Generative Muscle Stimulation: Providing Users with Physical Assistance by Constraining Multimodal-AI with Embodied Knowledge

Yun Ho*
University of Chicago
Chicago, Illinois, USA
yunho@uchicago.edu

Romain Nith*
University of Chicago
Chicago, Illinois, USA
rnith@uchicago.edu

Peili Jiang
University of Chicago
Chicago, Illinois, USA
peilij@uchicago.edu

Steven He
University of Chicago
Chicago, Illinois, USA
stevenhestudent@gmail.com

Bruno Felalaga
University of Chicago
Chicago, Illinois, USA
brunofelalaga@uchicago.edu

Shan-Yuan Teng
University of Chicago
Chicago, Illinois, USA
tengshanyuan@uchicago.edu

Rhea Seeralan
University of Chicago
Chicago, Illinois, USA
rheaseeralan@gmail.com

Pedro Lopes
University of Chicago
Chicago, Illinois, USA
pedrolopes@uchicago.edu

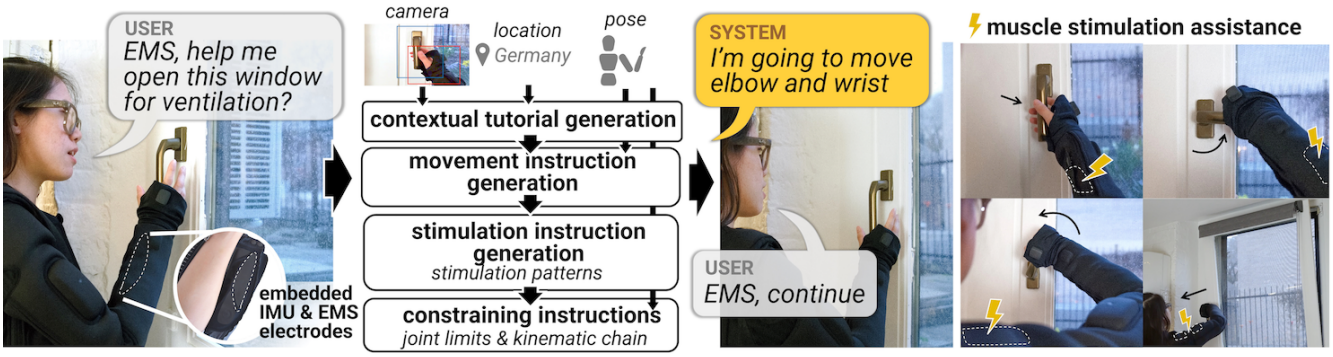

Figure 1: Our system assists users via electrical muscle stimulation; except, it does not deliver fixed-instructions but generates them. This user requests help to open a window in a foreign country. Our system gathers information: body pose (IMU suit) + point-of-view (objects, hand-poses) + location (e.g., "Germany"), etc. These are fed to a multimodal-AI that creates textual-instructions—allowing it to infer that this window opens vertically. As AI-models have no information to respect joint-limits of the user's body, our system constrains instructions with embodied knowledge. Thus, opening this window (used in our user study).

Abstract

Electrical muscle stimulation (EMS) can support physical-assistance (e.g., shaking a spray-can before painting). However, EMS-assistance is highly-specialized because it is (1) fixed (e.g., one program for shaking spray-cans, another for opening windows); and (2) non-contextual (e.g., a spray-can for cooking dispenses cooking-oil, not paint—shaking it is unnecessary). Instead, we explore a different approach where muscle-stimulation instructions are generated considering the user's context (e.g., pose, location, surroundings). The resulting system is more general—enabling unprecedented EMS interactions (e.g., opening a pill bottle) yet also replicating existing systems (e.g., *Affordance++*) without task-specific programming. It uses computer-vision/large-language-models to generate EMS-instructions, constraining these to a muscle-stimulation knowledge-base & joint-limits. In our user-study, we found participants successfully completed physical-tasks while guided by generative-EMS, even when EMS-instructions were (purposely) erroneous. Participants understood generated gestures and, even during forced-errors, understood partial-instructions, identified errors, and re-prompted the system. We believe our concept marks a shift toward more general-purpose EMS-interfaces.

*These authors contributed equally and are ordered alphabetically.

CHI '26, Barcelona, Spain

ACM ISBN 979-8-4007-2278-3/26/04
https://doi.org/10.1145/3772318.3790817

CCS Concepts

• **Human-centered computing** → **Haptic devices**; • **Hardware** → **Emerging interfaces**.

Keywords

Electrical Muscle Stimulation, Embodied AI, Physical AI, Force-feedback, Haptics, Affordance

ACM Reference Format:
Yun Ho, Romain Nith, Peili Jiang, Steven He, Bruno Felalaga, Shan-Yuan Teng, Rhea Seeralan, and Pedro Lopes. 2026. Generative Muscle Stimulation: Providing Users with Physical Assistance by Constraining Multimodal-AI with Embodied Knowledge. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 22 pages. https://doi.org/10.1145/3772318.3790817

1 Introduction

Procedural-knowledge is denoted as the "know-how" of performing a task, in contrast to knowing facts [17, 105]; it includes the *sequences of movements* needed to achieve the task. For instance, knowing the mechanism of a pickle jar is not synonymous with being able to open it effortlessly (it requires a specific gesture and force, e.g., leveraging grip friction to enable a strong twist). Supporting users in these situations is very challenging, as procedural-knowledge is often acquired from lived-experiences, and simply relaying facts is often insufficient [58, 74, 123].

To tackle this, much effort has been dedicated to engineering interfaces for physical assistance via *force-feedback* [12, 21, 32, 63, 72, 75, 93, 99]. One interface that has gained popularity in the last two decades, with ~150 publications on the topic in HCI alone [24], is electrical muscle stimulation (EMS) due to its wearability. Interactive systems based on EMS move the user's body via electrically-actuated movements. Examples of interactive EMS-assistance include: sign-language [76], eyes-free navigation [98], piano-playing [73, 97], and demonstrating movement instructions to assist users in operating objects that users have not used before [63]—in *Affordance++* [63], EMS is used to help users discover actions, e.g., force users to shake a spray-can before painting.

Unfortunately, the ~150 systems in HCI using EMS for physical assistance [24] are highly-specialized due to their **(1) rigid programming:** movement instructions are crafted in advance rather than generated in real-time (e.g., *Affordance++* [63] provides only six fixed EMS-instruction sequences) resulting in limited sets of predefined movements; and **(2) non-contextual**: EMS-systems ignore contextual-cues that could enable assisting users in more complex situations (e.g., *Affordance++* shakes a spray-can for painting, but it would fail to recognize that a cooking oil spray-can does not require shaking).

Towards our goal of advancing the research in interactive systems based on EMS, we explored a new avenue: we engineered a system that *generates* EMS-instructions given the user's request & context. As depicted in Figure 1, our system takes advantage of multimodal-AI: it takes in a user's spoken-request and images from their point-of-view (POV); it uses computer-vision (e.g., detect objects/hands) large-language-models (e.g., reason about objects, situations, surroundings), EMS-knowledge (e.g., feasible movements) and joint-information (e.g., joint-limits, kinematic-chain) to *generate* EMS-instructions that were *not* part of the system's built-in programming. In this particular example, where the user is facing a tilt-turn window (common in parts of Europe) that can be opened by turning the handle vertically, our EMS-system generated EMS movements that actuated the user's body, depicting the movements needed to open the window (i.e., stimulate the shoulder to turn the handle, then biceps to pull).

1.1 Scope and contribution

Our contribution is two-fold, one *technical* and one *conceptual*.

Our conceptual contribution is proposing a type of *physical* assistance based on multimodal-AI. Unlike traditional AI-based interfaces, the feedback that our AI provides to its users is primarily in the form of (electrically actuated) *muscle movements*. This type of unique AI-feedback allowed us to put forward a novel form of **embodied-AI**: an AI that understands & assists with body movements, rather than with language.

Our technical contribution is improving EMS-based interactive systems by leveraging multimodal-AI reasoning to *generate* electrical muscle stimulation instructions that were *not* part of the system's built-in programming. As such, our work builds on EMS research [24] and contributes a generative-EMS architecture that yields new benefits and insights into EMS research. As depicted in Figure 2, we complement our (a) open-source implementation which we hope will accelerate new research in this area with (b) an ablation study that characterized our implementation; and (c) a user study, where we found that, even when such systems fail (e.g., generated incorrect instructions), participants were able to recover and interpret gestures and successfully perform the intended tasks.

Finally, it is important to underscore that our contribution is centered only around proposing, implementing, and measuring the generation of EMS instructions rather than comparing EMS-based assistance against other modalities (e.g., visuals, auditory), or improving upon the existing limitations of EMS (e.g., its practicality). Moreover, while there are many technical routes to implement our concept (e.g., other multimodal-AI techniques, reinforcement learning, simulators), our key contribution is not to exhaustively implement all these alternatives, but to examine the value of instruction generation *for interactive EMS-systems*.

2 Related work

Our work is focused on advancing interactive systems based on EMS. While EMS is not the only actuator capable of physical assistance, we focus on it due to its wearability when compared to mechanical actuators [7, 14, 15, 31, 39, 71, 125]. As we aim to generate EMS-instructions, we review key works in multimodal-AI. Given the wide range of this area, we provide only a succinct review—refer to [5, 18, 109, 111, 124] for extensive reviews.

2.1 Conveying *procedural-knowledge* through haptic feedback

Procedural-knowledge is denoted as the "know-how" of performing a task [17, 105]. It supports the acquisition/retention of manual tasks, and it does not depend just on declarative knowledge or vision [45, 88, 113]. Studies found that procedural-knowledge is learned implicitly via experience—people can learn it even if they

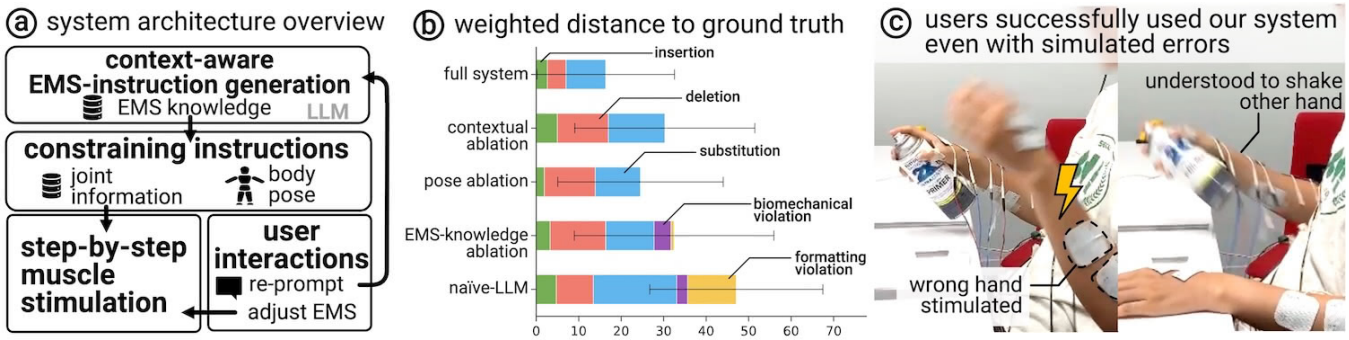

Figure 2: Preview of contributions: (a) our system architecture; (b) ablation study suggests it produced ~3x fewer errors; (c) results from user study suggest they were able to use our system to perform physical tasks, even if instructions contained simulated errors.

cannot articulate it [58]—physical know-how is usually shown or felt rather than described [74]. People tend to *learn by doing*—hands-on experiences involving active manipulation allow them to associate movement instructions and more easily remember the task [26, 67, 69].

One promising way to deliver a hands-on user experience is through *haptic feedback*. For instance, providing movement guidance with haptic feedback benefits learning physical tasks, such as surgical procedures [80, 123], sports [91, 104], welding [121], as well as memorizing movement sequences [16, 74]. In fact, several studies confirmed the added value of haptic feedback with respect to motor learning [28, 65, 70, 94]. With respect to the scope of our work—i.e., a form of haptics known as electrical muscle stimulation—many systems demonstrated the benefits of using EMS to actuate the user's body to demonstrate procedural knowledge [22, 23, 36, 63, 103]. In fact, recent findings suggest that EMS can, in certain contexts, outperform other types of haptic feedback for skill transfer [56]. It is important to underscore that many other modalities can be used to convey procedural knowledge (e.g., visuals, audio); however, these lay outside the scope of our paper, which is focused on *electrical muscle stimulation*.

2.2 Electrical muscle stimulation (EMS) as force feedback

EMS actuates the body by electrically contracting muscles. While it originated in rehabilitation [96], it became a popular choice for force-feedback, as its hardware is smaller than mechanical-actuators [61]. This led to ~150 EMS-publications in HCI [24], such as AR/VR [13, 49, 62, 64], or skill-acquisition [22, 36, 63, 103]. Given our scope, we focus on assistance-based EMS-systems (see [24] for a more general discussion of application areas in interactive-EMS).

Typically, assistance-based EMS-systems convey information by moving the body. For instance, *Affordance++* [63] provides EMS movements for six objects the user might encounter (e.g., movements to shake a spray-can before painting). However, these EMS-instructions are fixed—when a user encounters a new object that requires instructions not found in the system's built-in programming (e.g., open a pill bottle), current generation EMS-systems cannot adjust their movements to tackle this. Similarly, the instructions delivered by traditional EMS-systems are non-contextual. For instance, while *Affordance++* [63] delivers EMS-instructions to operate a spray-can while painting (e.g., it shakes the can), yet, these would fail to assist a user who is spraying cooking oil (no shaking is needed)—in this case, contextual cues could aid in resolving this by including the visuals of the tools or surroundings (e.g., kitchen, frying pan). These contextual cues need not be visual; they could be based on the user's geolocation (e.g., some tools work differently in different regions of the globe) and so forth.

Our research question is as follows: if EMS-systems were given the ability to provide flexible instructions and understand contextual cues, could it widen their scope? (e.g., support users even in situations not explicitly included in their programming?). To embed this novel level of understanding in EMS, we turned to a range of technical approaches, namely, multimodal AI and constraining AI output with additional domain-knowledge (e.g., EMS knowledge).

2.3 Multimodal-AI to reason on unknown objects/situations

Integrating large-language-models (LLMs) with computer-vision has enabled reasoning on more complex visual contexts—leading to vision-language-models (VLMs) [118, 122]. Many models exist that can either reason on visual concepts learned from natural language descriptions (e.g., *CLIP* [90]) or hybridize object-detection with language models (e.g., [35, 117]). An area of rapid progress is refining multimodal pipelines to improve accuracy (e.g., *InstructGPT* [81] and *Self-RAG* [4]— to cite a few). Given the breadth of multimodal-AI, we recommend detailed surveys [5, 41].

Multimodal-AI was leveraged in many interfaces for assistance (an exhaustive list of LLM applications in HCI is provided in [82]), including assisting blind users in recognizing objects [9, 55] or guiding procedural tasks via audiovisual feedback [2, 42, 54]. While audiovisual guidance is important, we focus on a different type of assistance: movement assistance via force-feedback—i.e., demonstrating a task not with audiovisuals but via *movements*.

With respect to generating force-feedback instructions, we can draw parallels to robotics. Multimodal-AI models, supplemented with robotic-specific training or fine-tuning, have been shown to generate task-planning for robotic manipulators [1, 77, 110]. We also use multimodal-AI to generate instructions, but we focus on generating EMS-instructions to be delivered directly to the user's body (not via an external robot), which requires enabling the system to consider the user's current body pose, joint-limits, the capabilities of EMS, etc.

2.4 Balancing general-reasoning with domain-specificity

Foundation models are suitable for reasoning on *general tasks* since, to some degree, they reason through abstract instructions [27], can parse these into steps [54, 68, 114], extrapolate to unseen examples [19, 51], find clues in text or images [118, 119], etc.—all of which are critical properties if a system desires to assist users in a task that is *not* part of its built-in programming. This is the key reason why we leverage LLMs/VLMs to generate instructions. However, as we will detail in *Implementation Study*, foundation models are not designed to output electrical muscle stimulation and can fail/hallucinate when selecting the correct joint to move, request moving a joint past its biomechanical limits, and so forth.

A popular approach to tackle this is to *constrain* the output of the model to generate domain-specific responses. This can be implemented in a number of ways, such as knowledge-graphs [101, 120], filters [59, 116], and so forth. For instance, *CoScript* [116] uses an LLM to generate instructions for tasks but filters the outputs using a similarity score, allowing it to pick options that best match constraints set in the prompt (e.g., "make a cake for diabetics" deprioritizes using sugar); *Scaling Up and Distilling Down* uses success-filtering for LLM-guided robot skill acquisition [33] (this filter learns from successful outputs, prioritizing these); *SHAPE-IT* [89] used retrieval augmented generation to map LLM textual-instructions to pin-based display animations, etc. Inspired by these, we leverage constraints to transform textual-instructions into muscle-stimulation instructions. In our case, the constraints consider *EMS-knowledge* (e.g., user's pose, joint-limits, kinematics).

