and assessed the technical correctness of its outputs. OpenAI ChatGPT was given the laboratory quality-management description of an established PhotoModeler/3ds Max stature-reconstruction workflow and used to formulate a requirements-oriented master prompt for the OpenAI Codex programming agent through multiple iterations using the GPT-6 Astra model at a “high” reasoning effort. The supplied dataset comprised a registered laser scan, a reference photograph with markers, and four anonymized images showing a test subject in different body positions, each with both feet in contact with the ground. Codex received the data, had full access to an experimental workstation, and was not supplied with any precomputed camera solutions or body model. It autonomously developed and executed a pipeline for camera reconstruction, fitted an articulated human model, performed measurements, and generated an illustrated report. During the final run, the only investigator intervention during execution was a 'continue' prompt after token allowances were exhausted. The four measured heights were 176.65, 175.96, 161.30, and 169.27 cm, compared with a reference height of 175 cm measured with shoes on. Two estimates met the predefined +/-3 cm target; the mean of 170.8 cm did not. Marked variation in fitted body dimensions and imperfect positioning in additional views indicated unresolved anatomical and depth ambiguity. This single-case experiment demonstrates fully autonomous workflow development and measurement execution, including reporting, while showing that technical completion does not ensure a correct reconstruction. These findings suggest that even complex forensic expert workflows may soon be within reach of fully autonomous AI systems, provided that their outputs are rigorously validated, regulatory requirements are met, and sufficiently capable models and computational resources are available.

1. Introduction

Most current applications of artificial intelligence (AI) in forensic science use AI as a specialized analytical component rather than as an autonomous examiner. Machine-learning and deep-learning systems are usually developed for defined tasks such as classification, identification, pattern recognition, or quantitative prediction. The human expert remains central. The expert formulates the forensic question, selects and adapts the method, applies it, evaluates the output, corrects failures, and interprets the result in the context of the case. Method selection and adaptation are therefore major parts of forensic expertise. Recent forensic literature accordingly treats AI mainly as a tool that supports expert decision-making, with human oversight and validation remaining central requirements [1, 2].

Generative AI and large language models (LLMs) have started to broaden this role. Recent forensic studies have explored generative AI for text-based workflows and digital-forensic assistance [3, 4]. These systems can be useful, but the operating model is still mainly assistive. A human defines the problem, chooses the method and tools, and decides how they should be applied. AI works within that human-

designed process, while the expert reviews the output and remains responsible for the final interpretation. For example, a Netherlands Forensic Institute pilot found practical value for lower-risk text work but distinguished this from forensic reasoning, which requires stronger validation and oversight [3].

This division of work is becoming less clear as general-purpose AI develops from answer-generating models into agents that can handle complex multi-step tasks. Such systems can write and execute code, use technical software, inspect intermediate results, and revise their approach. Outside forensic science, LLM-based systems have already generated procedural 3D workflows and executable Blender scenes [5, 6]. Current frontier agents are also designed for coding, computer use, and end-to-end professional tasks [7]. This raises a broader question than whether AI can automate one forensic subtask: can a general-purpose agent take a complex forensic problem, determine how to solve it, build the required technical workflow, and execute, validate and report it with minimal human intervention?

We chose forensic stature reconstruction from surveillance imagery as a controlled but demanding test case. The underlying geometry is well established. Single-view metrology provides a basis for recovering metric information from images [8], and forensic studies have examined posture and activity [9], laser-scanner-supported height estimation [10], PhotoModeler-based workflows [11], highly distorted surveillance imagery [12], and combined video/3D-laser-scanning approaches [13]. AI-assisted person analysis in a 3D forensic environment has also been explored [14]. The reconstruction itself, however, is underdetermined. One 2D image can be explained by several different 3D body configurations. Changes in depth, posture, joint position, and body proportions can produce similar image projections while yielding different reconstructed statures. A useful solution therefore requires more than a good overlay. It must also be anatomically and biomechanically plausible and, where video is available, consistent with the observed movement. This combination of geometric ambiguity and expert contextual knowledge makes stature reconstruction a suitable test of whether AI can move beyond executing a predefined algorithm.

In our established laboratory workflow, surveillance images or images extracted from video are combined with a metric 3D model of the scene. The scene is documented by laser scanning, and distributed reference markers provide 3D control information. A reference photograph containing the same markers is processed in PhotoModeler (PhotoModeler Technologies/Eos Systems Inc., Vancouver, BC, Canada). Corresponding marker positions in the image and metric scene are used to reconstruct the camera position, orientation, projection geometry, and relevant lens-distortion parameters. The image is then corrected for lens distortion. The reconstructed camera position, corrected image, and 3D scene are transferred to 3ds Max (Autodesk, Inc., San Francisco, CA, USA). An experienced examiner inserts the standard 3ds Max Biped and adjusts its pose and body proportions until its projection matches the person in the evidence image. [15] At the same time, the body must remain anatomically and spatially plausible in 3D. Adjacent video images can provide additional information about the movement into and out of the selected pose. PhotoModeler-based and laser-scan-supported approaches of this general type have been evaluated in previous suspect-height studies [10–13]. More recently, calibrated reverse-

projection photogrammetry has been combined with a parametric 3D human body model for real-scale body-shape and height estimation [16].

A single convincing overlay is not sufficient. The internal laboratory procedure requires at least three evaluable positions, with additional reconstructions used when suitable images are available. Stature is derived from the reconstructed body geometry and interpreted from visible external endpoints. Footwear, hair, headwear, clothing, posture, and landmark placement can all shift the measured external extent away from true anatomical stature. Inter-observer variation in head and foot point selection has also been quantified in forensic height analysis [17]. The available reconstructions are therefore considered together rather than treating one image as definitive.

The established workflow also created a deliberate transfer problem. Camera reconstruction and image correction are normally performed in PhotoModeler, while the 3D reconstruction uses 3ds Max and its standard Biped. As an additional complication, neither program was available to the agent. Codex had access only to Blender and Python, and it was not given a precomputed PhotoModeler camera or a ready-made 3ds Max Biped. It therefore had to reproduce the required functions with the available tools: recover the camera from the marked reference image and metric scene, correct the test subject images, build an articulated and parameterized human model, control pose and body proportions, test 3D plausibility, measure stature, and generate auditable outputs.

The task also extended beyond a single stature value. The four images extracted from video were to be treated as separately fitted single-image reconstructions, with intermediate iterations preserved and documented. The system was asked to calculate the resulting measurements, perform the required descriptive or sensitivity analyses, and produce an illustrated report covering the method, individual reconstructions, measurements, relevant error information, and limitations. The central question was therefore whether a general-purpose AI agent could turn an expert forensic specification into an executable workflow: select or adapt suitable methods, implement missing components, run the analysis, inspect and revise intermediate results, and deliver an auditable measurement and report without interactive human reconstruction. We use the deliberately provocative phrase 'No Humans!' as shorthand for this human-out-of-the-loop execution phase after task specification. It does not imply autonomous evidence acquisition, scientific validation, legal responsibility, or final expert accountability.

