International Journal of Human–Computer Interaction

ISSN: 1044-7318 (Print) 1532-7590 (Online) Journal homepage: www.tandfonline.com/journals/hihc20

Grounding Pedagogical Intents in Embodied AI Teachers: A Human-Centered Framework for Designing and Evaluating Instructional Gestures

Lai Wei, Sark Pangrui Xing, Kenny K. N. Chow & Stephen Jia Wang

To cite this article: Lai Wei, Sark Pangrui Xing, Kenny K. N. Chow & Stephen Jia Wang (14 May 2026): Grounding Pedagogical Intents in Embodied AI Teachers: A Human-Centered Framework for Designing and Evaluating Instructional Gestures, International Journal of Human–Computer Interaction, DOI: 10.1080/10447318.2026.2664083

To link to this article: https://doi.org/10.1080/10447318.2026.2664083

INTERNATIONAL JOURNAL OF HUMAN–COMPUTER INTERACTION
https://doi.org/10.1080/10447318.2026.2664083

Grounding Pedagogical Intents in Embodied AI Teachers: A Human-Centered Framework for Designing and Evaluating Instructional Gestures

Lai Wei[a*], Sark Pangrui Xing[b*], Kenny K. N. Chow[c] and Stephen Jia Wang[b]

[a]Brain, Language, and Computation Lab, Department of Language Science and Technology, The Hong Kong Polytechnic University, Hong Kong, China; [b]School of Design, The Hong Kong Polytechnic University, Hong Kong, China; [c]School of Communication, Hong Kong Baptist University, Hong Kong, China

ABSTRACT

Generative AI expands opportunities for embodied agents in HCI, yet a gap persists between human-centered AI principles and practical design methods, particularly for pedagogical agents' (PAs) co-speech gestures. Automated text-to-gesture systems lack the instructional nuance needed for effective teaching. To address this, we used a Research-through-Design approach to develop a human-centered framework that translates pedagogical intent into gesture specifications for embodied AI teachers. The framework includes four iterative stages: Preparation, which analyzes gesture patterns and instructional functions; Human PA Acting, where educators and designers rehearse gestures through performance; Embodied PA Acting, which transfers human motion to agents using video-based pose estimation; and EPA-assisted Course Delivery, which evaluates student experiences through interviews. Findings indicate that these gestures enhanced students' perception of the PA's professionalism, approachability, and instructional rhythm, while boosting overall engagement. This work contributes a design framework and insights for pedagogical gesture design, and exploratory guidance for generative-AI prompting.

KEYWORDS

Human-Centered AI; design methodology; pedagogical agent; gesture; generative AI

1. Introduction

Generative AI (GenAI) has become a novel tool in Human-Computer Interaction (HCI) and design research, transforming the ways in which we conduct HCI research and design practices (Zhou et al., 2024). Notably, scholars have employed GenAI to generate Embodied Agents (EAs) with co-speech gestures, and it is progressively becoming capable of performing conversational tasks (Nyatsanga et al., 2023; Wolfert et al., 2022). Recently, joint efforts have advanced to leverage GenAI for developing gesture synthesis systems that generate semantically relevant gestures aligned with speech rhythm and content for Embodied Conversational Agents (ECAs) (Zhang et al., 2024; Zhi et al., 2023). Meanwhile, in the landscape of digital learning, the application of Embodied Pedagogical Agents (EPAs)[1] has emerged as a transformative field for enhancing educational engagement and learning performance (Kizilkaya & Askar, 2008; Li et al., 2024; Savin-Baden et al., 2015). The development and design of these agents encompass several dimensions, including EPAs' appearance (Beege et al., 2022; Shiban et al., 2015), acoustic attributes (Ceha & Law, 2022), interaction (Serras Pereira et al., 2018), personality (Castillo et al., 2018; Hahn et al., 2018; Thomas et al., 2022), and voices (Bonfert et al., 2021; Lee et al., 2020; Mildner et al., 2024; Reicherts et al., 2022). Additionally, particular attention has been given to nonverbal cues (Baylor & Kim, 2009; Lin et al., 2020), which have been extensively studied across

CONTACT Stephen Jia Wang stephen.j.wang@polyu.edu.hk School of Design, The Hong Kong Polytechnic University, Hong Kong, China
*These authors contributed equally to this work.

various disciplines, including education, psychology, and computing (Kersey et al., 2024; Lawson et al., 2021; Li et al., 2022; Yoon et al., 2021).