3 Walkthrough

To illustrate the flexibility that our approach adds to AI interactions, we demonstrate via a walkthrough. In this example, our system acts as a form of **embodied-AI**—a system that, rather than assisting its user with (audiovisual) textual instructions, assists its user by means of generated (electrically actuated) muscle movements.

Figure 3 (a) depicts our user wearing a rendition of our concept (see *Implementation Study*), comprised of an electrical muscle stimulation & motion-tracking suit (*Teslasuit*), alongside glasses with a camera and a microphone (*Meta Ray-Ban*). Our user, who is traveling to a foreign country, is struggling to operate objects. Confused by a window, she asks our system, "EMS, help me open this window for ventilation?" ("EMS" is the wake-up word for invocation).

All interactions in our examples depict unmodified EMS-outputs of running our system as shown (e.g., using these voice prompts, these POV photos, electrodes placed as is, etc.). These two examples (opening a non-traditional window and a pill bottle with a locking-mechanism) are also featured in our studies (see *Ablation Study* and *User Study*.

To assist the user, our system gathers information. It captures her point of view and location (e.g., "Germany", the country she is visiting). Using computer-vision, it extracts objects ("window" and "handle") and hand poses. These are passed to our architecture, which uses multimodal-AI and our EMS-knowledge to generate instructions (see *Implementation Study*). Now, as depicted in Figure 4, our system is ready-to-act and informs the user ("I will move fingers and arm") and *waits for confirmation.* After she confirms

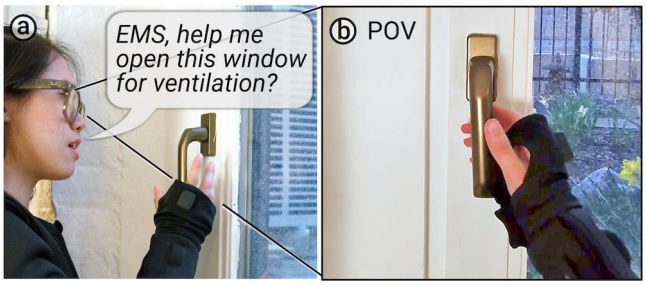

Figure 3: (a) User asks for help with window. (b) Their POV.

("EMS, continue"), it delivers movement instructions via EMS, causing their body to grasp the handle (finger flexion), turn the handle to the upwards position (wrist pronation & shoulder abduction), and pull the handle to tilt the window (elbow flexion).

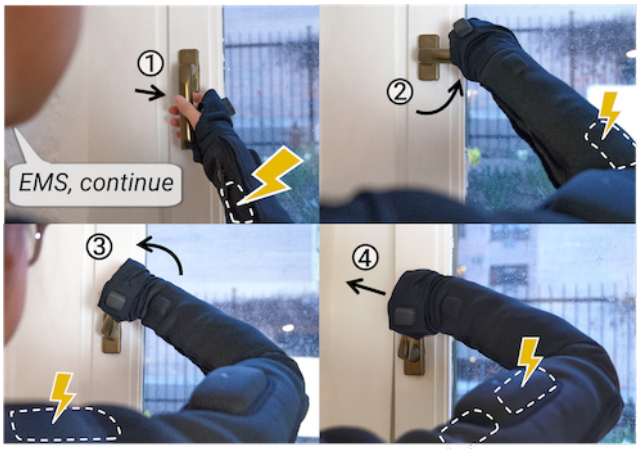

Figure 4: EMS sequence that opens the window in tilt mode.

After this, in Figure 5, our user tries to open a pill bottle they got at the pharmacy, featuring a locking-mechanism unfamiliar to them. Failing to open it, she asks, "EMS, help me". This time, we opted for an example where the user's request is vague (no mention of goal/object); yet our system gathers contextual-cues to generate EMS-instructions.

Figure 5 (c) depicts that, after the user's confirmation, our system delivers the EMS-instructions that cause her hands to squeeze the cap (finger flexion) and turn the cap (wrist abduction). During EMS, our system uses checkpoints to select the next EMS-instruction as needed. For instance, when not seeing that the cap was removed, it adjusts the EMS plan to open the cap in *repetitive* movements (grasp fingers, turn clockwise, release fingers, turn counter-clockwise—repeat). It executes this until it reaches the checkpoint (i.e., determines the cap is out) or the user verbally stops the interaction (e.g., "EMS stop") or the user moves against the EMS feedback (e.g., employed also by [63] as a quick-stop).

However, note that we intentionally complicated this example to illustrate a failure-case and how the added flexibility might allow our system to recover. Despite the EMS-assistance in the last step, the pill bottle did not open. Thus, not seeing the cap removed, the user intervenes: "EMS, it did not open. Try something different?". In response, our system restarts, but appends the output of the

Figure 5: (a) User asks for help with a pill bottle. (b) Their POV. (c) The system can understand some degree of the state of objects, which is used to add "checkpoints" to the muscle-stimulation instructions, e.g., repeat until the cap is removed. (Although, as we will see in the next step of this walkthrough, this will not open this type of pill bottle.)

previous interaction. This allows the reasoning to generate another solution, which is depicted in Figure 6 ("I can push fingers and turn wrist"). The user confirms, and the EMS actuates the body, which causes the bottle cap to open—turns out it was a *push-turn* cap mechanism rather than a *squeeze-turn*.

Figure 6: (a) User intervenes and asks for a different approach, which triggers our system to generate a new sequence of electrical muscle stimulation assistance.

Finally, we purposely illustrated extremely verbose examples in which the muscle stimulation plan is read aloud to the user via text-to-speech. This is not the only way such an embodied-AI system can preview its muscle-based instructions to the user. In fact, our system features two additional operational modes that preview the EMS-instructions using (1) low-intensity EMS (i.e., creates very subtle movements) or (2) sub-threshold EMS (i.e., the simulation creates electro-tactile feedback that suggests movements without physical limb displacements).

4 Implementation

To help readers replicate our design, we provide technical details in this section and open-source all of our codebase in our repository[1] (also available in the *Supplementary Material*). Furthermore, to illustrate key aspects of our implementation, we provide system diagrams and, in *Appendix*, examples that showcase the contribution of a particular module/feature (all examples in this section were created using our complete system; for our technical evaluation, see *Ablation Study*.

First, to illustrate the difference between naïvely generating EMS-instructions using an LLM vs. by leveraging our system, Figure 7 provides detailed output from the *Walkthrough* (i.e., opening the window to air the room).

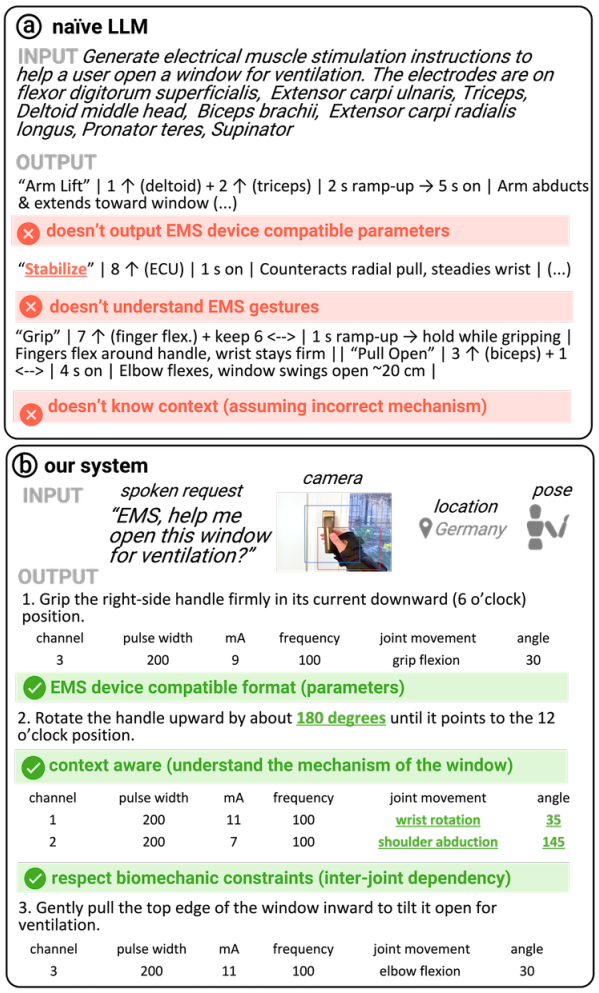

Figure 7: Juxtaposition of naïve OpenAI GPT-4.1 vs. our system.

In both cases above, the LLM is the same (*OpenAI GPT-4.1*); however, the naïve case depicts how foundation models excel at general-reasoning but fail to consider EMS-knowledge (e.g., incorrect use of wrist extensor), contextual-cues (e.g., failed to identify the tilt-turn window), or joint-information (e.g., current wrist pose). Adding these EMS considerations is a technical challenge that our

[1]Code available at https://embodied-ai.tech

system solves. Figure 8 details the system modules needed to implement these capabilities. Also, prompts and EMS-knowledge-base are in *Supplementary Material*.

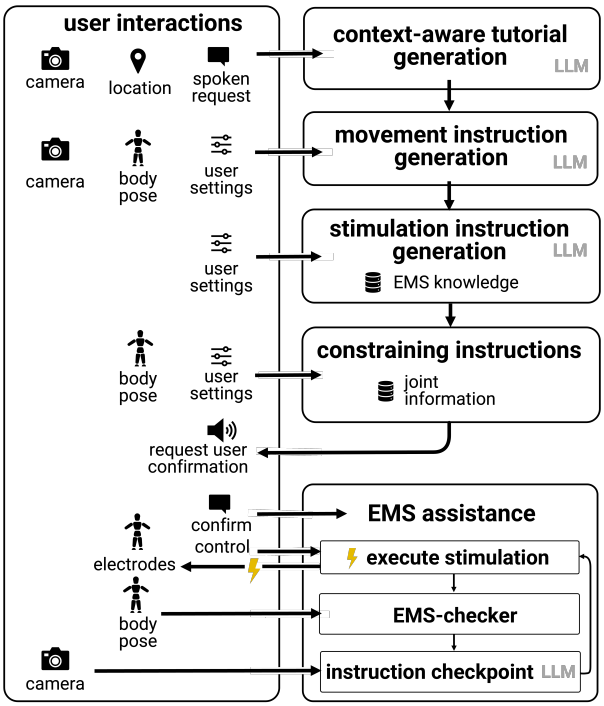

Figure 8: Outline of our system's architecture and each of its implemented modules.

4.1 User interactions

As depicted in Figure 9, this module handles interactions between users and submodules. By piping video and audio from the smartglasses (similar to [38]), our system transcribes speech-to-text (upon detection of the "EMS" wake-up word). Using speech, users can make requests (e.g., "EMS, help me with this") or adjust the delivery of EMS-instructions (e.g., saying "repeat", "slow down/speed up", "pause", "stop", "resume"). Before EMS takes place, the system reads out a summary of the soon-to-be stimulated muscles and awaits user confirmation.

Advanced settings. Users can also configure how the stimulation operates. They can adjust the stimulation mode by saying "EMS stimulation mode: "*actuate*, *nudge*, or *tactile*." These modes change the overall intensity of the stimulation: "actuate" sets the stimulation at the maximum intensity calibrated by the user, "nudge" sets it at a medium-intensity value, and "tactile" applies sub-motor threshold stimulation (not strong enough to cause an involuntary contraction). Users can also change how the system behaves when movements are beyond EMS capabilities by saying "EMS, completion mode: *partial* or *full*". In partial-mode, our system will execute all instructions but skip the ones that it cannot find a suitable EMS-instruction in our EMS-knowledge base (e.g., if there are no electrodes on the left hand, but the task would require left-hand movements)—in this case, it will still say these instructions out loud. Conversely, in the full-mode, the system will stop if the

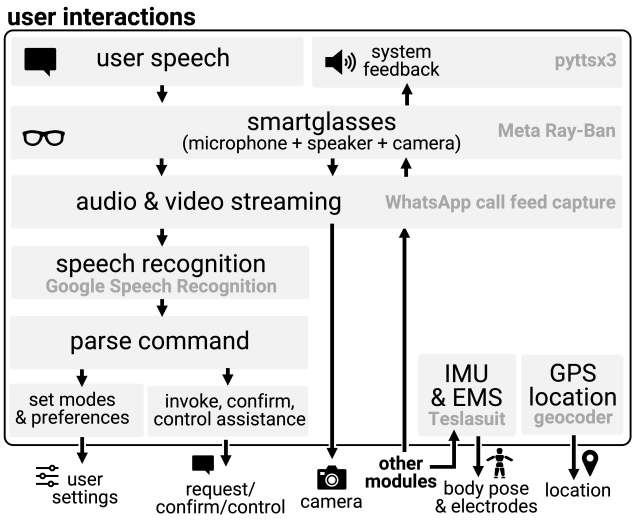

Figure 9: Information flow in the User interactions module.

sequence contains any instructions without an EMS counterpart. Finally, users can add custom settings, which are saved in the system's prompt (e.g., "EMS, user-settings: my right hand has a weak grasp").

4.2 Context-aware tutorial generation

Figure 10 depicts the flow of information in this module, which uses multimodal-AI to *generate* tutorial-like descriptions to perform the requested task. Rather than just taking the user's spoken request, it also appends the following (if available): POV image + location + output of previous interaction. Most of the contextual-analysis is done by the LLM (OpenAI *GPT-4.1*), but to infer object-in-hand, we leverage hand tracking (*MediaPipe)* and object classification (*YOLO*) as LLMs are statistically more biased [40] (e.g., only 10.5% of the population is left-handed [83] and LLMs could hallucinate right-handed instructions, even if an object was grasped on the left). See *Appendix* for comparisons of contextual-analysis outcomes with and without GPS locations, visual contexts, and user requests.

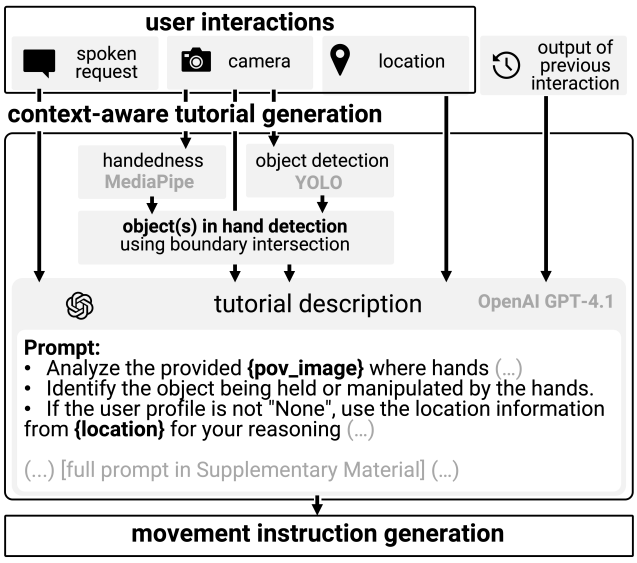

Figure 10: Information flow in the Context-aware module.

4.3 Movement-instruction generation

Figure 11 depicts this module, which uses an LLM to parse the output of the previous module (i.e., a tutorial description) and generate a sequence of specific movements, taking into account body-pose and any added preferences. Current body pose is obtained by a mocap-suit (*Teslasuit*) and by the glasses' camera (*MediaPipe*) for hand pose. Then, to convert poses to text (for LLM), we utilize a rule-based system that describes the pose of each limb along an X-Y-Z axis projected from the user's egocentric-view (e.g., "palm up" if the palm faces the ceiling, and so forth), all computed in Unity3D. See *Appendix* for a side-by-side example of generated movement instructions with and without body pose.

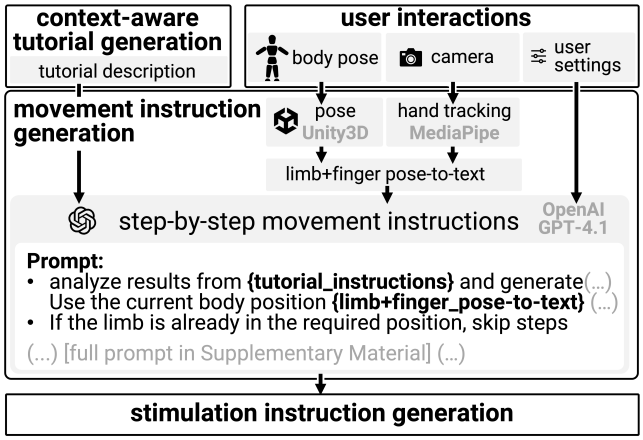

Figure 11: Information flow in Movement instruction generation module.

4.4 Stimulation instruction generation

This module uses an LLM to *select* a set of EMS-instructions from a knowledge-base of feasible EMS-actuations. As depicted in Figure 12, it achieves this by taking the output of the previous module (i.e., movement-based instructions), and for each of these, it selects any number of gestures from the EMS-actuations knowledge-base to achieve this movement instruction. It is worth noting that there are a number of technical ways to achieve this step (including without LLMs, using more traditional text-filtering methods).

Also, as depicted in Figure 7, without such an EMS knowledge-base, a naïve-LLM will not output valid EMS commands that a stimulator can execute (as these have specific EMS-stimulation parameters, e.g., pulse-width, amplitude, etc.).