2. Materials and Methods

2.1 Study design and task formulation

The experiment was designed to test whether a general-purpose AI agent could translate an established forensic procedure into a working technical pipeline and then execute it without case-specific human interaction. The established laboratory workflow uses PhotoModeler Premium (PhotoModeler Technologies/Eos Systems Inc., Vancouver, BC, Canada) for camera reconstruction and Autodesk 3ds Max (Autodesk, Inc., San Francisco, CA, USA) for 3D body reconstruction. As a preparatory step, the

laboratory quality-management document for forensic stature reconstruction was provided to ChatGPT (OpenAI, San Francisco, CA, USA). ChatGPT was used to formulate a requirements-oriented master prompt for Codex with GPT-6 Astra. The prompt defined the forensic objective, required workflow behavior, checks, and final deliverables. It did not contain case-specific executable code or a Blender implementation. Codex therefore had to decide how to implement the procedure with the software and data available to it. The complete master prompt is provided as Supplementary Material S1.

The experiment addressed the 'unknown person' variant of the stature-reconstruction task. No suspect body scan or known reference morphology was supplied. The primary question was whether Codex could reconstruct the test subject's stature from the controlled test material. The system was also required to preserve its intermediate processing and fitting steps and to produce a final report for subsequent technical assessment.

2.2 Controlled acquisition and test dataset

A controlled dataset was acquired in an indoor room using a Blackmagic Pocket Cinema Camera 6K Pro (Blackmagic Design Pty Ltd, Melbourne, VIC, Australia). The test subject moved through the scene and adopted different body positions, including a forward-bent posture. Four images showing the test subject from head to foot were selected, each with both feet in contact with the ground and two of them with heels in elevated positions. The face was anonymized using GIMP 3.2.6 (GIMP Development Team, international open-source project) before the images were supplied to the AI system.

Photogrammetric markers were distributed throughout the room to provide geometric reference information. A reference photograph was then acquired without changing the camera position or rotation and without changing the camera body, lens, or lens settings used for the test subject recording. The camera geometry and optical settings were therefore identical to the walkthrough acquisition. The marked scene was documented with a Leica RTC360 terrestrial laser scanner (Leica Geosystems AG, Heerbrugg, Switzerland). The registered point cloud, supplied in E57 format, provided true-to-scale 3D scene geometry and contained the same physical marker configuration visible in the photograph. The test subject's stature was measured independently with shoes on and retained as the reference for post-run comparison.

2.3 Input package and autonomous task

The autonomous run received the registered laser scan, the marked reference photograph, the four anonymized test subject images, the laboratory quality-management document, and the requirements-oriented master prompt. No precomputed PhotoModeler camera solution or ready-made 3ds Max reconstruction was supplied. Codex received only the four selected images, not the full video or adjacent images. This reflects casework in which only individual images or recordings with very low frame rates are available. The test was intended to show whether the agent could reproduce the required processing in the available environment.

The agent was instructed to implement the complete case-processing pipeline. It had to recover the camera and relevant image-distortion parameters from the marked reference image and metric scene, create the image representation needed for reconstruction, import and manage the metric scene, build an articulated biped-like human model, and process all four test subject images as separately fitted single-image reconstructions. It also had to calculate the individual and combined stature results, perform the requested statistical or sensitivity analyses, preserve intermediate iterations, and generate an illustrated report describing the processing steps, measurements, error information, and limitations. The code also had to be retained as a reusable pipeline package so that future runs would not require the workflow to be developed again.

The master prompt imposed several non-negotiable constraints. Both feet had to remain in contact with the measured floor in every selected image. Where the image provided enough information, the projected 3D joints had to match the visible anatomical joint locations. The camera and metric scene had to be reconstructed and checked before body fitting and then remain fixed. Camera calibration and the distortion-correction solution were determined once from the reference photograph by the agent and reused for all four images to keep the 3D reconstructions comparable. The system was not allowed to improve the body overlay by moving the camera or scene. The human model had to remain articulated and anatomically plausible, with symmetric left-right segment lengths and physiologically reasonable joint motion. A convincing evidence-camera overlay alone was not sufficient. Each candidate had to be inspected from additional frontal, lateral, oblique, and top views and revised or rejected if its 3D geometry, anatomy, ground contact, or pose was implausible. Only pose information visible in the supplied images was available; the agent had no adjacent-image or full-video movement context. Intermediate states had to be preserved for audit.

2.4 General-purpose AI agent and execution environment

The autonomous run used GPT-6 Astra at High reasoning effort through Codex (OpenAI, San Francisco, CA, USA). Blender 5.1 (Blender Foundation, Amsterdam, The Netherlands) was the target 3D environment. Codex had full access to the experimental workstation, including the local file system, command line, installed applications, case directory, Python environment, Blender executable, Blender's bpy API, and local computational resources. PhotoModeler Premium and Autodesk 3ds Max were not available for execution on this workstation. A Blender MCP interface was installed, but its use during the run could not be verified and is therefore not considered part of the experimental method. The archived implementation shows reproducible geometry operations performed through Python/bpy and Blender execution.

The master prompt had been prepared in advance with ChatGPT from the laboratory workflow and the experimental requirements. The aim was to let the AI carry out the established laboratory workflow autonomously, including detailed documentation of each processing step and preparation of a report. The prompt specified what the system had to achieve, not a case-specific implementation. After the Codex run started, the only investigator intervention was a 'continue' prompt whenever the available token allowance was exhausted. No technical corrections, case-specific Blender code, or manual biped

adjustments were supplied during execution. Codex selected or constructed the required computational methods, executed them, inspected intermediate results, and revised the workflow when needed.

2.5 Reference stature and technical assessment

The test subject's stature was measured independently with a physical measuring scale while wearing the shoes used in the recording and withheld from Codex during reconstruction. After the autonomous run, the per-image and combined AI estimates were compared with this reference. In routine laboratory reporting, the established stature-reconstruction method uses an error range of +/-3 cm based on internal validation. For this proof-of-concept experiment, an AI-derived stature within +/-3 cm of the measured reference was predefined as the target for metric success. Meeting this target in one controlled reconstruction would not validate the AI-based method, but it would provide an initial indication that the approach can produce a quantitatively accurate result.

Technical correctness was assessed from the camera comparison, agreement with the measured reference stature, variation in reconstructed body dimensions, and the final model views. The additional views were used to inspect anatomical proportions and spatial placement, but they did not provide independent observations of the true pose.

2.6 Documentation and post-run inspection

All generated Python and Blender code, Blender scenes, run logs, intermediate checkpoints, AI review decisions, numerical outputs, and reports were retained after execution. The archived material was inspected to document how Codex had implemented the requested workflow and to distinguish operations performed by local deterministic code from tasks that still required interpretation by the cloud-based AI model during case processing. This inspection was descriptive and did not modify the reported reconstructions or generated code.

The complete master prompt, final AI-generated report and iteration report file are provided as Supplementary Materials S1-S3, respectively. Together, these materials document the task specification, processing history, reconstruction instructions, and technical handoff. Supplementary Materials S1 to S3 are archived on Zenodo, as specified in the Data and code availability statement.