EPAs in digital learning environments have been further explored in areas such as educational virtual reality (VR) (Petersen et al., 2021; Tsai et al., 2019) and intelligent EPAs (Dai et al., 2024). Among these applications, anthropomorphic characteristics, such as co-speech gestures, have become a primary focus (Dai et al., 2022; Davis, 2018; Mayer & DaPra, 2012; Wang & Ruiz, 2021). These gestural cues direct learner attention, clarify concepts, support retention, and guide transitions between knowledge points—making their pedagogically grounded design central to effective learning outcomes (Davis et al., 2021; Li et al., 2019; Moon & Ryu, 2021). Regarding existing motion synthesis for generating PA, there are two approaches. One reconstructs motion through pose estimation and retargeting, inferring full-body movement from sparse data and refining articulation through kinematic modeling, as seen in DeepMotion[2]—this approach is time-intensive and closely dependent on the motion-tracked performer's behavior (Peng et al., 2018; Xu et al., 2020). The other generates motion through video synthesis, predicting coherent spatiotemporal sequences for fluid, context-aware movement, exemplified by large vision models such as KlingAI and Sora (Liu et al., 2024). This paradigm operates through prompt-driven generation (text or image to video), producing outputs that vary considerably across runs, with limited user control over gesture semantics, timing, or pedagogical alignment (Zhang et al., 2023; Zhu et al., 2024). Leveraging these technologies, commercial platforms such as HeyGen[3], Synthesia[4], and iFlyTek[5] offer AI-generated instructors featuring anthropomorphic characteristics—including facial expressions, lip synchronization, and natural body movements—through text-to-speech (TTS) and automated animation.

While a substantial body of literature affirms that gestures performed by EPAs positively enhance student learning outcomes (Beege et al., 2020; Li et al., 2024; McNeill, 2011; Pi et al., 2022), recent findings suggest that these benefits are often inconsistent. This variability indicates that the pedagogical value of gestures is highly contingent upon specific gesture categories (Li et al., 2019), their interaction with other modalities (e.g., agent appearance (Davis, 2018)), and the context of use (Mayer & DaPra, 2012). Such complexity underscores an urgent need for foundational design frameworks that remain largely unestablished. Although recent advancements have integrated computer vision with biomechanical skeletal modeling to achieve high-fidelity movement reconstruction (Shaw et al., 2024), significantly enhancing the alignment between visual pose estimations and skeletal rigging, this technical precision has yet to be guided by the pedagogical design standards necessary for effective instructional behavior.

Specifically, while existing research has made significant technical advances in gesture synthesis (pose estimation (Shaw et al., 2024), semantic retrieval (Zhang et al., 2024)), and documented gesture effectiveness in education (Li et al., 2022; Schneider et al., 2022), a critical gap remains between technical precision and pedagogical authenticity: no systematic framework translates pedagogical intent into implementable gesture specifications grounded in authentic teaching practices. Current research has yet to address whether the proportional distribution of gesture types—as observed in naturalistic teaching—is essential for EPA realism and efficacy, or how specific gesture designs influence students' perceptions of instructional quality. Without such ecological behavioral benchmarks, Pedagogical Agent (PA) gesture design remains predominantly intuitive and fragmented. Consequently, current generative technologies have not resolved the fundamental challenge of creating PAs that are both behaviorally realistic and pedagogically effective. We identify three critical, interrelated problems:

1. *Lack of a Design Framework:* There is no structured, replicable framework to guide the design of instructional gestures. This absence impedes collaboration between educators and designers, making it difficult to systematically translate specific *instructional needs* into meaningful, co-speech gestures that are aligned with pedagogical goals.
2. *The Connectivity Gap Between Human Behavioral Data and PA Design:* A significant gap persists in translating empirical research on human instructional gestures into PA design. Specifically, few studies correlate authentic teacher behaviors (e.g., cohesive or beat gestures) with measurable student perceptions in the learning process. Without understanding how a PA's gestures influence key

HCI metrics, such as learner perceptions of credibility and professionalism, evidence-based design choices remain difficult. Consequently, the lack of behavioral benchmarks leaves even advanced generative models disconnected from the nuanced pedagogical functions of human movement.