4.5 Constraining instructions with EMS-knowledge

This module uses a **rule-based system** (no-LLM) to *modify* the output of the previous module (i.e., sequence of stimulation instructions), and takes into account hard-constraints (i.e., a table of joint-limits obtained from [8, 11, 86, 87], the current pose from our tracking, and a list of the human joints as a kinematic-chain) as depicted in Figure 13. For each stimulation instruction: (1) *stop* the instruction if it is not biomechanically possible (e.g., stops if the EMS-instruction requires "turn neck, clockwise, 180 degrees"); (2) *stop* the instruction if the joint's current-angle would not allow it

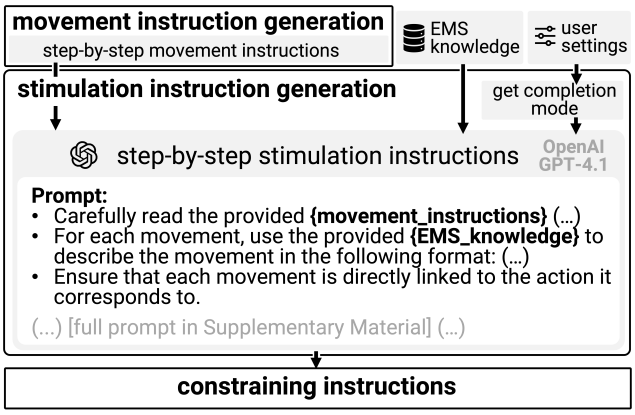

Figure 12: Information flow in Stimulation instruction generation module.

(e.g., stops for the instruction "turn neck, clockwise, 10 degrees" if the neck is already at its maximum angle); and, more generally-speaking, (3) **attempts to *adjust* the instruction** if there is a parent joint (i.e., the joint that is closer to the torso from the current joint) in the kinematic-chain. If this is the case, it adds up the angles of successive movements in upward joints until the target angle was reached or is deemed impossible—for instance, for the instruction "wrist, abduction, 45 degrees" while the user's wrist is fully-extended, it would proceed to the parent joint (i.e., elbow) and examine its current pose (and if not maxed out) would extend it by 45 degrees. This repeats iteratively up the entire kinematic-chain. As we show in our ablation study, this adjustment-step is important since a multimodal-AI might ignore or hallucinate some pose aspects (e.g., individual joint-limits, assume poses, etc.) during instruction generation.

4.6 EMS assistance

This module, depicted in Figure 14, uses a rule-based system and an LLM to, respectively, check if the EMS is moving a joint, and determine if instruction-checkpoints were reached.

Execute stimulation. EMS is delivered via a *Teslasuit* with 80 EMS channels: 16 per each of the four limbs and 16 on the torso (see electrode-placement map in [106]) which is controlled using *Teslasuit*'s *Unity3D* API [107]. Like most EMS devices, this API allows control of the frequency, amplitude, pulse-width, and duration of a stimulation—these are per-user pre-calibrated settings for the desired muscle (see *Constraining* module).

EMS-checker. From the currently actuated-joint and its target-direction, a rule-based system checks if this joint is: **(1) moving in the expected direction:** system continues; **(2) not moving significantly** against a pre-calibrated IMU-threshold: after 3-seconds without movement, EMS is halted and depending on the user-selected completion mode, this stimulation is skipped or the entire interaction stops; or, **(3) moving in the opposite direction**: user opted to cancel the EMS by moving in the direction opposite to EMS-actuation, which stops the interaction. This applies when a collision occurs, as the observed movement is not the same as the instructed one.

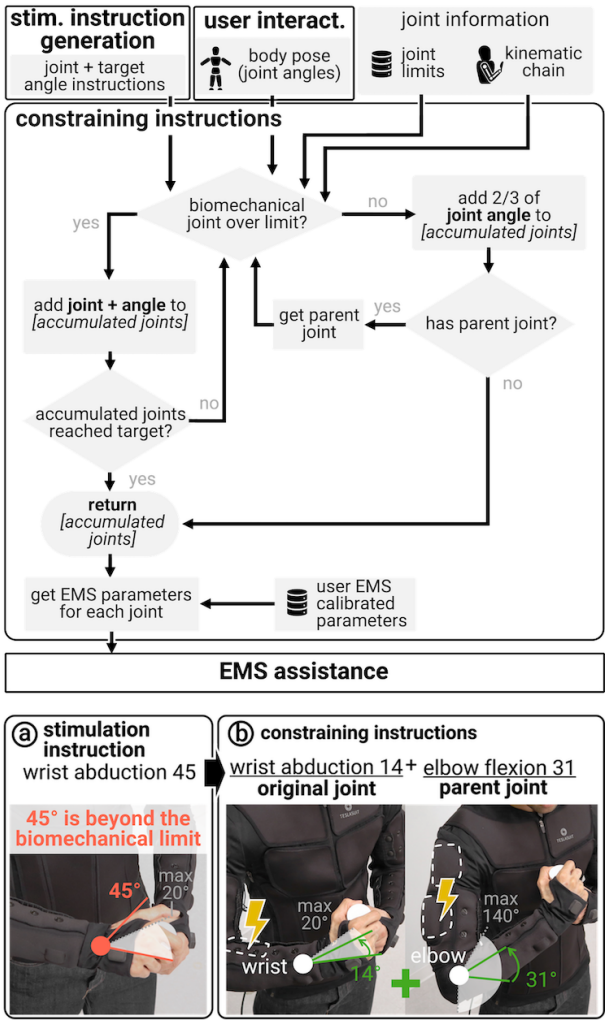

Figure 13: Information flow in the Constraining module and an example of joint constraints in action: (a) without constraints, a generated EMS-instruction might actuate a joint beyond its limits (e.g., wrist past 45°); conversely, (b) the constraining module distributes the angles over the kinematic chain (e.g., moves the wrist by 14° and the elbow by 31°—adding up to the intended 45°).

Checkpoints. After all the EMS-instructions for this movement instruction have been delivered, our system takes a photo of the outcome and uses its LLM to select the next movement instruction (from the previously generated list). This allows the system to engage in a "checkpoint"-like behavior, where it can skip instructions if two steps were achieved in one-go (e.g., potentially due to user's own intervention), or even repeat instructions up to five times (e.g., as seen when the pill bottle cap did not open in the *Walkthrough*).

4.7 Additional safety-considerations

Apart from joint-limit considerations, we implemented typical EMS safety-considerations (e.g., movements calibrated per-user to operate pain-free). Further, the *user interactions* module runs in a separate process, providing always-available control (i.e., "EMS pause/stop" or moving against EMS-actuation, detected by body-pose tracker). Finally, interactions via our full system are always user-initiated and require user-confirmation before stimulation.

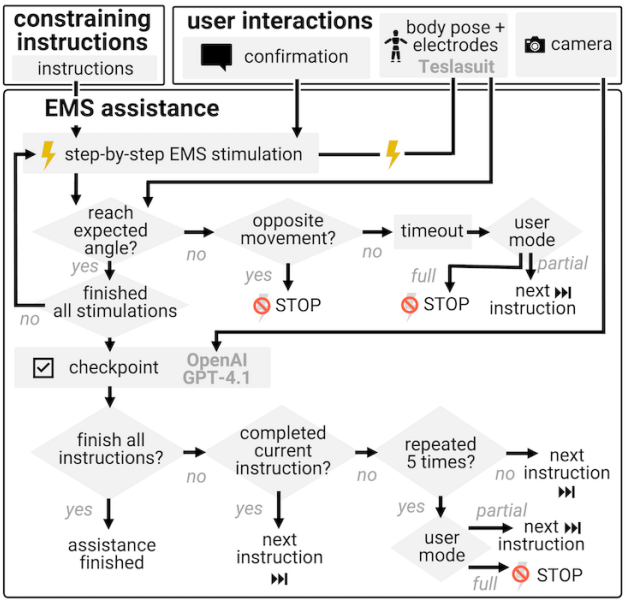

Figure 14: Information flow in the EMS assistance module.

5 Technical evaluation via an ablation study

In this evaluation, we break down the contributions of the modules through an ablation study, where we compare instructions generated by one version of our system (e.g., complete system, ablation of a module, and a Naïve-VLM with our EMS-knowledge base) against a set of ground-truth EMS-instructions. To accelerate research in this area, we provide ground-truth and generated EMS-instructions in *Supplementary Materials*—this is unprecedented in this thriving research area, as no such datasets exist.

5.1 Study design

Ground truth data. These EMS-instructions were handwritten by the authors before the evaluation. The three authors authoring these instructions jointly have 23 years of experience with EMS. For each task's starting point (a photo of a user's hand poses and the objects they see in front), we designed the smallest possible set of EMS-instructions to deliver procedural-knowledge required for the task—as mentioned, accuracy of EMS is not matched with that of humans [20]; thus, our system focuses on gestures that approximate the procedural-knowledge, but do not match human dexterity. Because the ground-truth is written in EMS commands, not in plain language, these were refined & validated using EMS code (i.e., we posed our bodies as shown in Figure 15 and applied EMS to match gestures that would provide procedural-knowledge to a user). This process took ~20 hours. Succinctly, the final instruction sets, provided in *Supplementary Material*, are structured as follows: T1 (task 1) with three gestures (index, wrist), T2 with one gesture (index flexion), T3 with four gestures (index, thumb), T4 with eight gestures (including shoulder, fingers), T5 with two gestures (ankle eversion/flexion), T6 with nine gestures (including shoulder, hand),

T7 with 17 gestures (including shoulder, elbow, hand), T8 with seven gestures (including shoulder, hand), T9 with nine gestures (including shoulder, elbow, hand), T10 with 14 gestures (including shoulder, elbow, hand), T11 with 17 gestures (including shoulder, hand), and T12 with one gesture (wrist extension).

Tasks. We ran our study on 12 physical tasks, depicted in Figure 15, selected from prior work (e.g., EMS [22, 46, 63] and *POV* dataset [29]) and designed to illustrate different aspects of procedural-knowledge: (T1-T4) focused on contextual-clues (e.g., spraying can for painting, spraying can for cooking, using a disposable film camera, picking up nails with a magnetic-sweeper), (T5-T8) focused on body-poses (e.g., unclipping right-bike pedal, opening push-twist pill bottle, using a miter saw with body-pose out of sight, positioning left-hand on golf club), and (T9-T12) focused on biomechanical constraints (e.g., opening a tilt-turn window, removing the pit from an avocado with left hand, cutting vegetables with left hand, opening a bucket of paint). Four tasks were adopted from *Affordance++* [63] (e.g., tilt-turn window, avocado tool, magnetic-sweeper, and paint spray-can), four tasks were from the *ego4D* dataset [29] (e.g., biking, miter saw, cutting vegetables, and paint bucket), three inspired by prior EMS work (e.g., camera [46], golf [22], and a variation on the spray-can [63]), and one crafted by the authors (e.g., opening a pill bottle). Images required as visual prompts for all these tasks were taken either directly from the papers referencing these tasks, including the golf task (T5, T7, T8, T11, and T12 from *ego4D* dataset [29]), or reconstructed by posing the objects on a table alongside the author's hand pose from a POV—see Figure 15. User poses were also reconstructed for each scenario by posing our system's skeletal-model inside *Unity3D*.

Ablation conditions. Per task, five sets of EMS-instructions were generated, one per condition: (1) **full system** (all modules, except checkpoints, since we only fed the starting position for each task in Figure 15); three different ablated-conditions corresponding to removal of a system's module: (2) **contextual ablation** (i.e., removed POV image for reasoning, object recognition, GPS data, and user instruction were replaced by generic "help me"), (3) **user-pose ablation** (i.e., removed user-pose), and (4) **EMS-knowledge ablation** (i.e., database contained EMS-movements but no joint-limits, kinematic-chain, or user preferences such as dominant-hand); finally, for contrast, we also generated instructions using a (5) **naïve-VLM**, our baseline (*OpenAI GPT-4.1* with user request "help me"). Outputs were standardized to *[handedness][limb][movement][target angle]*. Thus, even in the naïve-VLM, we added instruction-format and a reduced-version of our EMS knowledge-base (i.e., only had access to valid EMS movements, but no joint-limits, etc.); otherwise, the VLM would output text that could not be understood by an EMS stimulator, nor compared to the remaining conditions in this study. All conditions had the same inputs, including contextual (POV image, user's spoken instructions, object recognition, geographical data), user-pose data (*Unity3D* generated poses, hand tracking), and EMS-knowledge (joint-limits, kinematic-chain, possible EMS movements), unless ignored in their respective ablated-conditions.

Latency measurement. On average, it took our system an average of 23.6s (SD=6.3) to generate instructions across these tasks, because it uses three LLM-API calls, compared to the average 7.2s (SD=2.7) of the Naïve-VLM (one API call).

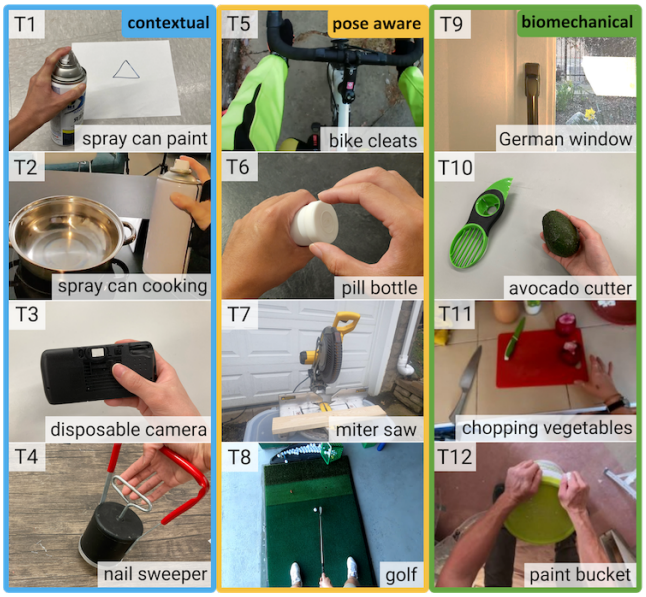

Figure 15: Our 12 physical tasks used in the ablation study, including, (T1) spray-can for painting, (T2) spray-can for cooking oil, (T3) using a disposable camera with film rewinding, (T4) picking up nails with a magnetic-sweeper, (T5) unclipping right foot from a clip-based bike-pedal, (T6) opening push-twist pill bottle, (T7) using a miter-saw, (T8) positioning left hand on golf-club, (T9) open a tilt -turn window (e.g., "German window"), (T10) remove avocado's pit, (T11) cutting vegetables, and (T12) opening a bucket of paint.

5.2 Metrics

To compare the differences between generated-instructions vs. ground-truth, we used a modified *Levenshtein distance* [57]. This depicts the **minimum cost** for transforming a set of generated instructions into those of the ground-truth set, accounting for *insertions*, *deletions*, and *substitutions* at the instruction-level. Each operation was assigned a penalty cost proportional to its impact on movement fidelity. *Insertions* were assigned a cost of 1, since additional instructions may embellish or reinforce a motion without altering it. *Deletions* were assigned a higher cost of 6, as omitting an instruction is likely to result in an incomplete outcome. As for *substitutions* (i.e., partially-matched instructions), a finer-grained comparison was applied by evaluating individual components of the instruction. We consider the following three metrics, each assigned weights to reflect their relative importance in the muscle stimulation output and are cumulative to express the substitution cost: (1) **incorrect-handedness** (cost 2) if the generated movement's instructions occur on the opposite side of the body compared to ground-truth (e.g., left vs. right elbow); (2) **incorrect-joint** (cost 2): if the generated movement's instructions occur on another joint of the body compared to ground-truth (e.g., elbow vs. wrist); and, (3) **incorrect-DOF**: (cost 1): if the generated movement's instructions actuate the wrong degree-of-freedom (DOF) for a target muscle (e.g., flexion vs. extension of the wrist). On top of these operations, two additional violation-types were added: **biomechanical violation** (cost 2) if a movement's angle exceeded the joint-limit, and a

formatting violation (cost 2) when deviating from the required format (e.g., specifying an angle range instead of a single value or omitting the target limb)—malformed instructions were still reformatted when possible for the sake of the comparison analysis but with the violation penalty.

Also, since the number and order of instructions may vary from the generated to the ground-truth, we used dynamic programming to align neighboring instructions and compute the overall minimal transformation cost per comparison.

Finally, while we applied weights to reflect the severity of mistakes, we also report the unweighted distances.

5.3 Results

Figure 16 depicts our key result. For all 12 tasks, we found that the distance between **full system** and **ground-truth** was the lowest. In the following, we present the breakdown for all conditions (i.e., full system, three ablations, and naïve-VLM) and detail the type mismatches between generated instructions and ground-truth.