2.7 Execution protocol

After the Codex run was launched, investigator input was limited to a 'continue' prompt whenever the available token allowance was exhausted. No technical corrections, manual biped adjustments, case-specific Python or Blender code, or changes to the camera solution were supplied during execution. The four test subject images were processed as separately fitted single-image reconstructions, and all intermediate states were retained for later inspection.

Post-run inspection of the archived code and logs was used to classify each operational step as user input, AI-mediated interpretation or review, local Python/Blender computation, or a combination of these. The resulting runtime division of work is reported in Results and summarized in Table 1.

Table 1
Allocation of tasks between the investigator, AI, and locally executed Python/Blender code during data acquisition and autonomous case processing.

Operational step	Performed by	Function during case execution
Data acquisition and preparation	User	Record the test subject, place reference markers, acquire the reference photograph and laser scan, and measure reference stature. Select and anonymize the four images before case processing.
Case input and launch	User	Provide the reference photograph, laser scan, four test subject images, and workflow instructions; then start the run. Subsequent input is limited to 'continue' prompts after token allowances are exhausted.
Camera correspondences and calibration	AI + local Python	AI interprets target identity and ambiguous correspondences. Python calculates the camera and distortion model, runs diagnostics, and locks the accepted camera.
Scene and support preparation	Local Python / Blender	Load the metric scan, preserve scale and coordinates, derive support surfaces, and prepare the scene for body placement.
Image observations	AI	Inspect each image and identify visible anatomical landmarks, left/right assignment, occlusion, foot contacts, and pose information that geometry alone cannot provide.
Initial biped placement	Local Python / Blender	Use the locked camera, support geometry, and AI observations to initialize the articulated model with both feet on the measured floor.
Staged body fitting	Local Python / Blender	Optimize global position, joint angles, and later body dimensions against the image observations while enforcing geometric constraints.
Multi-view plausibility review	AI	Inspect the evidence-camera overlay and frontal, lateral, oblique, top, contact, and skeleton views. Return ACCEPT, REFINE, or REJECT and, when needed, specify revised constraints for another local fit.
Stature measurement and statistics	Local Python	Place the body model in an upright pose without rescaling the accepted morphology, calculate support-to-head stature, repeat for all four images, and compute descriptive statistics and sensitivity outputs.
Final report and package	AI + local code	Local code assembles figures, tables, numerical files, hashes, and Blender scenes. AI review decisions and explanatory text are incorporated into the generated report.

2.8 End points and evaluation

The experiment was evaluated against three end points: development of the workflow; conduct of the measurement, including reporting; and technical correctness of the outputs. Development required an

executable pipeline implementing the specified camera reconstruction, scene handling, articulated body model, fitting, and measurement functions. Conduct of the measurement required processing all four images and producing measurements, intermediate audit material, and an illustrated report, with investigator input limited to the continuation prompts described above.

Technical correctness was considered separately from successful development and execution. The reconstructed camera was compared with an independent PhotoModeler solution from the established laboratory workflow. Individual and combined model heights were compared with the measured reference using the predefined +/-3 cm target. Variation in fitted body dimensions and the final 3D review views were used to assess reconstruction consistency and identify limitations of anatomical and spatial interpretation.

3. Results

The experiment produced two types of results. First, Codex's technical solution was itself an outcome: the agent had to decide how to reproduce the PhotoModeler/3ds Max workflow with Python and Blender and how to divide the work between local computation and AI-mediated interpretation during execution. Second, the resulting pipeline produced the camera reconstruction, four body reconstructions, stature measurements, intermediate audit material, and final report. The following sections therefore describe the generated workflow and its runtime behavior before assessing the technical correctness of its outputs.

3.1 Workflow development

Codex developed a local Python/Blender pipeline rather than reproducing the laboratory procedure as a sequence of manual software operations. Inspection of the archived implementation identified six main functions: evidence and provenance management; camera and image-geometry recovery; metric scene and support-surface handling; construction of an articulated human model; numerical body fitting with three-dimensional plausibility checks; and stature measurement, sensitivity analysis, and automated documentation.

The generated human model used a 23-bone articulated rig. Body shape and pose were represented separately. Morphology variables described segment lengths and body dimensions, whereas pose variables described global placement and major joint rotations. As required by the generated master prompt, the implementation constrained corresponding left and right segments to equal length, broad anatomical ranges discouraged clearly unrealistic proportions, and no target stature was supplied to the fitting process.

Codex replaced manual biped positioning with a numerical search implemented with SciPy's least-squares optimizer. [18] The local software repeatedly changed the model, projected its 3D landmarks into the evidence image, and minimized the differences from the observed joint positions while also enforcing the specified geometric constraints. These included support by both feet, plausible joint

configuration and body proportions, left-right consistency, and avoidance of self-intersection and scene penetration. Silhouette agreement was used mainly for visual review because clothing and facial anonymization made the visible outline an unreliable anatomical target.

Fitting was organized in stages. The generated pipeline first established the overall body position, then refined feet including heel position and lower limbs, torso and head, upper limbs, body dimensions, and the final joint configuration. The locked camera projected the 3D landmarks back into the image at each stage, while the laser-scan surface constrained ground contact. Intermediate checkpoints and review renderings were saved throughout for the four separately fitted single-image reconstructions.

3.2 Conduct of the measurement and reporting

The generated workflow processed all four test subject images and produced camera and body reconstructions, intermediate checkpoints, multi-view review images, stature measurements, sensitivity outputs, and the final illustrated report. The separate iteration report preserves the reconstruction history and the reasons recorded for accepted, refined, and rejected intermediate candidates.

Execution combined locally executed AI-written source code with direct AI analysis. Local Python and Blender code performed the reproducible numerical calculations, optimization, geometry operations, rendering, and measurement. Codex remained active during the run for tasks that required interpretation rather than deterministic calculation. These included identifying anatomical landmarks and occlusions, assigning anatomical left and right, interpreting the pose, and reviewing candidate reconstructions from multiple views. Table 1 summarizes the division of work in case-processing order.

Camera reconstruction was completed and locked before body fitting. For each image, local code generated the evidence-camera overlay and additional frontal, lateral, oblique, top, contact, and skeleton views. Codex inspected these outputs and returned ACCEPT, REFINE, or REJECT decisions. When a candidate was refined, the affected fitting stage was repeated with revised constraints while the camera and scene remained fixed. Thus, AI-mediated review remained part of the operational pipeline throughout the ambiguous 3D body-positioning stages.

The final measurement stage ran locally. The accepted morphology was copied without rescaling, the body model was placed in an upright pose for height measurement, and support-to-head stature was calculated for each reconstruction. Local code also assembled numerical files, figures, hashes, and Blender scenes, while the AI-generated review decisions and explanatory text were incorporated into the final report. Completion of these outputs demonstrated conduct of the requested measurement and reporting workflow separately from the accuracy of the resulting measurements.

3.3 Technical correctness

3.3.1 Camera comparison

The same dataset was calibrated independently in PhotoModeler Premium using the established laboratory procedure. The Codex camera center differed from the PhotoModeler solution by 32.8 mm in 3D. Expressed in the PhotoModeler camera coordinate system, approximately 32.7 mm of this difference lay along the viewing direction, with about 3.0 mm vertical displacement and negligible lateral displacement.