3. *Unexplored Potential and Pitfalls of Generative AI:* Current text-to-video generation tools for instructional content lack systematic evaluation. It remains unclear how to structure prompts to capture pedagogical intent and temporal coherence. A critical assessment of generation workflows is necessary to establish effective design strategies for pedagogically grounded gesture creation.

Research has demonstrated that instructional gestures for embodied agents can foster stronger social connections, which in turn enhance student learning performance (Li et al., 2022; Schneider et al., 2022). By enriching the social agency of virtual instructors, these gestural behaviors directly cater to the burgeoning trend toward AI-driven personalized and immersive learning environments (Zhang, Li, et al., 2025; Zhang, Liu, et al., 2025). Given the increasing demand for such high-fidelity interaction, this study seeks to establish formalized design frameworks to support the systematic development of gestures in AI teachers. As such, we investigate the collaborative instructional gesture design process between PA designers and educators, explore how generative technologies can be leveraged within this workflow, and examine student perceptions of the resulting multimodal instructional information.

This study is grounded in Research-through-Design (RtD) (Zimmerman et al., 2007), where an educator and a PA designer collaboratively explore the design of embodied PAs with instructional meaning, using state-of-the-art technologies to fine-tune PAs as desired. Initially, we observed thirteen design-related lectures to analyze the frequency of different gesture types used by educators. This was followed by the formation of a team consisting of a PA designer and educator, who worked together through collaborative acting and rehearsal. Throughout the gesture design process, the team aligned their objectives to ensure the gestures met educators' instructional needs and improved the clarity of the PA's expressions. Using the RtD framework, the team went through four stages of the design process: Preparation, Human PA Acting, Embodied PA Acting, and EPA-Assisted Course Delivery. During each stage, both team members engaged in iterative practices, delivering the prepared presentation, making gestures intuitively, and emulating real-life teaching scenarios. The PA designer observed the educator's gestures, and both reflected on how to better express co-speech gestures to convey instructional meaning. Building on insights from previous research on word-synchronized gestures (Ali et al., 2020; Kucherenko et al., 2020; Liu et al., 2022) and the alignment of gestures with parts of speech (PoS) (Wei & Chow, 2023; Yoon et al., 2021)[6], we explored the use of DeepMotion, Sora, and KlingAI to evaluate the quality of animated instructional gestures for PAs. The result of this design and acting process was a 14-minute course featuring instructional gesture-based PAs. To evaluate learning efficacy, we tested the course with 38 validated Hong Kong-based university students.

This paper makes the following contributions to the field of HCAI methodologies:

1. *A Human-Centered Framework for Gesture Design:* We detail a replicable *human-centered design framework* for creating pedagogically effective gestures. Unlike prior work that focuses on technical naturalness or general gesture generation, our framework operationalizes the translation of educators' tacit pedagogical knowledge into implementable specifications through structured corpus analysis, collaborative rehearsal, and student-centered validation. This four-stage process is grounded in the analysis of authentic teaching and uses reflective, collaborative practices between educators and designers to translate pedagogical intent into agent embodiment, serving as a practical benchmark for advancing teaching practices via multimodal instructional design and educational technology development.
2. *An Empirical Evaluation of the PA and Its Impact:* We conducted a formal *empirical evaluation* of a pedagogical agent built using our framework. The results validate our approach: the agent's use of cohesive and beat gestures led to high student ratings of its professionalism, credibility, and approachability, with students describing enriched engagement and positive learning experiences.

3. *Insights into Generative AI and Future Directions:* Through exploratory testing and quantitative evaluation, we demonstrate that structured prompts with temporal phases and linguistic anchors substantially improve instructional coherence in text-to-video generation. This perspective supports a "human-in-control" workflow and reveals practical considerations for developing future instruction-focused AI systems.

2. Background and related work

2.1. Human-centered AI and the methodological gap

The notion of Human-Centered AI (HCAI) advocates for a paradigm shift, emphasizing AI systems that amplify and augment human capabilities rather than merely replace them (Xu, 2019; Zhou et al., 2024). Key principles of HCAI include ensuring human control, transparency, fairness, and promoting well-being (Shneiderman, 2020). These principles are further substantiated by evidence showing that outcomes of decision-making augmented by algorithmic systems are strongly influenced by how humans perceive, trust, and engage with these systems (Burton et al., 2020; Ozmen Garibay et al., 2023). However, many organizations and researchers have noted the difficulty in translating these high-level principles into concrete design actions (Xu & Gao, 2023). The core challenge lies in developing methodologies that are both structured enough to be repeatable and flexible enough to handle the complexity of human contexts.