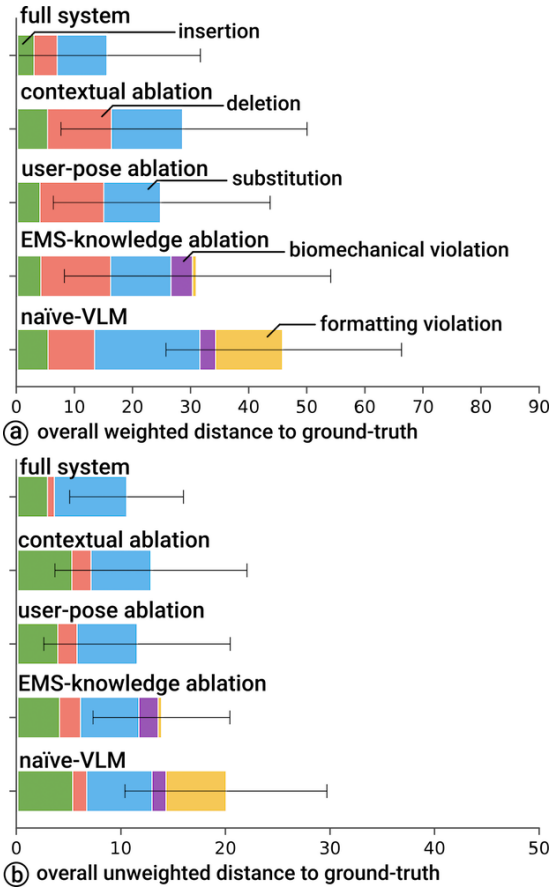

Figure 16: Our ablation evaluation results for all 12 scenarios.

Overall. Our analysis (Figure 16) revealed the following distances, in descending order (i.e., best to worst), between generated-instructions vs. ground-truth: **full system** with an average of 15.4 (10 unweighted), **user-pose ablation** with an average of 24.6 (11 unweighted), **contextual ablation** with an average of 28.4 (12 unweighted), **biomechanical ablation** with an average of 30.8 (13 unweighted), and **naïve-VLM condition** with an average of 45.6 (19 unweighted). These suggest that the full system outperforms the ablations, and the Naïve-VLM (even when it was assisted with the correct syntax). Additionally, ablated conditions also outperformed the Naïve-VLM, showcasing each module's value.

Contextual-tasks. In these, we selected scenarios where cues were useful to derive ground-truth movements. Figure 17 reveals the following distances, in descending order: **full system** with an average of 6.8 (6 unweighted), **EMS-knowledge ablation** with 11.0 (8 unweighted), **user-pose ablation** with 11.3 (6 unweighted), **contextual ablation** with 23.8 (15 unweighted), and **naïve-VLM** with 25.3 (12 unweighted). While performing slightly better than the naïve-VLM, the contextual ablation performed the worst in terms of raw operation count, mostly due to *inserting* more instructions than needed—this was expected, since contextual cues were removed in that condition. To illustrate one such exemplary task, we focus on the disposable-camera task (T3), where a user takes a photo but must first advance the film. In this case, when the contextual-module was ablated (no POV or object recognition), the system failed to recognize that the camera was a disposable analog model requiring film advancement. Instead, it assumed just pressing the shutter. By contrast, all other conditions correctly instructed the user to advance the film before taking a photo.

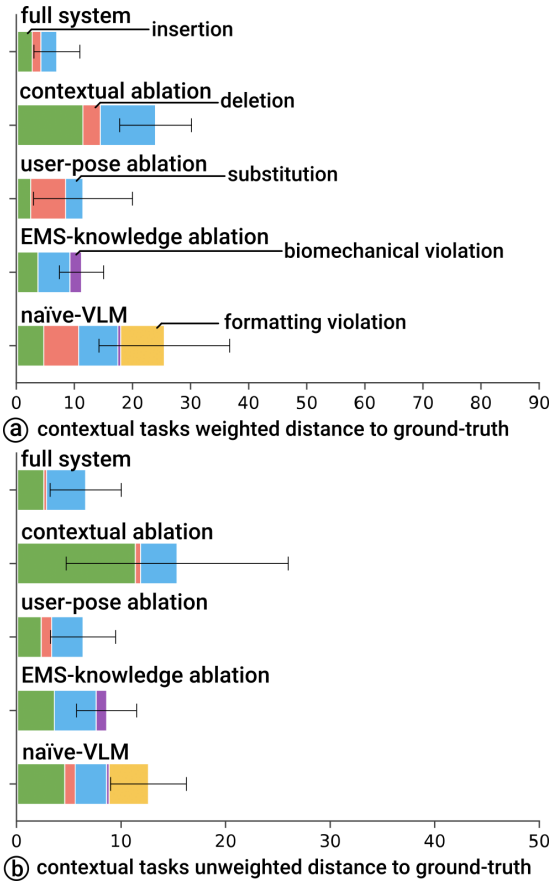

Figure 17: Our ablation results for all four contextual tasks.

Pose-aware-tasks. In these, we examined scenarios where access to pose information was useful for generating correct instructions. Figure 18 depicts the average edit distance, in descending order: **full system** with an average of 19.3 (11 unweighted), **contextual ablation** with 30.8 (9 unweighted), **user-pose ablation** with 35.0 (9 unweighted), **EMS-knowledge ablation** with 39.8 (16 unweighted), and **naïve-VLM** with 49.8 (18 unweighted). While the EMS-knowledge ablation had the highest distance amongst the ablated models, a portion of the penalty was attributed to biomechanical and formatting violations; without those, the user-pose ablation had the highest score—this is mostly due to the *deletion* operations (missing instructions), which could be attributed to the lack of body-pose knowledge. To illustrate one such exemplary task, we take a closer look at T8 (i.e., place left-hand on a golf club, as the right-hand is already on the club). In the user-pose ablation condition, where no pose information was available to the system, it assumed that both hands were already on the club—even when provided with a POV image—and therefore produced no instruction (no correction was deemed necessary). By contrast, all other systems (except for naïve-VLM) suggested actuating the left-hand.

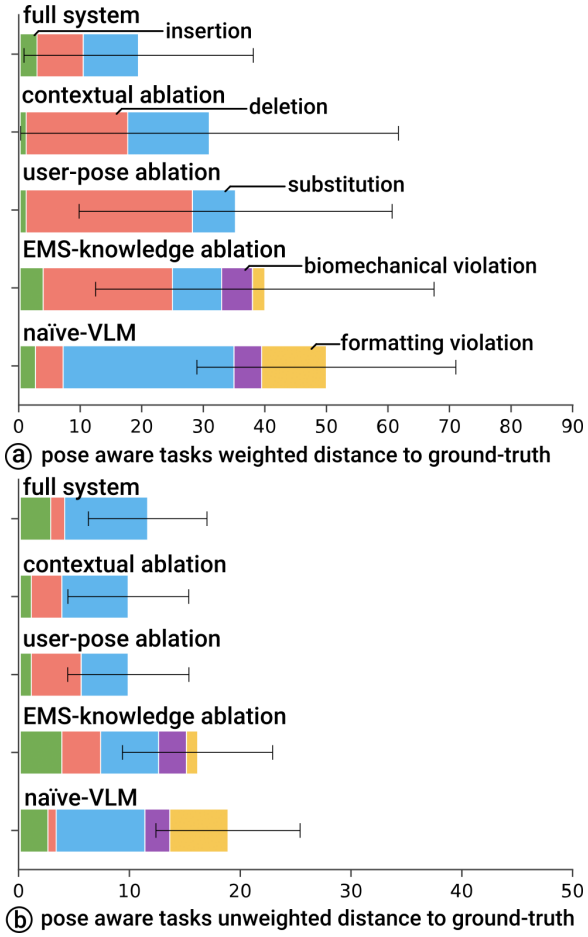

Figure 18: Our ablation results for all four pose-aware tasks.

EMS-knowledge-tasks. We examined scenarios where EMS knowledge was useful for generating correct movements. Figure 19 revealed the following distances, in descending order: **full system** with an average of 20.3 (13 unweighted), **user-pose-ablation** with 27.5 (18 unweighted), **contextual ablation** with 30.8 (13 unweighted), **EMS-knowledge ablation** with 41.3 (16 unweighted), and **naïve-VLM** with 61.8 (28 unweighted). As expected, only EMS-knowledge-ablation showed the highest overall distance, largely driven by biomechanical violations, resulting in the worst among the ablated-conditions. We further noted that only EMS-knowledge-ablation and naïve-VLM had biomechanical violations since they were the only conditions without this knowledge. To illustrate one such exemplary task, we consider T9 (opening a tilt-turn window). This type of window can be opened either from the side—by rotating the handle to the right (3 o'clock position)—or from the top, by rotating the handle upwards (12 o'clock). In this scenario, the user intends to open the window from the top, requiring the handle to rotate from 6 o'clock to 12 o'clock (180° rotation). After correctly identifying this requirement, our system's LLM suggested rotating the wrist 180°. However, it is biomechanically impossible to perform this motion using only the wrist. As expected, the **EMS-knowledge ablated** condition failed to catch this joint-limit violation and proceeded with the instruction. By contrast, all other conditions (except naïve-VLM,) detected this and compensated (via our constraints implementation) by incorporating shoulder and elbow movements.

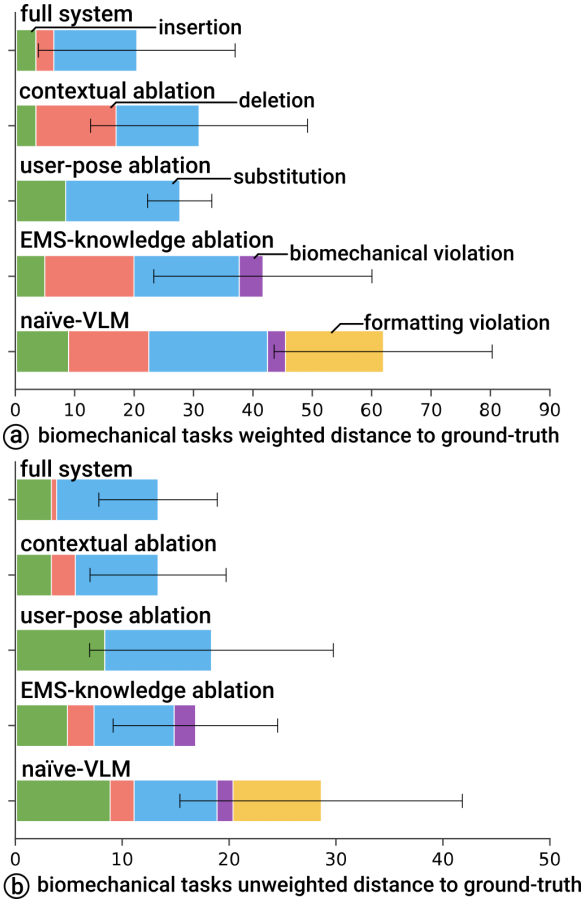

Figure 19: Our ablation evaluation for all four biomechanical tasks.

Evaluation limitations. In this evaluation, we measured output using a distance metric. However, the human body is complex, and different motion trajectories or joint configurations can still achieve the same end goal. There are also more possible ways of defining the ground truth, which is itself subject to the complexity discussed above. Finally, as is typical with these evaluations, the real-time aspects of our system were disabled, such as our *checkpoints* feature.

Summary of findings. We found that our system outperformed the baseline. Moreover, ablations suggest that each module provides a net-positive contribution: (1) absence of contextual-cues led to movement-errors, (2) absence of pose-information generated movements that did not respect the current body-pose; and, finally, (3) absence of EMS-knowledge leads to incorrect EMS-instructions, sometimes even violating physical human constraints (i.e., unergonomic). These technical results highlight that our system never perfectly matched the EMS-expert's ground-truth. In fact, next, we will use these types of errors found in our ablation study, to (purposely) simulate errors in our user study.

6 User Study

In our user study, we examined participants' responses to our system in a laboratory-setting. Since no interactive system will ever perfectly capture the intent of all possible users, and certainly any system based on LLMs is susceptible to hallucinations, we designed this study to **examine not only what happens when our system generates the correct instructions**, but especially, **what happens when the generated instructions are erroneous** e.g., by injecting simulated errors, similar to those observed in our technical evaluation, see *Ablation Study*).

To control so that all participants experience the same set of erroneous instructions, we systematically created different failures-modes by injecting incorrect instructions (e.g., wrong order, wrong limb, nonsensical movement, etc.). This allowed us to observe if participants can not only use our system when its output matches the task, but also in extreme cases where a user is required to be critical of the muscle stimulation output. This allowed us to ask:

(1) *Are users able to detect errors?*

(2) *Upon detecting errors, are users able to ignore errors to make sense of partially-correct instructions*

(2) *Upon detecting errors, do users attempt to refine the output by re-prompting the system?*

(4) *When exploring the system's output, what strategies do users employ?* (e.g., repeat gestures, mimic, slow down)

6.1 Study design

We closely followed the study design of *Affordance++* [63], which was set up to observe users' reactions to EMS-based physical assistance. In this design, for each trial, participants are shown objects and asked to perform a task (e.g., "paint with a spray-can") using the assistance of EMS. Important to this design is that participants are instructed to think-out-loud—this canonical method is often employed to generate more insights that are invisible to experimenters, namely to understand a user's mental model of an interface[10, 37, 78, 95].

User's voice control. While our complete system allows users to control it using real-time voice recognition (i.e., saying the wake-up word "EMS" followed by an instruction or command such as "help me"), we did not want errors from the voice recognition to interfere with this study—our goal was to observe participants' reactions to EMS-instructions, not to measure the accuracy of voice recognition. Thus, we employed the canonical Wizard-of-Oz methodology—*Affordance++* also used this [63]. The only role the experimenter denoted as Wizard took was to trigger pre-cached outputs (i.e., sequences of muscle gestures) as a response to any valid-verbal commands provided by a participant (e.g., "EMS, help me", "EMS, repeat", etc.—as in *Implementation Study*). For instance, when a participant said "EMS, help me", the Wizard initiated the pre-cached stimulation obtained from our system to that prompt. Similarly, when a participant said "EMS, slow down", the Wizard simply pressed a key that calls the slow down function in our code. Additionally, since instructions were pre-cached, this minimized the waiting time of our system when running online (~23.6s).

Instruction generation. All EMS-instructions were generated by our complete system ahead of time and pre-cached for each task. To generate the instructions, our system heard the same verbal prompts related to the task goals that the participants were instructed to complete. Alongside this, our system was shown an image of the object in that trial, including the hands/body-pose of an experimenter next to the object (and if needed, a user's simulated-location, e.g., Germany, when the user encountered the tilt-turn window). Participants were asked to take the starting poses in each task right before the start of trials, as shown in Figure 20.

Trial. For each trial, we asked participants to *think-out-loud* while letting the EMS-instructions guide them, with procedural knowledge, to perform a task in front of them. At the end, they rate how well they understand the steps on a 7-item Likert scale (scale used also in [63]).

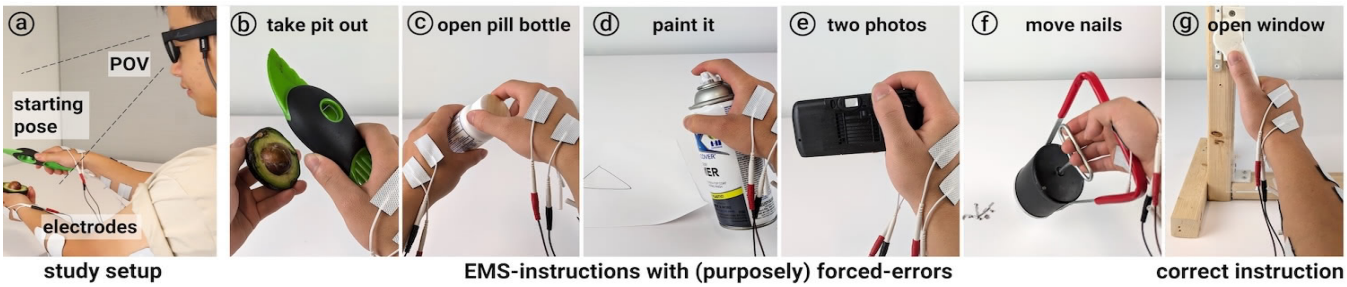

Figure 20: Participants completed six tasks in a random order. (a) shows the study setup. Participants wore smart-glasses and pre-calibrated electrodes and were instructed to start each task with specific starting poses. (b)-(g) shows the starting poses of each task. Task (b)-(f) presented EMS-instructions with purposely forced-errors, and task (g) presented correct EMS-instruction.

Tasks. Figure 20 (a) depicts our six tasks, which are simplified versions of tasks from the *Ablation Study*: (b) "take the pit out of the avocado" (avocado tool)—this was found to be the most confusing step for the participants in *Affordance++* [63], (c) "open a pill bottle" (child-lock pill bottle), (d) "paint this" (spray-can), (e) "take two pictures" (disposable camera), (f) "move the nails to the right" (magnetic-sweeper), and (g) "open the window" (tilt-turn window).