To compare the geometric effect of the two solutions, the 91 PhotoModeler-exported 3D reference points were projected through both camera models onto the same 1920 x 1080 image raster. The mean distance between corresponding projected positions was 9.18 pixels (median, 9.11 pixels; range, 7.36–12.17 pixels). The shift was systematic, averaging approximately 6.35 pixels horizontally and 6.50 pixels vertically (Fig. 1). These values describe the difference between the Codex and PhotoModeler solutions. They are not residuals against independently measured image points and therefore do not establish which calibration is more accurate.

3.3.2 Stature and body reconstruction

The system completed four separately fitted single-image reconstructions. Each image started from the same generic body dimensions, but pose and morphology were fitted separately for each image. The documented fitting procedure did not explicitly use previous body dimensions, fitted poses, target stature, or a cross-image mean to guide an individual fit. The reported values are external support-to-head model endpoints, compared here with reference stature measured with shoes on. Hair and model endpoint placement may still contribute to differences from the physical measurement.

The final model heights were 176.65 cm for image 2295, 175.96 cm for image 2336, 161.30 cm for image 2353, and 169.27 cm for image 2415. The test subject's measured stature was 175 cm. The mean autonomous estimate was 170.8 cm, with a sample standard deviation of 7.2 cm (n = 4). The corresponding errors were + 1.65, + 0.96, -13.70, and − 5.73 cm. Body-landmark RMS residuals were 6.92, 7.70, 14.78, and 10.58 pixels. All four saved reconstructions passed the implemented geometry checks. The estimates spanned 15.35 cm; image 2353 produced both the lowest stature and the largest landmark residual. The per-image results are summarized in Table 2.

Table 2
Final stature estimates, body-landmark RMS residuals, and geometry checks for the four separately fitted single-image reconstructions.

Image	Model height (cm)	Landmark RMS (px)	Geometry check
2295	176.65	6.92	PASS
2336	175.96	7.70	PASS
2353	161.30	14.78	PASS
2415	169.27	10.58	PASS
Mean / SD	170.8 / 7.2	-	-

Two of the four reconstructions fell within the laboratory benchmark of +/-3 cm around the measured reference. The mean was 4.2 cm below the reference and therefore 1.2 cm outside this benchmark. The sample standard deviation describes disagreement among the four image-based fits only; it is not a validated forensic uncertainty estimate or confidence interval. Averaging does not remove common camera, scan, body-model, clothing, hair, or footwear effects. An example of the automatically generated multi-view output is shown in Fig. 2, and Fig. 3 summarizes the four stature estimates against the measured reference.

The reconstructed body dimensions varied even more strongly between images. Because all four reconstructions represent the same person, measures such as femur length, tibia length, trunk length, pelvis width, and shoulder width should remain broadly similar. Instead, the fitted ranges were 34.19–45.78 cm for the femur, 40.23-53.00 cm for the tibia, 32.00-50.83 cm for the trunk, 15.00-25.93 cm for pelvis width, and 33.25-49.00 cm for shoulder width. These auxiliary dimensions are not reported in our routine stature opinions, and accurate recovery of them was not a task requirement. They are shown here as diagnostic indicators of reconstruction instability. The pattern suggests that the system sometimes absorbed unresolved 3D pose ambiguity by changing body proportions rather than recovering stable anatomy. Figure 4 summarizes the between-image variation in these fitted dimensions.

Computational sensitivity analyses further indicated instability. Of six landmark-localization perturbations per image, five were accepted for image 2295, none for image 2336, two for image 2353, and none for image 2415. For image 2353, the accepted perturbations yielded heights of 172.31 to 174.70 cm, compared with the main estimate of 161.30 cm. These conditional ranges do not constitute validated measurement uncertainty intervals (Supplementary Material S2).

4. Discussion

4.1 Study aim and technical feasibility

This Technical Note does not propose a new method for forensic stature estimation. It asks whether a current general-purpose AI agent can implement and execute a complex, underdetermined forensic workflow that normally requires substantial expert input. Three questions are central: can the agent develop the required workflow, can it conduct the measurement and produce the report, and how technically correct are the resulting reconstructions and measurements?

The first two questions can be answered positively at the level of this single-case experiment. Codex translated the specified PhotoModeler/3ds Max procedure into an executable Python/Blender pipeline and used it to process all four images and produce measurements and documentation. Deterministic calculations ran locally, while Codex remained involved in anatomical interpretation, inspection of candidate models, and decisions to accept, refine, or reject fitting stages. Investigator input was limited to 'continue' prompts when token allowances were exhausted. These results demonstrate development and execution capability, while the third question requires a separate assessment of the outputs.

The combination of code generation, tool use, visual review, and iterative revision has precedents outside forensic science. SceneCraft, for example, generates Blender code and uses visual feedback to refine 3D scenes [6]. The contribution here is the application of these capabilities to an expert-specified forensic measurement workflow with metric scene constraints and an independently measured reference. The automatically generated review figure also illustrates that technically complex execution can coexist with simple errors: the labels 'left' and 'right' appear to be inverted in Fig. 2.

4.2 Technical correctness and the limits of autonomous 3D interpretation

The camera comparison showed a small difference between the Codex and PhotoModeler solutions: 32.8 mm in camera-center position, almost entirely along the viewing direction, and a mean projected-point displacement of 9.18 pixels. As an illustrative first-order estimate, if the 32.7 mm depth difference acted alone as an uncompensated change in camera-to-person distance at approximately 4.8 m, with focal length, orientation, distortion, and body pose otherwise unchanged, the implied scale difference would be about 0.7%, equivalent to approximately 1.2 cm for a 175 cm person. The actual solutions differ in interacting parameters, so this is not a measured stature error. It nevertheless suggests a modest effect under these assumptions, rather than an effect large enough to explain the full spread of the body reconstructions.

Body fitting depends more strongly on interpretation. Different combinations of pose, depth, joint position, and body dimensions can produce similar projections from the evidence camera. Additional views can reveal unusual anatomy or spatial placement, but they do not supply the missing observations. This stage therefore depends on anatomical, biomechanical, and movement-related interpretation in addition to geometric fitting.

The test subject's reference stature, measured with shoes on, was 175 cm. Two of the four autonomous reconstructions were within the laboratory benchmark of +/-3 cm. The mean was 170.8 cm, 4.2 cm below the reference, and one image underestimated stature by 13.7 cm. Thus, the workflow completed the measurement and reporting tasks but did not meet the predefined target for the combined estimate. Passing the implemented geometry checks did not ensure that a model correctly represented the test subject.

The variation in reconstructed segment dimensions helps explain this limitation. The same test subject was assigned substantially different femur, tibia, trunk, pelvis, and shoulder dimensions across images, while the additional views showed imperfect positioning of the articulated model. The system could satisfy the camera-view fit without resolving the underlying anatomy and depth consistently. This pattern is consistent with limitations in the anatomical and movement-related interpretation applied during fitting, although the experiment cannot isolate their contribution from other sources of error. Experts using the established workflow have access to a ready-made, readily parameterized 3ds Max Biped, whereas Codex had to construct its own articulated body model. This placed the autonomous workflow at a disadvantage and limits direct comparison with expert reconstruction.