2.2. Generative AI for HCI and design

GenAI refers to a class of artificial intelligence models designed to create new content, such as images, text, and music, by learning patterns from existing data. One of its most prominent applications is AI image synthesis, where techniques such as Generative Adversarial Networks (GANs) and diffusion models play crucial roles. GANs are particularly known for generating high-fidelity images through a competitive learning process between a generator and a discriminator (Brock, 2018). However, with the advancement of diffusion models and large language models (LLMs), modern Text-to-Image Models, such as DALL·E, Stable Diffusion, and MidJourney, have outperformed GANs especially in image diversity, enabling a broader audience to create high-quality, stylized images from textual descriptions (Dhariwal & Nichol, 2021).

GenAI has become an integral tool in HCI and design research, shaping design workflows (Petridis et al., 2024; Tholander & Jonsson, 2023), decision-making (Park et al., 2024; Zhou et al., 2024), and user experience studies (Uusitalo et al., 2024). It is widely applied in design research and practice across various fields, both within and beyond the HCI community (Weisz et al., 2024). While widely applied in ideation (Hollmén Larsen & Zhu, 2024) and artifact generation (Kun et al., 2024; Zhang et al., 2024), traditional text-to-x models often struggle to capture the fluid and multifaceted nature of design intent. Consequently, multimodal AI systems have emerged as a significant trend, offering integrated frameworks that align diverse data streams—such as visual, textual, and spatial inputs—to support more nuanced explorations (Liu et al., 2025; Wu et al., 2023; Zhu et al., 2023). By fuzing these modalities, these systems provide designers with higher degrees of creative agency and more intuitive ways to translate complex, non-verbal intentions into tangible artifacts (Manesh et al., 2024; Peng et al., 2024; Tilekbay et al., 2024).

Despite these advances, GenAI poses significant challenges. One major issue is *control*, designers often struggle to steer AI outputs due to the opacity of AI decision-making, leading to results that require iterative refinement as stated in Gao et al. (2025). Additionally, while *multi-modal* inputs such as audio and images are being explored to facilitate more intuitive designer-AI interactions, text-based prompts remain the dominant input method, potentially limiting creativity and flexibility.

2.3. Co-speech gesture synthesis

Co-speech gesture synthesis refers to the automatic generation of realistic and contextually appropriate hand and body gestures that accompany spoken language in EAs. Recently, GenAI has been widely

adopted to generate co-speech gestures for these agents. There are two long lasting methodologies to create such gestures: hybrid rule-based techniques and data-driven methods (Chen et al., 2025; Nyatsanga et al., 2023; Wolfert et al., 2022; Yoon et al., 2022). The earliest study (Cassell, 1998) on the hybrid rule-based approach employed probabilistic models to generate gestures based on linguistic and contextual cues. In contrast, the data-driven approach utilizes deep learning to train agents' gestures using Mocap or video data from sources like dyadic conversations and TED presentations (Lee et al., 2019; Wolfert et al., 2022; Yoon et al., 2019; 2020). Both methods have demonstrated effectiveness in producing naturalistic bodily movements. To develop AI-generated EAs that enhance authenticity and immersion, integrating co-speech gestures with semantic meaning is essential. However, the aforementioned approaches are inherently labor-intensive and face scalability challenges due to technological constraints. Moreover, research on generative methods for *gesture synthesis*, especially on defining gesture rules for AI-generated EAs remains limited.

In recent years, a few co-speech gesture generation models with semantic retrieval have emerged, combining rule-based and deep learning-based systems (Sadoughi & Busso, 2019; Zhang et al., 2024; Zhi et al., 2023). Evaluations of these semantics-aware gesture synthesis systems demonstrate that these systems outperform purely deep-learning approaches in generating semantically meaningful and rhythmical gestures, thereby enhancing the human-like expressiveness of AI-generated EAs. For instance, a semantic-aware co-speech gesture synthesis system leverages a GPT-based motion generator and a LLM-driven retrieval framework to ensure high-quality, semantically relevant gestures that align with speech rhythm and content (Pang et al., 2025; Zhang et al., 2024). Inserting retrieved semantic gesture identifiers into rhythmic gestures enhances the perceived expressiveness of AI-generated ECAs in social interactions. Despite the advantages of using Mocap datasets for precise full-body movement tracking, inconsistencies in semantic annotations across datasets pose a challenge (Zhang et al., 2024; Zhi et al., 2023). Additionally, while these systems employ semantic motion generation via new text prompts to replicate several classic semantic gestures (Cheng et al., 2024; Zhang et al., 2024; Zhi et al., 2023), such as "pointing left," "disagreement," and "a large amount," human-produced semantic gestures are often more nuanced and hierarchical. Consequently, the current prompt-based segmentation of gestures is relatively simplified compared to McNeill's Gesture Phases and functional gesture classifications (McNeill, 2011). However, this simplification may introduce inconsistencies, such as incomplete or unclear gestures, or conflicts in the temporal sequencing of semantic and rhythmic gestures.