EMS-instructions with forced-errors. Important to our study was to not only observe participants' reactions when our system generates correct instructions, but especially, when the generated instructions are not suitable for the task. Thus, we systematically added five simulated failure-modes that include the errors similar to those we found in our *Ablation Study* and system glitches (e.g., tracking-errors): **(1) Misunderstood-mechanism** (e.g., a child-lock pill bottle was recognized as a standard pill bottle, thus the system did not push the cap before turning it)—this simulates an error in the context-aware tutorial generation module; **(2) Wrong-order** (e.g., system pulled the lever of the magnetic-sweeper, which is a release mechanism, before catching the nails), simulating movement instruction generation errors. **(3) Nonsensical** (e.g., adding unrelated movements to a task)—simulating hallucinations during instruction generation; **(4) Wrong-joint** (e.g., rotation of the hand with the avocado cutter becomes a rotation of the neck)—simulating possible hallucinations/tracking-errors; and **(5) Wrong-handedness** (e.g., actuating the left hand with the shaking movements while the participant holds the spray-can in the right)—simulating hallucinations/tracking-errors.

6.2 Apparatus

Our study used pre-cached EMS-instructions without any sound output, i.e., we did not let participants hear the gesture-confirmation in the study, so that we could observe their reactions to EMS alone. For this reason, we also set EMS stimulation mode to "actuate".

Observations. With participants' prior consent, we filmed their interactions using cameras and microphones.

EMS setup. Electrical stimulation was delivered via a medically-compliant stimulator (*HASOMED*, RehaStim) and pre-gelled electrodes. Prior to the study, the experimenter calibrated the EMS to ensure pain-free operation and reliable movements. A total of 26 electrodes were placed on a single participant's body: right *flexor digitorum superficialis*, right and left *extensor carpi ulnaris*, right *deltoid*, right and left *biceps brachii*, *right extensor digitorum*, right and left *dorsal interosseous* 1^{st} (as in [97]), right *flexor pollicis brevis*, left and right *triceps*, and *splenius capitis* (as in [102]). We added electrodes to muscles that were not needed to perform the tasks; these were used as a *sham* during calibration (i.e., to prevent participants from inferring the set of needed muscles; thus, we calibrated more muscles than needed to perform the tasks—e.g., all tasks could be performed without the *left extensor*).

6.3 Participants

We recruited 12 participants (five female, six male, and one non-binary; average age=25.3; SD=2.2). All participants were right-handed. They receive $12 USD per hour as compensation. Our study was approved by our ethics review board (IRB23-0025). All participants agreed to video, but some requested to blur their faces.

6.4 Results of task completion, usage of verbal commands, and task-comprehension

Overall success rate. Overall, in 92% of all trials, participants accomplished the task goal by following EMS-instructions (e.g., took the avocado pit out). Since we utilized two types of tasks—in one, we first induced a forced-error, and only provided correct instructions if participants re-prompt (tasks 2-6), and in the other, we directly provided the correct instructions (task 1)—we break down this analysis further into its constituents.

Success rate when the correct instruction set was delivered first. Figure 21 (a) shows that all participants succeeded in opening a tilt-turn window (which is an unfamiliar window type in their context). All participants stated they understood from the EMS-instruction how to turn the handle all the way up to pull it open, as P10 stated "[stimulation] in my shoulder is telling me to go up (...) until vertical" (similarly, P3 and P4).

Success rate when forced-error instruction set was delivered first. For the remainder of the tasks (tasks 2-6), first, the system delivered a forced-error EMS-instruction set. Still, participants succeeded in 90% of the trials. In 28% of these cases, participants succeeded without needing to experience the correct instructions (i.e., participants succeeded without re-prompting the system in any way).

Failed trials. We observed one (out of 72) trial where the participant failed to meet the task goal—P8 could not, even after experiencing the EMS assistance, open the pill bottle (which contains a non-trivial locking mechanism). Additionally, during analysis, we

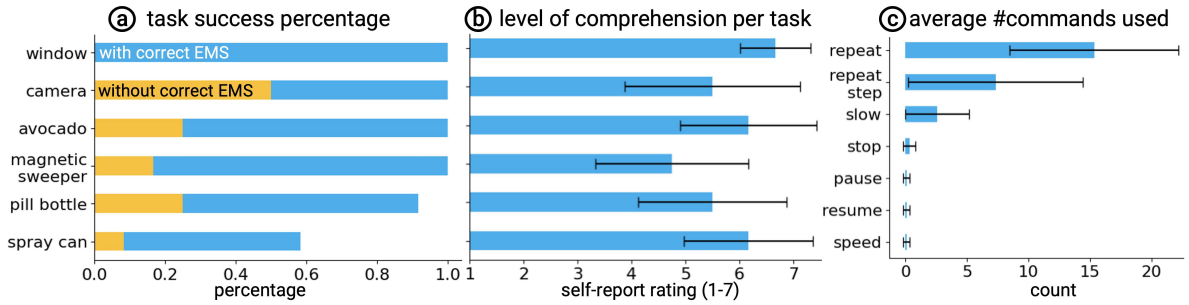

Figure 21: Quantitative results of user study. (a) Task success (%) (b) Task comprehension (self-report) (c) Usage of verbal commands.

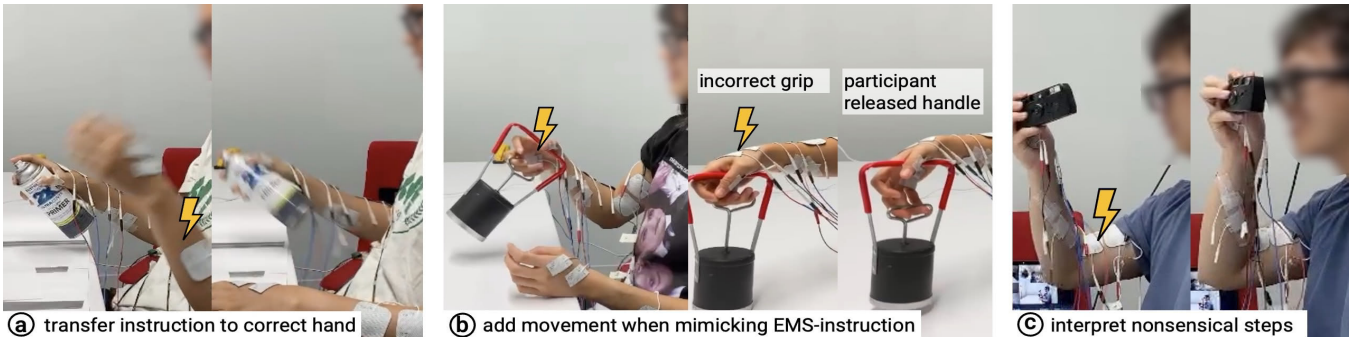

Figure 22: Participants recovered from forced-errors. (a) The participant noticed the instruction was on the wrong hand and transferred the movements to the other. (b) An example of participants adding a release move to recover from the error when mimicking EMS-instructions. (c) A participant making sense of nonsensical elbow movements—looking through the viewfinder.

deemed five spray-can tasks as failed, despite the fact that participants technically met the goal of painting, but they failed to report understanding that shaking was required (P1-3, P6, P12).

Participants did not take EMS at face-value, they explored them. Participants reported they understood the EMS-instructions (M=5.8, SD=1.39), as depicted in Figure 21 (b). We analyzed how participants interacted with our system by counting the number of commands they used across all tasks. Figure 21 (c) shows that participants often used "repeat" (M=15.3 times, SD=6.87), "repeat a step" (M=7.3, SD=7.11), and sometimes "slow down" (M=2.6, SD=2.61) commands to explore EMS-instructions. We further analyzed the correlation between task comprehension and the number of commands used per task, and found no significance (R^2=0.26, p=4.42). This indicates that the participants utilized verbal commands not just for challenging tasks, but also to explore the system (e.g., after successfully opening the window with EMS, P5 closed it and said he would do it again with EMS, visibly exploring our system).

6.5 Participants' strategy for recovering from errors and understanding EMS-instructions

We transcribed participants' comments when thinking aloud and analyzed their strategies of recovering from erroneous EMS-instructions and how they understood correct EMS-instructions.

Participants recovered without re-prompting. We observed that in almost one-third of trials (28%), participants succeeded by experiencing the forced-error EMS-instructions and finding their way around them. We analyzed participants' think-out-loud processes and found three main strategies of recovering from erroneous instructions. **(1) Add movements when mimicking EMS-instructions** (P3, P9, P10, P12). For instance, P9 found that the magnetic sweeper did not pick up the nails when EMS made her pull the handle, so she added a release action on her own to figure out the mechanism. **(2) Used partially correct instructions and/or ignored errors** (P1, P3, P4, P9, P10, P12). After succeeding in the camera task, P3 stated, "I do not understand why it also moves the camera [shows the wrong gesture], (...) I understand it tells me to move the dial, click to take a photo". **(3) Participants made sense of erroneous steps** (P1, P4, P6, P11). These participants found new meanings in what we intended to be a forced-error, for example, P4 interpreted the nonsensical elbow movements in the camera task as "it's trying to let me look through this window [viewfinder]", P6 thought they might be for "checking something," and P11 described them as "put camera down" and "put it back". **(4) Transfer the instruction to the correct joint** (P10). In the spray-can task, P10 explicitly pointed out the error, ignored it, and corrected it, stating, "EMS demonstrates it on my left hand, so I'll do that on my right hand". See Figure 22 for study footage of how participants recovered from forced-errors.

Participants noticed errors and re-prompted. In roughly two-thirds (65%) of the trials, participants recovered by re-prompting the system (i.e., saying "EMS think again"). Before re-prompting, we observed behaviors that strongly suggest detected errors: **(1) Judge if EMS-instructions lead to task outcome** (P1, P2, P4, P8, P10, P11, P12), participants would let EMS perform the task, then, as they saw the task was not completed, they would re-prompt. In the spray-can task, P2 found that the movements in the left hand were just "shifting the paper around" and said, "I am pretty sure this isn't right (…) EMS, think again." P4 tried following EMS-instruction to turn the pill bottle open, and found "it's not working, so EMS think again." **(2) Identify error, re-prompt immediately** (P1, P2, P3, P5, P6, P7, P8, P12), participants would find a logical mistake or an unexpected movement for a task, they would instantly re-prompt. For instance, in the camera task, P2 stated "EMS repeat second step [the erroneous step], okay, this makes no sense. (…) Maybe EMS, think again", or, in the magnetic-sweeper task, after EMS mistakenly indicated a pull action at the first step, P6 pointed out "it's a pull (...) pull this now? (…) no (…) EMS think again".

Participants' understanding of correct EMS-instructions. Finally, we turn our analysis to what participants understood from experiencing the correct EMS-instructions. We found that, despite never being extensively explained the limitations of EMS in force-production or pose-accuracy, they: **(1) Follow EMS-cues to explore challenging situations** (P2, P6, P8). Many of our participants (~25 years old) were not familiar with disposable cameras, which require rewinding the film after every shot. This made the task challenging for them. For instance, after experiencing EMS-instructions for the camera, P2 noticed that "it [the mechanism] has something to do with these two fingers [index and thumb] (...) Oh, I guess there's this little dial here [!]" Similarly, P8, who has never used an analog camera before, asked EMS to repeat the

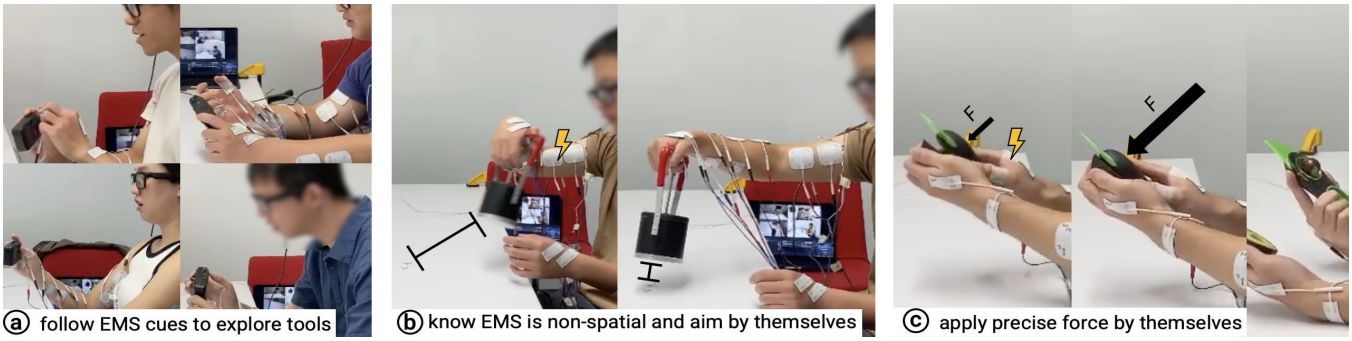

Figure 23: Participants understood correct instructions. (a) Examples of participants following EMS-instructions to discover the mechanism on an analog camera. (b) Participants aiming the nails by themselves as they understood EMS-instructions are non-spatial. (c) Participants adjusted the force on avocado pit as they realized EMS-instructions did not indicate precise forces.

rewind gesture and found "There's a wheel over here". **(2) Understood EMS-instructions were non-spatial, adjusted aim** (all 12 participants). For instance, when mimicking EMS-instructions to pick up the nails with a magnetic sweeper, P12 said, "Previously, maybe I didn't move the tool close enough to the nail, and that's why it didn't stick." **(3) Understood EMS-instructions did not indicate precise force and adjusted the force** (P6, P8, P12). P8 and P12 depicted canonical examples of this, with both mentioning that although "EMS did not indicate it", they found the first step of taking the avocado pit out "requires more force", which they voluntarily enacted. Similarly, for the pill bottle (e.g., EMS often does not have the required force to completely open the mechanism), participants mimicked the EMS-instruction by themselves with additional force. For instance, P6 succeeded this way. **(4) Understood EMS-instructions differ from their own** (P4). It was clear to P4 that EMS' movement trajectories are not human-like (e.g., EMS movements are much more isolated and typically use fewer muscle groups). To this end, P4 pointed out he usually uses the wrist, rather than EMS's choice (elbow), to turn bottle-caps, but still tried the EMS movement nonetheless to successfully open the cap. See Figure 23 for study footage of how participants understood EMS-instructions.

6.6 User study discussion

We observed how participants utilized EMS-instructions and worked around system limitations (e.g., hallucinations, EMS imprecision, etc.). We found that participants not only made use of correct instructions (e.g., all participants opened a tilt-turn window) but also detected and recovered from simulated system errors, inspired by those we observed in our *Ablation Study*. Our study showed the value of providing procedural knowledge through users' own physical movements—it is understandable (participants rated 5.8 out of 7 points on how well they understood the instructions), and it made error-detection easier for the participants (roughly two-thirds (65%) of the trials, participants were able to meet the task goal by detecting errors and re-prompt.) P12 further stated that the "body's intuitions help notice errors right away", whereas text instruction requires reading it and sometimes referring to a fact-check to detect errors. Participants also noted our system is helpful for unfamiliar tools (P1, P5, P8), and appreciated its general-purpose aim (P5, P12).

Study limitations. Our in lab-study is not without limitations. To let all participants experience the same set of simulated-errors, we pre-cached outputs and asked them to assume starting positions. Any variations outside the study will affect results (e.g., different spoken requests and poses), so we advise mindful extrapolations of our in-lab results.

7 Generate new EMS-assistance with a single EMS-system

We further demonstrate how our system can provide physical assistance without task-specific programming. In the example of Figure 24, a user is practicing cycling with cycling-shoes for the first time—these lock into the pedals and require a heel twist to disengage. The user is also wearing our full system (*Teslasuit* and smart-glasses).

Figure 24: Our system uses (a) user's spoken request; and (b) POV image, to (c) provide physical assistance through muscle stimulation of her heel, enabling this user to (d) successfully disengage from the bike pedal.

To practice safely, she clips only on her right foot, keeping the other on the ground. As anticipated, she cannot get her right foot off the clip-pedal, so she asks, "EMS help me disengage from the bike cleat" (Figure 24). Using contextual clues (e.g., POV/pose), our system recognizes that she is wearing cycling-shoes and the right foot is clipped into the pedal; it infers that the heel needs to be twisted. It confirms by asking, "I am going to move your right foot". Upon confirmation, the system proceeds to stimulate her calf (*peroneus longus*) to twist her ankle, disengaging the foot from the pedal.

Figure 25: Our system assists a user in placing a bike on a bus rack. (This mechanism is so challenging that the local transportation agency has created an entire life-sized practice location for citizens to practice—where we used our system.)

Still feeling like a novice in these cycling-shoes, she decides to take the bus. However, this is also the first time that she is loading her bike onto the bus and she is unfamiliar with the bike rack mechanism. Once again, she asks for assistance: "EMS help me with this bike rack" as shown in Figure 25 (a). After confirmation, the system proceeds to (1) stimulate the biceps to reach the handle of the bike rack, (2) flexes her hand to grab the handle, (3) squeezes the fingers to unlatch the rack, and finally (4) contracts the triceps to lower the rack; similar to our user study, she understands that the EMS cannot actuate her entire body and completes the remaining movements, such as bending the legs to help lower the rack (a video version of this example is available at https://embodied-ai.tech).