4.3 Implications for forensic practice and the expert role

General-purpose agents may make case-specific computation and specialized tools practical for problems that would not otherwise justify dedicated software development. Potential applications include data conversion, registration, visualization, measurement, iterative review, and reporting. Whether these capabilities improve forensic practice will depend on their reliability and the effort needed to verify them. Easier software development would increase the importance of provenance, testing, version control, software lifecycle management, review, and clearly defined validation boundaries [19].

The forensic expert's role could consequently change. This experiment shows that an agent can undertake substantial implementation and execution work after expert task formulation. If future systems become more reliable, expert effort may shift toward problem formulation, meaningful constraints, supervision, validation, exception handling, and final interpretation. In routine laboratory practice, an expert requires approximately 2–3 days for image correction, stature reconstruction from at least three images, and report preparation. Active processing time for an autonomous run was estimated at approximately 3–4 hours, excluding interruptions caused by token limits. These estimates were not obtained in a controlled timing comparison. The direction and extent of such a shift remain uncertain, and neither expert time savings nor equivalence to expert reconstruction was established in this study.

This experiment is a snapshot of a rapidly developing technology. GPT-6 Astra was publicly released on 3 September 2026, shortly before the study [7]. Further improvements in AI capability are a reasonable expectation, but this experiment cannot establish their pace, their relevance to forensic reconstruction, or whether they will reduce the specific errors observed here.

4.4 Oversight, explainability, and operational constraints

Greater autonomy creates corresponding quality-management and legal questions. If an agent can select or modify methods during a case, laboratories must define which decisions require human approval, who is responsible for those decisions, and how the full technical history is preserved. Generated Python and Blender code make many numerical operations inspectable, but the AI itself is neither deterministic nor fully transparent. A defensible implementation should therefore preserve model and software versions, prompts, source hashes, generated code, numerical outputs, review decisions, and intermediate checkpoints. Where practical, deterministic calculations should remain separate from AI judgment. For fixed input and an unchanged implementation and software environment, the deterministic code produces the same output. The AI-mediated stages have a stochastic component, so repeated end-to-end runs need not produce identical interpretations or reconstructions.

The same distinction matters in court. A technically convincing AI-generated report does not make the method explainable or admissible by itself. The laboratory must still show what was calculated, what the AI inferred, what was independently checked, what uncertainty was established, and where a human expert accepted responsibility for the conclusion. Repeated apparent success may lead to deskilling and

reduced vigilance, while expectations of faster turnaround may erode the time available for careful expert review.

Practical limits were also substantial. The experiment used a standard ChatGPT Plus account costing approximately 20 USD/CHF per month. The token allowance was repeatedly exhausted after about 30–45 minutes of intensive execution, and the workflow therefore extended over several days. Each interruption was followed by a 'continue' prompt; no additional technical guidance was supplied. Exact token use could not be reconstructed reliably because execution resumed with renewed context loading, so token count was not treated as a comparative metric. A low-cost subscription was sufficient to complete the experiment, but with repeated interruptions and substantial elapsed time. Any future efficiency benefit will depend on inference cost, secure infrastructure, and the effort required for validation and human oversight.

Data governance is a separate barrier. This study used controlled, anonymized internally generated test material. Operational forensic data may contain sensitive personal, medical, or law-enforcement information and cannot simply be transferred to external cloud services. Local deployment could reduce this problem, but an equivalent local frontier-agent workflow was not evaluated here and may require substantial hardware or future developments.

4.5 Limitations

This experiment used one test subject, four images from one controlled recording, one room, and relatively favorable image and scan data. It is a proof of concept, not a validation study, and does not establish general measurement performance. We did not test low-resolution, compressed, poorly illuminated, partially occluded, or otherwise degraded surveillance imagery, although such conditions are common in casework. The prompt required separate image-based fits, and the system treated them as such in the documented processing. However, information transfer between reconstructions through retained agent context or accessible processing history could not be excluded. Independence was therefore a procedural instruction, not a demonstrated isolation of information.

The physical measurement provides an external reference for stature with shoes on, but the true unobserved 3D pose cannot be recovered directly from the supplied images. Additional model views allow qualitative assessment of plausibility, not comparison with geometric ground truth. The findings are specific to one model, one prompt-development process, and one tool environment. The supplied instructions and iteration history support repetition of the procedure, but repeated agent runs may produce different implementations or solutions. Formal evidential validation would require a larger blinded study across test subjects, scenes, camera geometries, image qualities, clothing, movement patterns, and repeated runs [19].

4.6 Future work

A useful next experiment would remove most of the methodological scaffolding used here. Rather than receiving the established stature workflow, the agent could receive only the forensic objective and a

development set of reference cases containing scene scans, multiple images, and independently measured stature. It would then have to infer, implement, and document its own method. The resulting procedure could be frozen and tested on separate blinded cases. This would move the question from translating an expert-designed workflow to developing a useful forensic strategy from the problem definition and reference examples alone. This could also lead to improved or more efficient workflows, which would need to be tested independently. Unless the model were explicitly fine-tuned, this should be described as development or in-context learning rather than model training.

5. Conclusion

This proof-of-concept shows that a general-purpose AI agent can develop and execute a complex forensic 3D workflow with minimal human intervention. Technical accuracy was not yet sufficient for routine forensic use, particularly for interpretation-dependent body fitting. Nevertheless, the level of autonomy achieved indicates that such systems may substantially change forensic practice. Technical implementation and execution may increasingly shift from the examiner to AI, while expert work focuses more on task definition, validation, and final interpretation. The present results suggest that this transition has already begun. Forensic methods and quality management will have to adapt accordingly.

Declarations

Ethics and consent

Ethical approval was not required for this study. The test subject gave informed consent to participate and to the publication of the anonymized images shown in this manuscript.

Funding

None.

Conflict of interest

Lars Christian Ebert and Dominik Haenni previously participated in OpenAI’s external red-teaming activities. This prior activity was unrelated to the present study. The study was conducted independently, and OpenAI had no role in its design, execution, analysis, interpretation, or reporting. The authors declare no other competing interests.

Use of generative AI in manuscript preparation

A general-purpose AI system was itself the subject of the experiment. ChatGPT also assisted with manuscript drafting and literature research. The human authors critically reviewed, corrected, and approved the manuscript.

Data and code availability

The master-prompt document (S1), AI-generated reconstruction report (S2), and iteration journal (S3) are available on Zenodo (version 1.0, DOI 10.5281/zenodo.23035504). This deposit contains supplementary documentation, including reported measurements and illustrative images. The complete input dataset, executable source code, editable Blender scenes, and full machine-readable execution logs are not included in this deposit.