8 Discussion

8.1 Ethics

As the first paper that demonstrates a more general-purpose EMS-system, our work inherits the ethical concerns of both physical assistance [3, 30] and AI [34, 48]. We thus designed our system with these in mind: while it uses involuntary movements, it still preserves users' autonomy. First, it is user-initiated—EMS does not activate unless the users invoke. Second, our *user interactions* module runs in a separate process, providing always-available control (e.g., users can say "EMS stop" when they detect errors, as we observed in the user study). Third, our system awaits user confirmation before stimulation. Moreover, our system design followed the established EMS guidelines [52, 79, 92] for safety.

8.2 Limitations

As the first exploration in generating EMS-instructions, our work is not without limitations. In this section, we discuss them under different lenses, such as conceptual limitations (e.g., alternative techniques to implement variants of our proposal, latency, etc.) and practical limitations (e.g., limited scope of our examples, or the limited practicality of EMS).

Alternative system implementations. There are many technical routes to achieve these goals (e.g., reinforcement-learning, advanced biomechanical simulators). Our key contribution is not to exhaustively implement all these options but to demonstrate the value of embedding instruction generation ability in EMS-systems.

LLM limitations. With any system based on LLMs, there is a chance for "hallucinations"; while our EMS-knowledge-base and joint-information constraints will block some errors in instruction generation, they will affect the selection of stimulation instructions. At the time of writing, the most advanced cloud-VLM (Gemini) scored a 61.66% at an image description benchmark, compared to 84.78% for human annotators [25]. Locally-run VLMs scored lower at 50.06%. Therefore, our system is limited by the reasoning capabilities of these VLM to give an accurate contextual description. Our work relies on OpenAI GPT-4.1, which will likely be surpassed by newer models. While this presents a limitation of our system, it is also an advantage—future models are expected to offer higher accuracy (fewer hallucinations) and reduced latency, both of which constrain our current approach. In light of this limitation, it is important to underscore that our contribution is not advancing multimodal-AI but using it to advance interactive-EMS.

Latency. We run a state-of-the-art model (~175 billion parameters [6]) with general-reasoning capabilities, but it does *not* run locally nor fast. Accessing it via an internet-API adds latency (e.g., in our *Ablation Study*, the average latency was ~23.6s). In fact, in our video figure, we cached the API responses to skip waiting on LLM-API calls. Future renditions might run smaller-models locally (e.g., Llama) in devices with high GPU-parallelism.

Example scope. No range of examples can cover all the ways in which a user might ask for assistance; our goal is not to cover these exhaustively but to explore how an EMS-system could *generate* instructions. Thus, we narrowed down examples to a subset of situations that (1) require only movement-based assistance (e.g., a user might find it useful to feel their muscles moving to learn how to operate an object)—conversely, we did not focus on examples where audiovisual instructions would have been more beneficial (these are covered by LLMs already); (2) fit the reasoning capabilities of our LLMs (e.g., do not need access to hidden information not available in training data or its sensors); (3) insensitive to our system's latency

(e.g., no fast actions, no need to react timely to external events, and so forth); and, (4) fit the capabilities of our EMS (e.g., precise-movements are still considered challenging with state-of-the-art EMS).

Practicality of EMS. Like any interactive system based on EMS, ours inherits the well-documented limitations found in prior work [24, 85]. There are two types of important EMS limitations: those that arise from the use of electrodes on the skin and those that arise from the underlying principle of how currents propagate inside the body to stimulate the muscles. The former includes limitations such as the need for manual electrode placements, the need for manual calibration, and "tingling" sensations at the electrodes. The latter includes the fact that many muscles are hard to access by means of electrical currents (e.g., deeper muscles) and that many complex movements that underlie everyday, manual skills have not yet been demonstrated with EMS. In our implementation, we explored technical avenues to circumvent some of these limitations, such as simplifying electrode placement by using an EMS *Teslasuit* (electrodes placed inside in fixed locations) to demonstrate applications, and leveraging the state-of-the-art in electrode layouts (e.g., [16, 97]) to achieve the hand movements required for all of our applications and study. However, as we discuss next, since these limitations are well-known in EMS research, they are also active areas of research where progress is underway.

8.3 Future work

The aforementioned limitations are not insurmountable. We discuss how the landscape of AI and EMS might evolve.

Possible improvements on AI reasoning. First, regarding AI, just a decade ago, much of the reasoning displayed by contemporary models was unthinkable. Now, it is expected that newer models will minimize hallucinations [66] and latency [115]. Despite projected advances, some argue that models might never be entirely free of hallucinations [44]. We expect that to convert textual-instructions to EMS, future systems might still rely on concepts from our implementation (e.g., constraining with EMS-knowledge).

Possible improvements on EMS practicality. Similarly, for EMS, the capabilities have improved remarkably. Two decades ago, the first interactive system making use of EMS only actuated a user's biceps without any interactive closed-loop control [53]. Just five years later, researchers demonstrated the interactive systems capable of actuating movements in smaller joints, such as in the user's wrists or fingers [100]. After that, EMS in HCI rapidly evolved, with improved control-loops (e.g., [43, 47, 112]), independent-finger actuation [97], and new body-areas (e.g., head [102], legs [84], and feet [36]). With each system, some technical hurdles were improved, from calibration [50, 108], form-factor [98], or control-loops [76]. While there are still considerable gaps in practicality and accuracy (see *Limitations*), we expect the processes of electrode placement, calibration, and the sophistication of control loop algorithms for EMS to improve significantly in the next decades, which, ultimately, might enable more complex EMS-movements. Interestingly, as EMS capabilities increase, programming the stimulation patterns becomes harder. Thus, a system such as ours that can generate EMS-instructions without specific-programming might act as a platform to accelerate research.

Contrasting EMS with other modalities. Multimodal-AI has been applied to different interfaces for assistance (e.g., visual and audio interfaces) [2, 9, 42, 54, 55]. Since our contribution was to endow EMS-systems with the ability to generate their own movement instructions, we did not study the efficacy of EMS against other modalities (e.g., visuals, auditory) with respect to physical assistance. We believe this is an important area for future EMS research, as it might shine light on where EMS is useful and where it breaks down for communicating movement instructions. Moreover, as is typical in multimodal haptics research (e.g., [28, 65, 70, 94]), it is very likely that combining EMS with additional modalities (e.g., visuals, audio) will yield synergistic effects that might surpass either modality in isolation.

Embodied-AI for skill acquisition. While we engineered an on-body system that leverages multimodal-AI for flexibly generating muscle instructions for physical assistance, we did not measure whether this new type of embodied-AI can improve physical skill acquisition and "muscle-memory" retention.

9 Conclusion

We engineered an embodied-AI system that assists users with electrical muscle stimulation (EMS) by delivering muscle-movements akin to the procedural-knowledge needed to complete a physical task; except, unlike prior EMS-systems, our assistance is not part of the system's programming. Instead, instructions are *generated* to *contextually* assist the user. In this paper, we provide the first instantiation of this novel concept for EMS, which we investigated via a technical evaluation and a user study.

Given the rapid evolution in interactive EMS [24, 60], we believe that such an architecture will yield a number of benefits for the field of HCI: (1) it provides a platform for researchers to generate EMS-instructions for tasks that have never been explored with this type of haptic assistance before; (2) it can replicate some of the prior work on EMS assistance (e.g., [58]) enabling newcomers to this field to explore these ideas rapidly; (3) it offers a new conceptual way to control EMS-systems—in our approach, the user controls the system with their voice (re-prompt it, rewind steps, slow it down—which we observed participants taking advantage of in our user study). Taken together, we believe that our EMS architecture enables a new territory for the area of embodied-AI, aiming to inspire researchers to strive for flexible and context-aware physical assistance. We believe this powers new modes of reasoning with AI that are not just purely audiovisual, but instead focus on delivering *direct* feedback to the user's body.

10 Use of AI in this paper

While this paper is about using multimodal-AI to advance EMS-systems, no AI was used in writing, figures, or ideas.

Acknowledgments

We would like to thank our colleagues: Jas Brooks, for helping with the video; Andre de la Cruz, for the help with user study; finally, Jasmine Lu and Lonnie Chien, for proofreading the paper. We would also like to extend our gratitude to Teslasuit (https://teslasuit.io/), especially Andrei Pyko, for providing the EMS-suits used in some of the figures and video (although the reader should note that the

presented system is agnostic to the muscle stimulation hardware. In fact, all studies were performed using our custom hardware.)

Funding. This work was supported by multiple funding sources at different times of our development cycle, including: NSF Grant 2047189, Sony Research Award Program, University of Chicago Women's Board, and the AI Pillar: Human-Machine Creativity at the University of Chicago.

References

[1] 2025. Introducing Gemini Robotics and Gemini Robotics-ER, AI models designed for robots to understand, act and react to the physical world. https://deepmind.google/discover/blog/gemini-robotics-brings-ai-into-the-physical-world/

[2] Riku Arakawa, Hiromu Yakura, and Mayank Goel. 2024. PrISM-Observer: Intervention Agent to Help Users Perform Everyday Procedures Sensed using a Smartwatch. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. ACM, Pittsburgh PA USA, 1–16. doi:10.1145/3654777.3676350

[3] Thomas Arnold and Matthias Scheutz. 2017. Beyond Moral Dilemmas: Exploring the Ethical Landscape in HRI. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction (HRI '17)*. Association for Computing Machinery, New York, NY, USA, 445–452. doi:10.1145/2909824.3020255

[4] Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. https://openreview.net/forum?id=hSyW5go0v8

[5] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2019. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 2 (Feb. 2019), 423–443. doi:10.1109/TPAMI.2018.2798607

[6] Fabian Barreto, Lalita Moharkar, Madhura Shirodkar, Vidya Sarode, Saniya Gonsalves, and Aaron Johns. 2023. Generative Artificial Intelligence: Opportunities and Challenges of Large Language Models. In *Intelligent Computing and Networking*, Valentina Emilia Balas, Vijay Bhaskar Semwal, and Anand Khandare (Eds.). Springer Nature, Singapore, 545–553. doi:10.1007/978-981-99-3177-4_41

[7] Craig Carignan, Jonathan Tang, and Stephen Roderick. 2009. Development of an exoskeleton haptic interface for virtual task training. In *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*. 3697–3702. doi:10.1109/IROS.2009.5354834

[8] Cristina Magda Racu Cazacu and Ioan Doroftei. 2014. Structural and Kinematic Aspects of a New Ankle Rehabilitation Device. *Applied Mechanics and Materials* 658 (2014), 507–512. doi:10.4028/www.scientific.net/AMM.658.507

[9] Ruei-Che Chang, Yuxuan Liu, and Anhong Guo. 2024. WorldScribe: Towards Context-Aware Live Visual Descriptions. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology (UIST '24)*. Association for Computing Machinery, New York, NY, USA, 1–18. doi:10.1145/3654777.3676375

[10] Elizabeth Charters. 2003. The Use of Think-aloud Methods in Qualitative Research An Introduction to Think-aloud Methods. *Brock Education Journal* 12, 2 (July 2003). doi:10.26522/brocked.v12i2.38

[11] Jasper Chen, Alan B. Solinger, Jacques F. Poncet, and Charles A. Lantz. 1999. Meta-Analysis of Normative Cervical Motion. *Spine* 24, 15 (Aug. 1999), 1571. https://journals.lww.com/spinejournal/fulltext/1999/08010/Meta_Analysis_of_Normative_Cervical_Motion.11.aspx

[12] Xi Chen and Sunil K. Agrawal. 2013. Assisting Versus Repelling Force-Feedback for Learning of a Line Following Task in a Wheelchair. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 21, 6 (Nov. 2013), 959–968. doi:10.1109/TNSRE.2013.2245917

[13] Chia-Yu Cheng, Yu Chen, Sitaresmi Wahyu Handani, Avijit Balabantaray, and Mike Y. Chen. 2024. Paired-EMS: Enhancing Electrical Muscle Stimulation (EMS)-based Force Feedback Experience by Stimulating Both Muscles in Antagonistic Pairs. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–7. doi:10.1145/3613904.3642841

[14] Inrak Choi, Nick Corson, Lizzie Peiros, Elliot W. Hawkes, Sean Keller, and Sean Follmer. 2018. A Soft, Controllable, High Force Density Linear Brake Utilizing Layer Jamming. *IEEE Robotics and Automation Letters* 3, 1 (Jan. 2018), 450–457. doi:10.1109/LRA.2017.2761938 Conference Name: IEEE Robotics and Automation Letters.

[15] Inrak Choi and Sean Follmer. 2016. Wolverine: A Wearable Haptic Interface for Grasping in VR. In *Adjunct Proceedings of the 29th Annual ACM Symposium on User Interface Software and Technology (UIST '16 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 117–119. doi:10.1145/2984751.2985725

[16] Siya Choudhary, Romain Nith, Yun Ho, Jas Brooks, Mithil Guruvugari, and Pedro Lopes. 2025. Adaptive Electrical Muscle Stimulation Improves Muscle Memory. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI'25)*. Association for Computing Machinery, Yokohama, Japan, 1–11. doi:10.1145/3706598.3713676

[17] Neal J. Cohen and Larry R. Squire. 1980. Preserved Learning and Retention of Pattern-Analyzing Skill in Amnesia: Dissociation of Knowing How and Knowing That. *Science* 210, 4466 (Oct. 1980), 207–210. doi:10.1126/science.7414331

[18] Xin Luna Dong, Seungwhan Moon, Yifan Ethan Xu, Kshitiz Malik, and Zhou Yu. 2023. Towards Next-Generation Intelligent Assistants Leveraging LLM Techniques. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. ACM, Long Beach CA USA, 5792–5793. doi:10.1145/3580305.3599572

[19] Vishnu Sashank Dorbala, James F. Mullen, and Dinesh Manocha. 2024. Can an Embodied Agent Find Your "Cat-shaped Mug"? LLM-Based Zero-Shot Object Navigation. *IEEE Robotics and Automation Letters* 9, 5 (May 2024), 4083–4090. doi:10.1109/LRA.2023.3346800

[20] Barbara M. Doucet, Amy Lam, and Lisa Griffin. 2012. Neuromuscular electrical stimulation for skeletal muscle function. *The Yale Journal of Biology and Medicine* 85, 2 (June 2012), 201–215.

[21] Cosima du Pasquier, Lavender Tessmer, Ian Scholl, Liana Tilton, Tian Chen, Skylar Tibbits, and Allison Okamura. 2024. Haptiknit: Distributed stiffness knitting for wearable haptics. *Science Robotics* 9, 97 (Dec. 2024), eado3887. doi:10.1126/scirobotics.ado3887

[22] Sarah Faltaous, Aya Abdulmaksoud, Markus Kempe, Florian Alt, and Stefan Schneegass. 2021. GeniePutt: Augmenting human motor skills through electrical muscle stimulation. *it - Information Technology* 63, 3 (July 2021), 157–166. doi:10.1515/itit-2020-0035

[23] Sarah Faltaous, Anna Hubert, Jakob Karolus, Steeven Villa, Thomas Kosch, and Pawel W. Wozniak. 2022. EMStriker: Potentials of Enhancing the Training Process of Racket-based Sports via Electrical Muscle Stimulation. In *Sixteenth International Conference on Tangible, Embedded, and Embodied Interaction*. ACM, Daejeon Republic of Korea, 1–6. doi:10.1145/3490149.3505578

[24] Sarah Faltaous, Marion Koelle, and Stefan Schneegass. 2022. From Perception to Action: A Review and Taxonomy on Electrical Muscle Stimulation in HCI. In *Proceedings of the 21st International Conference on Mobile and Ubiquitous Multimedia*. ACM, Lisbon Portugal, 159–171. doi:10.1145/3568444.3568460

[25] Aaron Foss, Chloe Evans, Sasha Mitts, Koustuv Sinha, Ammar Rizvi, and Justine T. Kao. 2025. CausalVQA: A Physically Grounded Causal Reasoning Benchmark for Video Models. doi:10.48550/arXiv.2506.09943 arXiv:2506.09943 [cs].

[26] John M. Franchak, Dina J. van der Zalm, and Karen E. Adolph. 2010. Learning by doing: Action performance facilitates affordance perception. *Vision Research* 50, 24 (Dec. 2010), 2758–2765. doi:10.1016/j.visres.2010.09.019

[27] Yingqiang Ge, Wenyue Hua, Kai Mei, Jianchao Ji, Juntao Tan, Shuyuan Xu, Zelong Li, and Yongfeng Zhang. 2023. OpenAGI: When LLM Meets Domain Experts. *Advances in Neural Information Processing Systems* 36 (Dec. 2023), 5539–5568. https://proceedings.neurips.cc/paper_files/paper/2023/hash/1190733f217404edc8a7f4e15a57f301-Abstract-Datasets_and_Benchmarks.html

[28] Brianna L. Grant, Paul C. Yielder, Tracey A. Patrick, Bill Kapralos, Michael Williams-Bell, and Bernadette A. Murphy. 2019. Audiohaptic Feedback Enhances Motor Performance in a Low-Fidelity Simulated Drilling Task. *Brain Sciences* 10, 1 (Dec. 2019), 21. doi:10.3390/brainsci10010021

[29] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, and others. 2022. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 18995–19012.