References

1. Bugelli V, Calabrò F, Donato L, Cecchi R, Camatti J, Di Paolo M et al (2026) Artificial intelligence in forensic science: a systematic review. Part I: personal identification. Int J Legal Med 140(5):3317–3329. https://doi.org/10.1007/s00414-026-03855-5
2. Stacey J, Fleming R, Sheppard D, Sheppard J, Dobbie G, Karunakaran D (2025) A responsible artificial intelligence framework for forensic science. Forensic Sci Int 375:112548. https://doi.org/10.1016/j.forsciint.2025.112548
3. Meconi TNT, Henseler H (2026) Integrating generative AI into forensic workflows: a case study from the Netherlands Forensic Institute. Forensic Sci Int 388:113074. https://doi.org/10.1016/j.forsciint.2026.113074
4. Chernyshev M, Baig ZA, Syed N, Doss R, Shore M (2026) Large language models in digital forensics: capabilities, challenges and future directions, Forensic Sci. Int Digit Investig 56:302043. https://doi.org/10.1016/j.fsidi.2025.302043
5. Sun C, Han J, Deng W, Wang X, Qin Z, Gould S (2025) 3D-GPT: procedural 3D modeling with large language models, 2025 International Conference on 3D Vision (3DV) 1253–1263. https://doi.org/10.1109/3DV66043.2025.00119
6. Hu Z, Iscen A, Jain A, Kipf T, Yue Y, Ross DA et al (2024) SceneCraft: an LLM agent for synthesizing 3D scenes as Blender code, Proc. Mach. Learn. Res. 235 19252–19282. https://proceedings.mlr.press/v235/hu24g.html
7. OpenAI (2026) GPT-6 Astra: a new generation of intelligence, 3 September https://openai.com/index/gpt-6-astra/ (accessed 10 September 2026)
8. Criminisi A, Reid ID, Zisserman A (1999) Single view metrology, in: Proceedings of the Seventh IEEE International Conference on Computer Vision, vol. 1, IEEE, pp. 434–441. https://doi.org/10.1109/ICCV.1999.791253
9. Ramstrand N, Ramstrand S, Brolund P, Norell K, Bergström P (2011) Relative effects of posture and activity on human height estimation from surveillance footage, Forensic Sci. Int. 212 (1–3) 27–31. https://doi.org/10.1016/j.forsciint.2011.05.002
10. Johnson M, Liscio E (2015) Suspect height estimation using the FARO Focus3D laser scanner. J Forensic Sci 60(6):1582–1588. https://doi.org/10.1111/1556-4029.12829
11. Olver AM, Guryn H, Liscio E (2021) The effects of camera resolution and distance on suspect height analysis using PhotoModeler. Forensic Sci Int 318:110601. https://doi.org/10.1016/j.forsciint.2020.110601

12. Tosti F, Nardinocchi C, Wahbeh W, Ciampini C, Marsella M, Lopes P et al (2022) Human height estimation from highly distorted surveillance image. J Forensic Sci 67(1):332–344. https://doi.org/10.1111/1556-4029.14888
13. Ciampini C, Petrillo A, Zomparelli F, Groutas S (2024) An innovative method for human height estimation combining video images and 3D laser scanning. J Forensic Sci 69(1):301–315. https://doi.org/10.1111/1556-4029.15378
14. Becker S, Heuschkel M, Richter S, Labudde D (2022) COMBI: artificial intelligence for computer-based forensic analysis of persons. Künstl Intell 36:171–180. https://doi.org/10.1007/s13218-022-00761-x
15. Edelman G, Alberink I (2009) Comparison of body height estimation using bipeds or cylinders. Forensic Sci Int 188(1–3):64–67. https://doi.org/10.1016/j.forsciint.2009.03.013
16. Imoto D, Honma M, Asano M, Sakurai W, Kurosawa K (2025) Method for estimating real-scale 3D human body shape from an image based on 3D camera calibration and computer graphics-based reverse projection photogrammetry. J Forensic Sci 70(6):2276–2296. https://doi.org/10.1111/1556-4029.70130
17. Liscio E, Lim J (2024) Inter-observer variation of head and foot point selection for subject height determination. J Forensic Sci 69(4):1268–1288. https://doi.org/10.1111/1556-4029.15529
18. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D et al (2020) SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 17(3):261–272. https://doi.org/10.1038/s41592-019-0686-2
19. Swofford H, Lund S, Taylor M, Guttman B, Romsos EL, Soons J et al (2025) Validation in forensic science: guiding principles for the collection and use of validation data, NISTIR 8589, National Institute of Standards and Technology, Gaithersburg, MD. https://doi.org/10.6028/NIST.IR.8589