[30] Dov Greenbaum. 2016. Ethical, legal and social concerns relating to exoskeletons. *SIGCAS Comput. Soc.* 45, 3 (Jan. 2016), 234–239. doi:10.1145/2874239.2874272

[31] Xiaochi Gu, Yifei Zhang, Weize Sun, Yuanzhe Bian, Dao Zhou, and Per Ola Kristensson. 2016. Dexmo: An Inexpensive and Lightweight Mechanical Exoskeleton for Motion Capture and Force Feedback in VR. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. Association for Computing Machinery, New York, NY, USA, 1991–1995. doi:10.1145/2858036.2858487

[32] Abhishek Gupta, Marcia K. O'Malley, Volkan Patoglu, and Charles Burgar. 2008. Design, Control and Performance of RiceWrist: A Force Feedback Wrist Exoskeleton for Rehabilitation and Training. *The International Journal of Robotics Research* 27, 2 (Feb. 2008), 233–251. doi:10.1177/0278364907084261

[33] Huy Ha, Pete Florence, and Shuran Song. 2023. Scaling Up and Distilling Down: Language-Guided Robot Skill Acquisition. In *Proceedings of The 7th Conference on Robot Learning*. PMLR, 3766–3777. https://proceedings.mlr.press/v229/ha23a.html

[34] Thilo Hagendorff. 2020. The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines* 30, 1 (March 2020), 99–120. doi:10.1007/s11023-020-09517-8

[35] Guangxing Han and Ser-Nam Lim. 2024. Few-Shot Object Detection with Foundation Models. (June 2024), 28608–28618. doi:10.1109/CVPR52733.2024.02703

[36] Mahmoud Hassan, Florian Daiber, Frederik Wiehr, Felix Kosmalla, and Antonio Krüger. 2017. FootStriker: An EMS-based Foot Strike Assistant for Running. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 1 (March 2017), 2:1–2:18. doi:10.1145/3053332

[37] Thomas T. Hewett and Sari Scott. 1987. The Use of Thinking-out-loud and Protocol Analysis in Development of a Process Model of Interactive Database Searching. In *Human–Computer Interaction–INTERACT '87*, H. J. Bullinger and B. Shackel (Eds.). North-Holland, Amsterdam, 51–56. doi:10.1016/B978-0-444-70304-0.50018-2

[38] Kashmir Hill. 2024. Two Students Created Face Recognition Glasses. It Wasn't Hard. *The New York Times* (Oct. 2024). https://www.nytimes.com/2024/10/24/technology/facial-recognition-glasses-privacy-harvard.html

[39] Ronan Hinchet, Velko Vechev, Herbert Shea, and Otmar Hilliges. 2018. DextrES: Wearable Haptic Feedback for Grasping in VR via a Thin Form-Factor Electrostatic Brake. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology (UIST '18)*. Association for Computing Machinery, New York, NY, USA, 901–912. doi:10.1145/3242587.3242657

[40] Tiancheng Hu, Yara Kyrychenko, Steve Rathje, Nigel Collier, Sander van der Linden, and Jon Roozenbeek. 2025. Generative language models exhibit social identity biases. *Nature Computational Science* 5, 1 (Jan. 2025), 65–75. doi:10.1038/s43588-024-00741-1

[41] Yongquan 'Owen' Hu, Jingyu Tang, Xinya Gong, Zhongyi Zhou, Shuning Zhang, Don Samitha Elvitigala, Florian 'Floyd' Mueller, Wen Hu, and Aaron J. Quigley. 2025. Vision-Based Multimodal Interfaces: A Survey and Taxonomy for Enhanced Context-Aware System Design. doi:10.1145/3706598.3714161 arXiv:2501.13443 [cs].

[42] Mina Huh, Zihui Xue, Ujjaini Das, Kumar Ashutosh, Kristen Grauman, and Amy Pavel. 2025. Vid2Coach: Transforming How-To Videos into Task Assistants. In *The 38th Annual ACM Symposium on User Interface Software and Technology (UIST '25)*. Association for Computing Machinery, New York, NY, USA. doi:10.1145/3746059.3747612

[43] Seokhyun Hwang, Seongjun Kang, Jeongseok Oh, Jeongju Park, Semoo Shin, Yiyue Luo, Joseph DelPreto, Sangbeom Lee, Kyoobin Lee, Wojciech Matusik, Daniela Rus, and SeungJun Kim. 2025. TelePulse: Enhancing the Teleoperation Experience through Biomechanical Simulation-Based Electrical Muscle Stimulation in Virtual Reality. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI'25)*. Association for Computing Machinery, Yokohama, Japan. doi:10.1145/3706598.3713767

[44] Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023. Towards Mitigating LLM Hallucination via Self Reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 1827–1843. doi:10.18653/v1/2023.findings-emnlp.123

[45] Addie Johnson. 2003. Procedural Memory and Skill Acquisition. In *Handbook of Psychology*. John Wiley & Sons, Ltd, 499–523. doi:10.1002/0471264385.wei0418 Section: 18 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/0471264385.wei0418.

[46] Shunichi Kasahara, Jun Nishida, and Pedro Lopes. 2019. Preemptive Action: Accelerating Human Reaction Using Electrical Muscle Stimulation Without Compromising Agency. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, 643:1–643:15. doi:10.1145/3290605.3300873

[47] Oliver Beren Kaul, Max Pfeiffer, and Michael Rohs. 2016. Follow the Force: Steering the Index Finger towards Targets using EMS. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '16)*. Association for Computing Machinery, New York, NY, USA, 2526–2532. doi:10.1145/2851581.2892352

[48] Emre Kazim and Adriano Soares Koshiyama. 2021. A high-level overview of AI ethics. *Patterns* 2, 9 (Sept. 2021). doi:10.1016/j.patter.2021.100314

[49] Jinwook Kim, Seonghyeon Kim, and Jeongmi Lee. 2022. The Effect of Multisensory Pseudo-Haptic Feedback on Perception of Virtual Weight. *IEEE Access* 10 (2022), 5129–5140. doi:10.1109/ACCESS.2022.3140438 Conference Name: IEEE Access.

[50] Jarrod Knibbe, Paul Strohmeier, Sebastian Boring, and Kasper Hornbæk. 2017. Automatic Calibration of High Density Electric Muscle Stimulation. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3 (Sept. 2017), 68:1–68:17. doi:10.1145/3130933

[51] Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large Language Models are Zero-Shot Reasoners. *Advances in Neural Information Processing Systems* 35 (Dec. 2022), 22199–22213. https://proceedings.neurips.cc/paper_files/paper/2022/hash/8bb0d291acd4acf06ef112099c16f326-Abstract-Conference.html

[52] Michinari Kono, Yoshio Ishiguro, Takashi Miyaki, and Jun Rekimoto. 2018. Design and Study of a Multi-Channel Electrical Muscle Stimulation Toolkit for Human Augmentation. In *Proceedings of the 9th Augmented Human International Conference (AH '18)*. Association for Computing Machinery, New York, NY, USA, 1–8. doi:10.1145/3174910.3174913

[53] Ernst Kruijff, Dieter Schmalstieg, and Steffi Beckhaus. 2006. Using neuromuscular electrical stimulation for pseudo-haptic feedback. In *Proceedings of the ACM symposium on Virtual reality software and technology*. ACM, Limassol Cyprus, 316–319. doi:10.1145/1180495.1180558

[54] Bolin Lai, Xiaoliang Dai, Lawrence Chen, Guan Pang, James M. Rehg, and Miao Liu. 2025. LEGO: Learning EGOcentric Action Frame Generation via Visual Instruction Tuning. In *Computer Vision – ECCV 2024*, Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (Eds.). Vol. 15067. Springer Nature Switzerland, Cham, 135–155. doi:10.1007/978-3-031-72673-6_8 Series Title: Lecture Notes in Computer Science.

[55] Jaewook Lee, Jun Wang, Elizabeth Brown, Liam Chu, Sebastian S. Rodriguez, and Jon E. Froehlich. 2024. GazePointAR: A Context-Aware Multimodal Voice Assistant for Pronoun Disambiguation in Wearable Augmented Reality. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–20. doi:10.1145/3613904.3642230

[56] Kyungyeon Lee, Daniel S Yang, Kriti Singh, and Jun Nishida. 2025. Hapticus: Exploring the Effects of Haptic Feedback and its Customization on Motor Skill Learning: Tactile, Haptic, and Somatosensory Approaches. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, 1–20. doi:10.1145/3706598.3713821

[57] V. I. Levenshtein. 1966. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady* 10 (Feb. 1966), 707. https://ui.adsabs.harvard.edu/abs/1966SPhD...10..707L ADS Bibcode: 1966SPhD...10..707L.

[58] Pawel Lewicki, Thomas Hill, and Elizabeth Bizot. 1988. Acquisition of procedural knowledge about a pattern of stimuli that cannot be articulated. *Cognitive Psychology* 20, 1 (Jan. 1988), 24–37. doi:10.1016/0010-0285(88)90023-0

[59] Michael Xieyang Liu, Frederick Liu, Alexander J. Fiannaca, Terry Koo, Lucas Dixon, Michael Terry, and Carrie J. Cai. 2024. "We Need Structured Output": Towards User-centered Constraints on Large Language Model Output. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, 1–9. doi:10.1145/3613905.3650756

[60] Pedro Lopes. 2025. What if the "I" in HCI stands for Integration?. In *Adjunct Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology (UIST Adjunct '25)*. Association for Computing Machinery, New York, NY, USA, Article 16, 5 pages. doi:10.1145/3746058.3762829

[61] Pedro Lopes and Patrick Baudisch. 2017. Immense Power in a Tiny Package: Wearables Based on Electrical Muscle Stimulation. *IEEE Pervasive Computing* 16, 3 (2017), 12–16. doi:10.1109/MPRV.2017.2940953 Conference Name: IEEE Pervasive Computing.

[62] Pedro Lopes, Alexandra Ion, and Patrick Baudisch. 2015. Impacto: Simulating Physical Impact by Combining Tactile Stimulation with Electrical Muscle Stimulation. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology (UIST '15)*. Association for Computing Machinery, New York, NY, USA, 11–19. doi:10.1145/2807442.2807443

[63] Pedro Lopes, Patrik Jonell, and Patrick Baudisch. 2015. Affordance++: Allowing Objects to Communicate Dynamic Use. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. Association for Computing Machinery, New York, NY, USA, 2515–2524. doi:10.1145/2702123.2702128

[64] Pedro Lopes, Sijing You, Lung-Pan Cheng, Sebastian Marwecki, and Patrick Baudisch. 2017. Providing Haptics to Walls & Heavy Objects in Virtual Reality by Means of Electrical Muscle Stimulation. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. Association for Computing Machinery, New York, NY, USA, 1471–1482. doi:10.1145/3025453.3025600

[65] Laura Marchal-Crespo, Mark van Raai, Georg Rauter, Peter Wolf, and Robert Riener. 2013. The effect of haptic guidance and visual feedback on learning a complex tennis task. *Experimental Brain Research* 231, 3 (Nov. 2013), 277–291. doi:10.1007/s00221-013-3690-2

[66] Ariana Martino, Michael Iannelli, and Coleen Truong. 2023. Knowledge Injection to Counter Large Language Model (LLM) Hallucination. In *The Semantic Web: ESWC 2023 Satellite Events*, Catia Pesquita, Hala Skaf-Molli, Vasilis Efthymiou, Sabrina Kirrane, Axel Ngonga, Diego Collarana, Renato Cerqueira, Mehwish Alam, Cassia Trojahn, and Sven Hertling (Eds.). Springer Nature Switzerland, Cham, 182–185. doi:10.1007/978-3-031-43458-7_34

[67] William P. McCarthy, Sean P. Anderson, and Judith E. Fan. 2024. How does assembling an object affect memory for it? *Proceedings of the Annual Meeting of the Cognitive Science Society* 46, 0 (2024). https://escholarship.org/uc/item/71h690dm

[68] Frederic Meyer, Lennart Freitag, Sven Hinrichsen, and Oliver Niggemann. 2024. Potentials of Large Language Models for Generating Assembly Instructions. In *2024 IEEE 29th International Conference on Emerging Technologies and Factory Automation (ETFA)*. 1–8. doi:10.1109/ETFA61755.2024.10710806

[69] James Minogue and M. Gail Jones. 2006. Haptics in Education: Exploring an Untapped Sensory Modality. *Review of Educational Research* 76, 3 (Sept. 2006), 317–348. doi:10.3102/00346543076003317

[70] Dan Morris, Hong Tan, Federico Barbagli, Timothy Chang, and Kenneth Salisbury. 2007. Haptic Feedback Enhances Force Skill Learning. In *Second Joint EuroHaptics Conference and Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems (WHC'07)*. 21–26. doi:10.1109/WHC.2007.65

[71] Kazuki Nagai, Soma Tanoue, Katsuhito Akahane, and Makoto Sato. 2015. Wearable 6-DoF wrist haptic device "SPIDAR-W". In *SIGGRAPH Asia 2015 Haptic Media And Contents Design (SA '15)*. Association for Computing Machinery, New York, NY, USA, 1–2. doi:10.1145/2818384.2818403

[72] Ken Nakagaki, Sean Follmer, and Hiroshi Ishii. 2015. LineFORM: Actuated Curve Interfaces for Display, Interaction, and Constraint. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*. ACM, Charlotte NC USA, 333–339. doi:10.1145/2807442.2807452

[73] Arinobu Niijima, Yukio Koike, and Hironori Ishikawa. 2024. Motor-Skill-Download System Using Electrical Muscle Stimulation for Enhancing Piano Playing. In *Companion of the 2024 on ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '24)*. Association for Computing Machinery, New York, NY, USA, 313–317. doi:10.1145/3675094.3681942

[74] Jun Nishida, Yudai Tanaka, Romain Nith, and Pedro Lopes. 2022. DigituSync: A Dual-User Passive Exoskeleton Glove That Adaptively Shares Hand Gestures. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. ACM, Bend OR USA, 1–12. doi:10.1145/3526113.3545630

[75] Romain Nith, Yun Ho, and Pedro Lopes. 2024. SplitBody: Reducing Mental Workload while Multitasking via Muscle Stimulation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–11. doi:10.1145/3613904.3642629

[76] Romain Nith, Shan-Yuan Teng, Pengyu Li, Yujie Tao, and Pedro Lopes. 2021. DextrEMS: Increasing Dexterity in Electrical Muscle Stimulation by Combining it with Brakes. In *The 34th Annual ACM Symposium on User Interface Software and Technology (UIST '21)*. Association for Computing Machinery, New York, NY, USA, 414–430. doi:10.1145/3472749.3474759

[77] Kuniaki Noda, Hiroaki Arie, Yuki Suga, and Tetsuya Ogata. 2014. Multimodal integration learning of robot behavior using deep neural networks. *Robotics and Autonomous Systems* 62, 6 (June 2014), 721–736. doi:10.1016/j.robot.2014.03.003

[78] Gary M. Olson, Susan A. Duffy, and Robert L. Mack. 1984. Thinking-Out-Loud as a Method for Studying Real-Time Comprehension Processes. In *New Methods in Reading Comprehension Research*. Routledge. Num Pages: 34.

[79] Y. Omura. 1985. Electrical parameters for safe and effective electro-acupuncture and transcutaneous electrical stimulation: threshold potentials for tingling, muscle contraction and pain; and how to prevent adverse effects of electrotherapy. Part 1. *Acupuncture & Electro-Therapeutics Research* 10, 4 (1985), 335–337. doi:10.3727/036012985816714405

[80] Yousi A. Oquendo, Margaret M. Coad, Sherry M. Wren, Thomas S. Lendvay, Ilana Nisky, Anthony M. Jarc, Allison M. Okamura, and Zonghe Chua. 2024. Haptic Guidance and Haptic Error Amplification in a Virtual Surgical Robotic Training Environment. *IEEE Transactions on Haptics* 17, 3 (July 2024), 417–428. doi:10.1109/TOH.2024.3350128

[81] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* 35 (Dec. 2022), 27730–27744. https://proceedings.neurips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html

[82] Rock Yuren Pang, Hope Schroeder, Kynnedy Simone Smith, Solon Barocas, Ziang Xiao, Emily Tseng, and Danielle Bragg. 2025. Understanding the LLM-ification of CHI: Unpacking the Impact of LLMs at CHI through a Systematic Literature Review. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, 1–20. doi:10.1145/3706598.3713726

[83] Marietta Papadatou-Pastou, Eleni Ntolka, Judith Schmitz, Maryanne Martin, Marcus R. Munafò, Sebastian Ocklenburg, and Silvia Paracchini. 2020. Human handedness: A meta-analysis. *Psychological Bulletin* 146, 6 (2020), 481–524. doi:10.1037/bul0000229 Place: US.