Figures

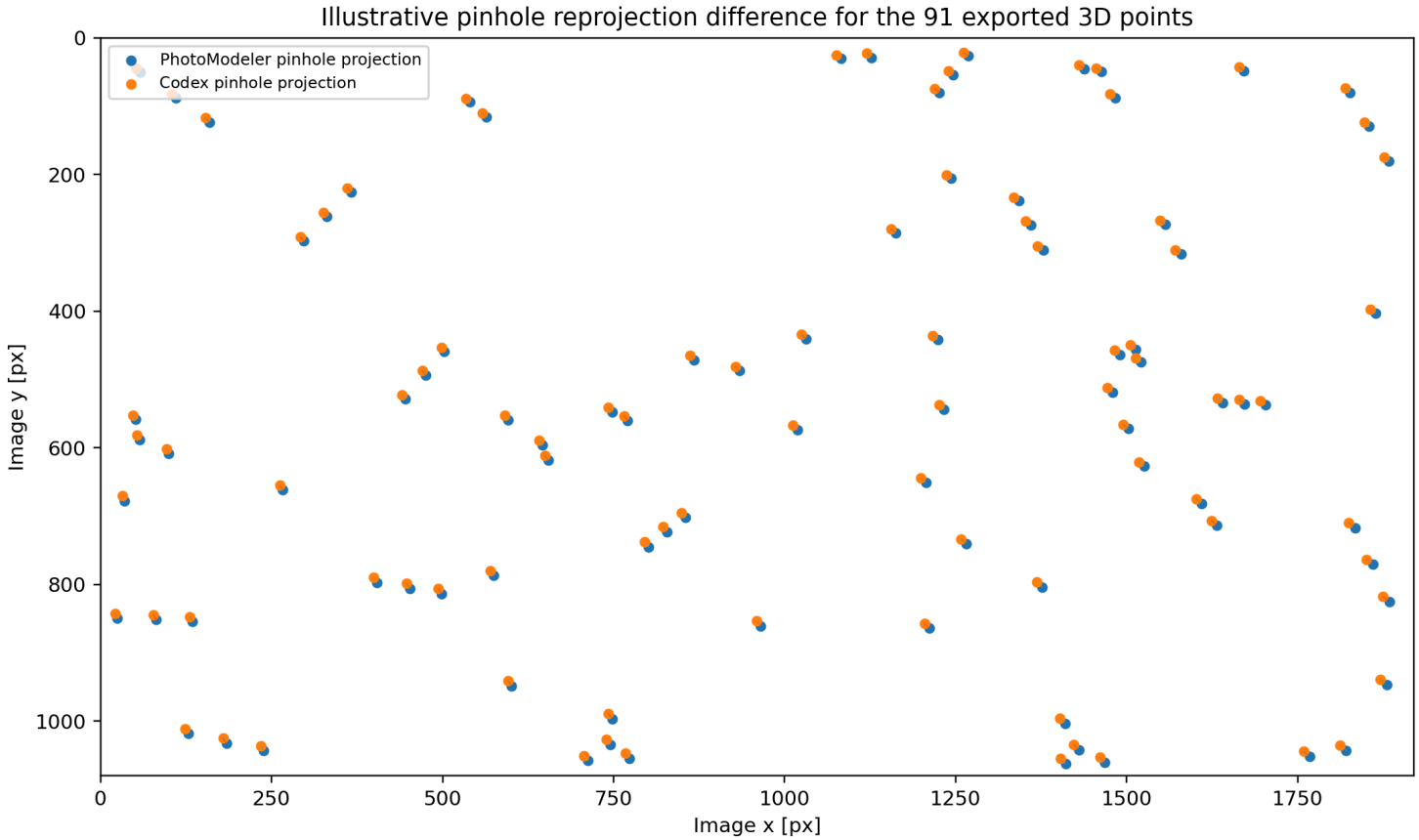


Figure 1

Comparison of the PhotoModeler and Codex camera projections. The same 91 exported 3D reference points were projected onto a common image raster; connecting lines show the displacement between solutions. These differences describe agreement between the two solutions, not residuals against independent ground-truth image coordinates.

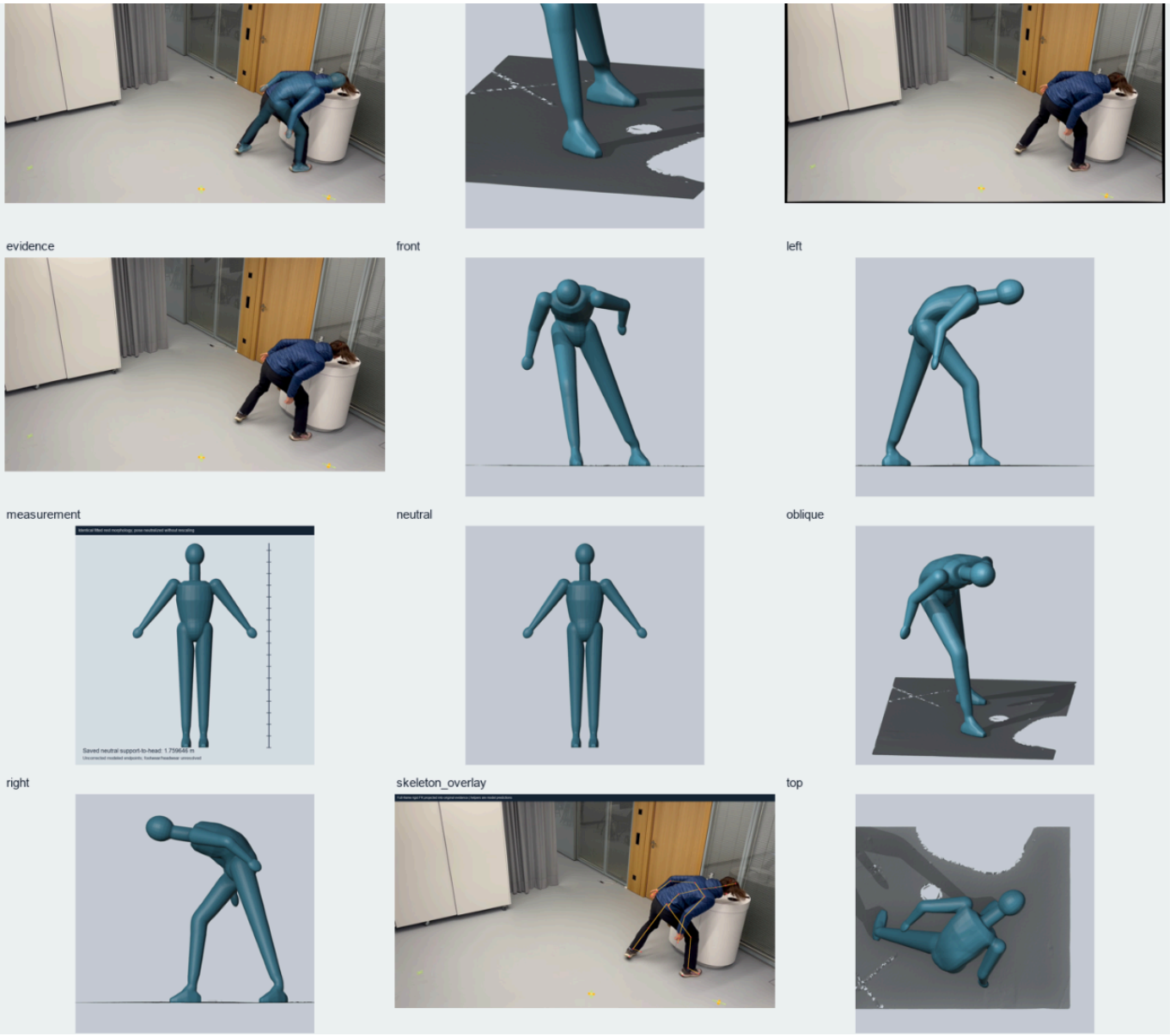


Figure 2

Automated multi-view plausibility and measurement output for image 2336, exported directly from the AI-generated report without editorial modification. The panel combines the evidence image, body and skeleton overlays, frontal and lateral model views, oblique and top views, floor-contact inspection, and the unscaled upright pose used for support-to-head measurement. The unobserved views are model-derived plausibility checks, not additional evidence. In the original AI output, the panels labeled 'left' and 'right' appear to be inverted; the labels are reproduced here unchanged.