[84] Max Pfeiffer, Tim Dünte, Stefan Schneegass, Florian Alt, and Michael Rohs. 2015. Cruise Control for Pedestrians: Controlling Walking Direction using Electrical Muscle Stimulation. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. Association for Computing Machinery, New York, NY, USA, 2505–2514. doi:10.1145/2702123.2702190

[85] Max Pfeiffer and Michael Rohs. 2017. Haptic Feedback for Wearables and Textiles Based on Electrical Muscle Stimulation. In *Smart Textiles: Fundamentals, Design, and Interaction*, Stefan Schneegass and Oliver Amft (Eds.). Springer International Publishing, Cham, 103–137. doi:10.1007/978-3-319-50124-6_6

[86] Physiopedia. 2025. Physiopedia. https://www.physio-pedia.com/home/ Last accessed April 9, 2025.

[87] Physiotutors. 2025. Wrist/Hand Active Range of Motion (AROM) Assessment. https://www.physiotutors.com/wiki/wrist-hand-active-range-of-motion/. Last accessed April 9, 2025.

[88] David Pinzon, Roberto Vega, Yerly Paola Sanchez, and Bin Zheng. 2017. Skill learning from kinesthetic feedback. *The American Journal of Surgery* 214, 4 (Oct. 2017), 721–725. doi:10.1016/j.amjsurg.2016.10.018

[89] Wanli Qian, Chenfeng Gao, Anup Sathya, Ryo Suzuki, and Ken Nakagaki. 2024. SHAPE-IT: Exploring Text-to-Shape-Display for Generative Shape-Changing Behaviors with LLMs. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology (UIST '24)*. Association for Computing Machinery, New York, NY, USA, 1–29. doi:10.1145/3654777.3676348

[90] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*. PMLR, 8748–8763. https://proceedings.mlr.press/v139/radford21a.html

[91] Gang Ren, Zhihuang Huang, Shijiang You, Wenshuo Lin, Tianyang Huang, Gang Wang, and Jee Hang Lee. 2025. Enhancing Motor Skills and Coordination With Visual-Haptic Feedback in Ball Sport Training. *IEEE Access* 13 (2025), 49214–49231. doi:10.1109/ACCESS.2025.3547159

[92] Stefan Schneegass, Albrecht Schmidt, and Max Pfeiffer. 2016. Creating user interfaces with electrical muscle stimulation. *interactions* 24, 1 (Dec. 2016), 74–77. doi:10.1145/3019606

[93] Vivian Shen and Chris Harrison. 2025. Kinethreads: Soft Full-Body Haptic Exosuit using Low-Cost Motor-Pulley Mechanisms. In *The 38th Annual ACM Symposium on User Interface Software and Technology (UIST '25)*. Association for Computing Machinery, New York, NY, USA. doi:10.1145/3746059.3747755

[94] Roland Sigrist, Georg Rauter, Robert Riener, and Peter Wolf. 2013. Augmented visual, auditory, haptic, and multimodal feedback in motor learning: A review. *Psychonomic Bulletin & Review* 20, 1 (Feb. 2013), 21–53. doi:10.3758/s13423-012-0333-8

[95] Paul Solomon. 1995. The think aloud method: A practical guide to modelling cognitive processes. *Information Processing & Management* 31, 6 (Nov. 1995), 906–907. doi:10.1016/0306-4573(95)90031-4

[96] P. Strojnik, A. Kralj, and I. Ursic. 1979. Programmed Six-Channel Electrical Stimulator for Complex Stimulation of Leg Muscles During Walking. *IEEE Transactions on Biomedical Engineering* BME-26, 2 (Feb. 1979), 112–116. doi:10.1109/TBME.1979.326520

[97] Akifumi Takahashi, Jas Brooks, Hiroyuki Kajimoto, and Pedro Lopes. 2021. Increasing Electrical Muscle Stimulation's Dexterity by means of Back of the Hand Actuation. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3411764.3445761

[98] Akifumi Takahashi, Yudai Tanaka, Archit Tamhane, Alan Shen, Shan-Yuan Teng, and Pedro Lopes. 2024. Can a Smartwatch Move Your Fingers? Compact and Practical Electrical Muscle Stimulation in a Smartwatch. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology (UIST '24)*. Association for Computing Machinery, New York, NY, USA, 1–15. doi:10.1145/3654777.3676373

[99] Nobuhiro Takahashi, Hayato Takahashi, and Hideki Koike. 2019. A Novel Soft Exoskeleton Glove for Motor Skill Acquisition Similar to Anatomical Structure of Forearm Muscles. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. 1568–1569. doi:10.1109/VR.2019.8797919

[100] Emi Tamaki, Takashi Miyaki, and Jun Rekimoto. 2011. PossessedHand: techniques for controlling human hands using electrical muscles stimuli. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI '11)*. ACM, 543–552. doi:10.1145/1978942.1979018

[101] Xingyu Tan, Xiaoyang Wang, Qing Liu, Xiwei Xu, Xin Yuan, and Wenjie Zhang. 2025. Paths-over-Graph: Knowledge Graph Empowered Large Language Model Reasoning. In *Proceedings of the ACM on Web Conference 2025 (WWW '25)*. Association for Computing Machinery, New York, NY, USA, 3505–3522. doi:10.1145/3696410.3714892

[102] Yudai Tanaka, Jun Nishida, and Pedro Lopes. 2022. Electrical Head Actuation: Enabling Interactive Systems to Directly Manipulate Head Orientation. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–15. doi:10.1145/3491102.3501910

[103] Sho Tatsuno, Tomohiko Hayakawa, and Masatoshi Ishikawa. 2017. Supportive training system for sports skill acquisition based on electrical stimulation. In *2017 IEEE World Haptics Conference (WHC)*. 466–471. doi:10.1109/WHC.2017.7989946

[104] Madhuri A. Tayal, Minal Deshmukh, Vijaya Pangave, Manjushri Joshi, Sulakshana Malwade, and Shraddha Ovale. 2023. VMLHST: Development of an Efficient Novel Virtual Reality ML Framework with Haptic Feedbacks for Improving Sports Training Scenarios. *International Journal of Electrical and Electronics Research* 11, 2 (June 2023), 601–608. doi:10.37391/IJEER.110249

[105] Timon ten Berge and Rene van Hezewijk. 1999. Procedural and Declarative Knowledge: An Evolutionary Perspective. *Theory & Psychology* 9, 5 (Oct. 1999), 605–624. doi:10.1177/0959354399095002

[106] Teslasuit. 2025. Teslasuit as a Self-standing FES Device. https://teslasuit.io/use-cases/teslasuit-as-a-self-standing-fes-device/. Last accessed April 8, 2025.

[107] TeslaSuit. 2025. TeslaSuit Developer API. https://developer.teslasuit.io/documentation/tag/apps-and-api/. Last accessed April 9, 2025. API requires a TeslaSuit license; public-facing versions are available at https://github.com/

teslasuit.

[108] Haider Usman, Yu Zhou, Benjamin Metcalfe, and Dingguo Zhang. 2020. A Functional Electrical Stimulation System of High-Density Electrodes With Auto-Calibration for Optimal Selectivity. *IEEE Sensors Journal* 20, 15 (Aug. 2020), 8833–8843. doi:10.1109/JSEN.2020.2983004

[109] Athanasios Voulodimos, Nikolaos Doulamis, Anastasios Doulamis, and Eftychios Protopapadakis. 2018. Deep Learning for Computer Vision: A Brief Review. *Computational Intelligence and Neuroscience* 2018, 1 (2018), 7068349. doi:10.1155/2018/7068349 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1155/2018/7068349.

[110] Jiaqi Wang, Enze Shi, Huawen Hu, Chong Ma, Yiheng Liu, Xuhui Wang, Yincheng Yao, Xuan Liu, Bao Ge, and Shu Zhang. 2025. Large language models for robotics: Opportunities, challenges, and perspectives. *Journal of Automation and Intelligence* 4, 1 (March 2025), 52–64. doi:10.1016/j.jai.2024.12.003

[111] Jinbao Wang, Shujie Tan, Xiantong Zhen, Shuo Xu, Feng Zheng, Zhenyu He, and Ling Shao. 2021. Deep 3D human pose estimation: A review. *Computer Vision and Image Understanding* 210 (Sept. 2021), 103225. doi:10.1016/j.cviu.2021.103225

[112] Kyosuke Watanabe, Makoto Oka, and Hirohiko Mori. 2019. Feedback Control to Target Joints Angle in Middle Finger PIP and MP Joint Using Functional Electrical Stimulation. In *Human Interface and the Management of Information. Information in Intelligent Systems*, Sakae Yamamoto and Hirohiko Mori (Eds.). Springer International Publishing, Cham, 440–454. doi:10.1007/978-3-030-22649-7_35

[113] Daniel B. Willingham, Mary J. Nissen, and Peter Bullemer. 1989. On the development of procedural knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 15, 6 (Nov. 1989), 1047–1060. doi:10.1037/0278-7393.15.6.1047

[114] Tongshuang Wu, Michael Terry, and Carrie Jun Cai. 2022. AI Chains: Transparent and Controllable Human-AI Interaction by Chaining Large Language Model Prompts. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–22. doi:10.1145/3491102.3517582

[115] Yuqing Yang, Lei Jiao, and Yuedong Xu. 2024. A Queueing Theoretic Perspective on Low-Latency LLM Inference with Variable Token Length. In *2024 22nd International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)*. 273–280. https://ieeexplore.ieee.org/document/10778367/

[116] Siyu Yuan, Jiangjie Chen, Ziquan Fu, Xuyang Ge, Soham Shah, Charles Jankowski, Yanghua Xiao, and Deqing Yang. 2023. Distilling Script Knowledge from Large Language Models for Constrained Language Planning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 4303–4325. doi:10.18653/v1/2023.acl-long.236

[117] Yuhang Zang, Wei Li, Jun Han, Kaiyang Zhou, and Chen Change Loy. 2025. Contextual Object Detection with Multimodal Large Language Models. *International Journal of Computer Vision* 133, 2 (Feb. 2025), 825–843. doi:10.1007/s11263-024-02214-4

[118] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. 2024. Vision-Language Models for Vision Tasks: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46, 8 (Aug. 2024), 5625–5644. doi:10.1109/TPAMI.2024.3369699

[119] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. VinVL: Revisiting Visual Representations in Vision-Language Models. 5579–5588. https://openaccess.thecvf.com/content/CVPR2021/html/Zhang_VinVL_Revisiting_Visual_Representations_in_Vision-Language_Models_CVPR_2021_paper.html

[120] Yutong Zhang, Lixing Chen, Shenghong Li, Nan Cao, Yang Shi, Jiaxin Ding, Zhe Qu, Pan Zhou, and Yang Bai. 2025. Way to Specialist: Closing Loop Between Specialized LLM and Evolving Domain Knowledge Graph. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1*. ACM, Toronto ON Canada, 1996–2007. doi:10.1145/3690624.3709187

[121] Peihan Zhong and Richard T. Stone. 2013. Automated kinesthetic trainer enhances kinesthetic memory development. *International Journal of Industrial Ergonomics* 43, 3 (May 2013), 238–245. doi:10.1016/j.ergon.2013.02.007

[122] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Learning to Prompt for Vision-Language Models. *International Journal of Computer Vision* 130, 9 (Sept. 2022), 2337–2348. doi:10.1007/s11263-022-01653-1

[123] M. Zhou, S. Tse, A. Derevianko, D. B. Jones, S. D. Schwaitzberg, and C. G. L. Cao. 2012. Effect of haptic feedback in laparoscopic surgery skill acquisition. *Surgical Endoscopy* 26, 4 (April 2012), 1128–1134. doi:10.1007/s00464-011-2011-8

[124] Xinrui Zou. 2019. A Review of Object Detection Techniques. In *2019 International Conference on Smart Grid and Electrical Automation (ICSGEA)*. 251–254. doi:10.1109/ICSGEA.2019.00065

[125] Ümit Önen, Fatih M. Botsalı, Mete Kalyoncu, Mustafa Tınkır, Nihat Yılmaz, and Yusuf Şahin. 2014. Design and Actuator Selection of a Lower Extremity Exoskeleton. *IEEE/ASME Transactions on Mechatronics* 19, 2 (April 2014), 623–632. doi:10.1109/TMECH.2013.2250295 Conference Name: IEEE/ASME Transactions on Mechatronics.

A Appendix

A.1 System design comparison

In this section, we detail situations that exemplify the value of specific aspects of our architecture (e.g., the value of using body-pose information to generate movement instructions). Each of these examples comprises two side-by-side comparisons, visually highlighting the difference in the output caused by a particular aspect of our architecture. While these are not essential in judging the accuracy of our architecture—this was reported in our *Ablation Study*—these provide visual examples to assist the reader.

Location information. Figure 26 depicts side-by-side examples where a sensor alone (e.g., here, GPS location) enabled an improved outcome (e.g., correctly understanding a turn-tilt window common in *Europe* vs. assuming a more-standard casement window).

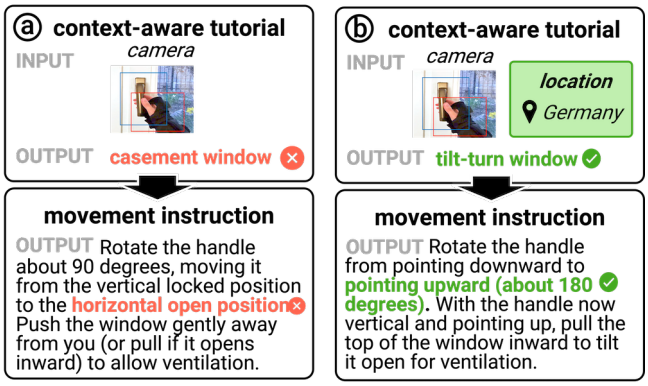

Figure 26: Generated instructions with vs. without location.

Visual context. Figure 27 illustrates how multimodal-AI can reason in visual scenes. In (a), we cropped the image of the spray-can and hand, while in (b), we provided the entire POV. The outcome differs, since without visual-context (e.g., spray and cooking pan) an incorrect "shake" (as in *Affordance++* [63]) was generated, though cooking-sprays do not need shaking.

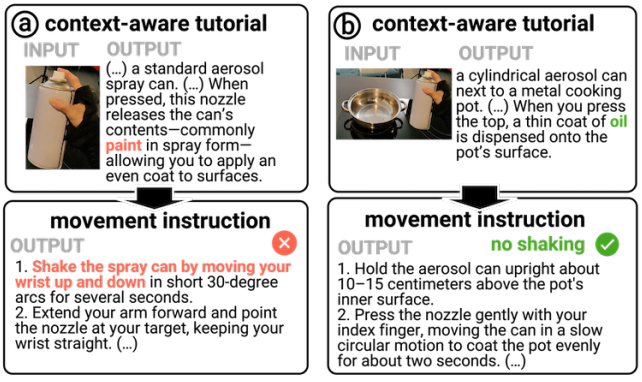

Figure 27: Generated instructions with vs. without POV.

Spoken request. Figure 28 illustrates how user's spoken requests can provide context. In (a), with the POV of a bike handlebar and a vague "EMS help me" request, the system generated instructions to brake. Conversely, in (b), the addition of "help me *disengage from the bike cleat*" provided clues to generate adequate instructions to unclip the shoe from the pedal.

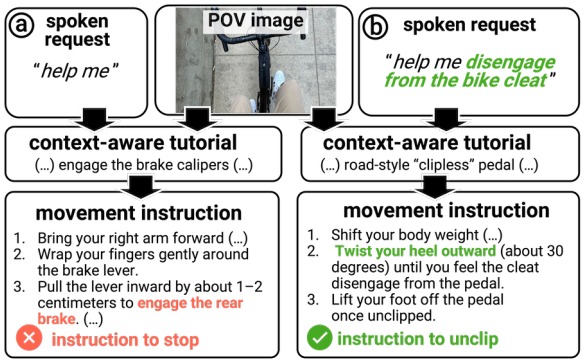

Figure 28: Generated instructions with vs. without details on the user's request.

Pose-information. Figure 29 illustrates side-by-side examples that exemplify the contribution of pose-information. In (a) without pose-information, the generated instructions do not direct the user's hands towards the handle; conversely, (b), with pose, the generated instructions highlight how the forearm needs to be rotated to reach the handle, given the current pose.

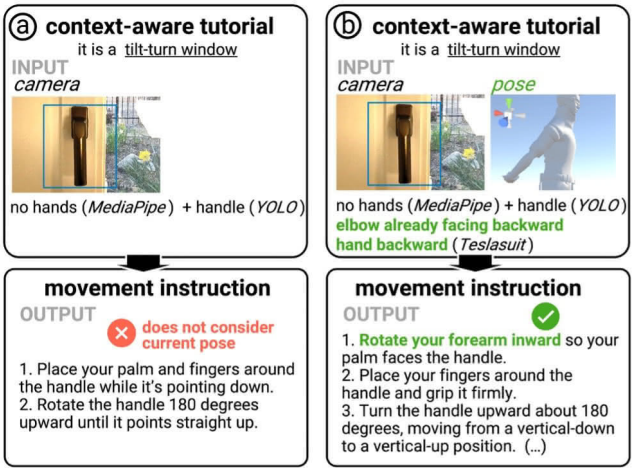

Figure 29: Generated instructions with vs. without body pose.

Additionally, readers can refer to our open-source repository (https://embodied-ai.tech) for the source-code and a database of outputs of our models.