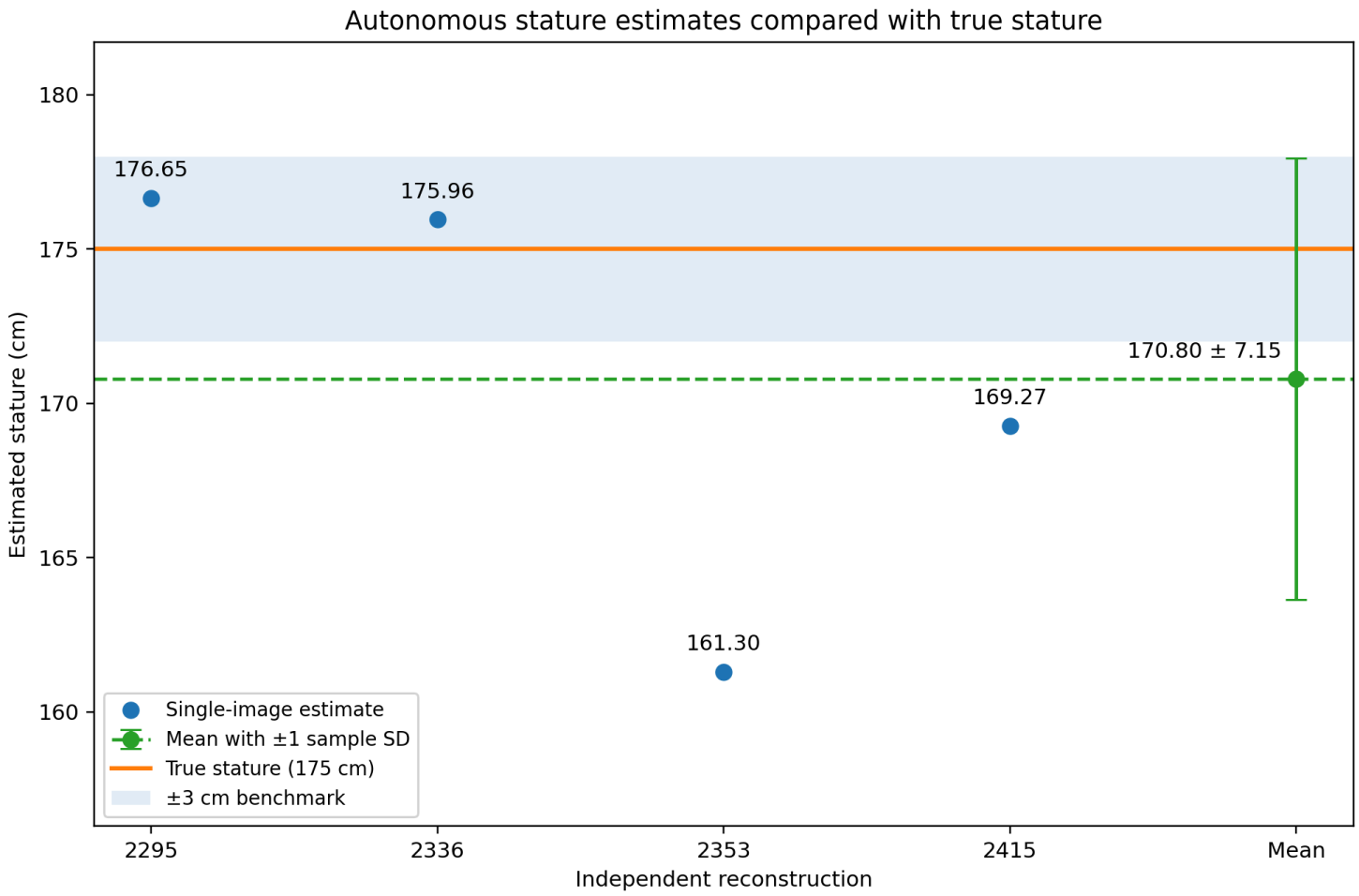


Figure 3

Four stature estimates from separately fitted single-image reconstructions shown as unconnected points, the mean and sample standard deviation, the measured stature of 175 cm, and the laboratory benchmark of +/-3 cm. Two individual estimates fell within the benchmark; the mean was 4.2 cm below the reference.

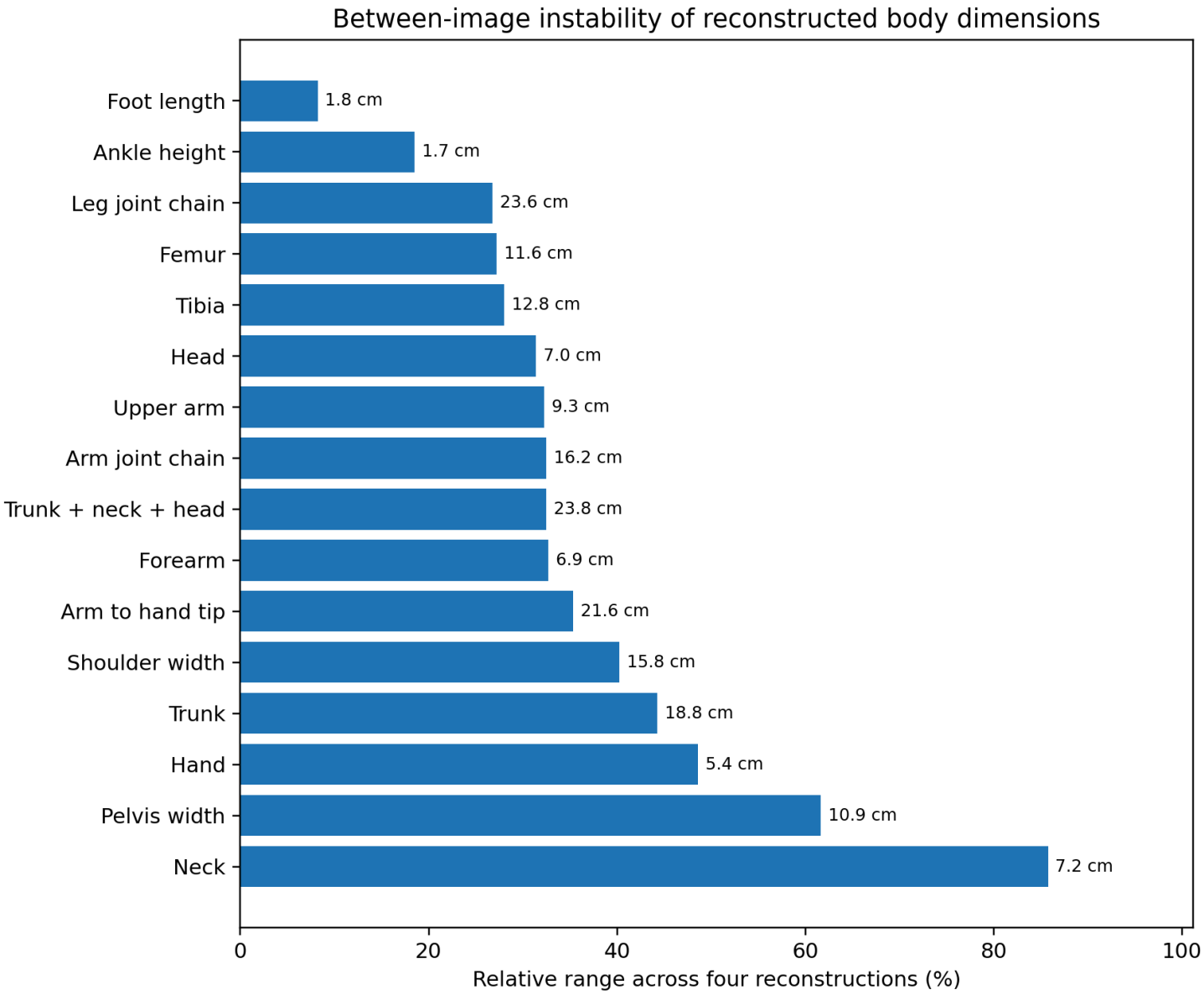


Figure 4

Between-image variation in reconstructed body dimensions across four separately fitted single-image reconstructions of the same test subject. Bars show the relative range for each dimension; labels give the absolute range in centimeters. The spread in several dimensions indicates that unresolved pose and depth ambiguity was partly absorbed by changes in fitted body proportions.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementarymateriallegends.docx
- S1Forensic3DStatureReconstructionMasterPrompt.txt

- S2Person1independentreconstructionreport.pdf
- S3Person1iterationjournalsmall.pdf