Moreover, most current datasets are primarily sourced from conversational gestures (Ginosar et al., 2019; Habibie et al., 2021; Liu et al., 2022; Yi et al., 2023). In contrast, generating co-speech instructional gestures necessitates an understanding of educators' instructional intentions to effectively convey educational content and account for students' interpretations of these gestures. For example, educators may deliberately adjust the rhythm, amplitude, and direction of their co-speech gestures to emphasize key concepts, while students may perceive these gestures differently in terms of their communicative intent. This highlights the need for further research on educators' gesture usage and students' perceptions of instructional gestures performed by AI-generated PAs. Past evaluations of user perception have often relied on scales measuring dimensions such as naturalness and human-likeness (Wolfert et al., 2022; Yang et al., 2023; Zhang et al., 2024; Zhi et al., 2023). However, these metrics may not fully capture how AI-generated PAs are perceived in real-world educational settings. Factors such as professionalism, approachability, and credibility are also crucial in educational contexts. Bridging this gap requires not only technical advances in synthesis but an understanding of educators' instructional intentions and learners' perceptual needs—dimensions that gesture synthesis research, operating largely within computational paradigms, has yet to address.

Though research has investigated data-driven co-speech gesture generation that emphasizes speech rhythm and lexical embedding (Ali et al., 2020; Kucherenko et al., 2020; Liu et al., 2022; Yoon et al., 2019), there remains substantial scope for examining the synchronization of coherent gestures, such as cohesive and beat gestures, with words. Word-synchronized gestures represent minimal units in the alignment of speech and gesture and hold promise for enhancing the design of gesture transitions and connections. A pioneering study conducted by Wei and Chow (Wei & Chow, 2023) has begun to show these alignments. They have discovered that cohesive gestures are associated with specific PoS, such as

adverbs and logical connectors, whereas beat gestures are associated with pronouns, adjectives, and action-oriented verbs. Cohesive gestures appear to strengthen the logical coherence and comprehension of speech, whereas beat gestures provide rhythmic emphasis and visual cues. The study identifies four patterns of alignment between educators' PoS, lexis, and gestures (Wei & Chow, 2023), providing insights into the design of instructional gestures. It also calls for a closer analysis of how human teachers use these gestures in delivering educational knowledge. Building on this, this study explores the design of PAs with instructionally coherent gestures, particularly cohesive and beat gestures, and evaluates their impact on the learning experience. Furthermore, it informs the intentional integration of word-synchronized gestures in PA design and extracts insights from evaluation findings to optimize semantic motion generation via text prompts.

2.4. Pedagogical agents' gesture design

Educators who employ gestures can enhance both social engagement and learning outcomes among students (Nakagawa et al., 2021; Schneider et al., 2022; Sinatra et al., 2021). These gestures transform cognitive understanding into multimodal information that students can easily process (Goldin-Meadow, 2015). McNeill categorized gestures into five distinct types—*cohesives*, *beats*, *deictics*, *iconics*, and *metaphorics*—each serving a distinct communicative role (McNeill, 2011). While iconic and metaphoric gestures are often tied to specific visual content, this study focuses on cohesive and beat gestures due to their frequent appearance in discourse and their fundamental role in reinforcing coherence and prosody. Cohesive gestures are movements that connect different parts of a narrative or discourse, creating logical links between concepts. They help to maintain continuity and structure the flow of information, for example, by using a hand to sweep from one idea to the next to show a transition; Beat gestures are small, rhythmic motions, such as a flick of the hand or a tap of the fingers, that are synchronized with the prosody of speech. Expanding on McNeill's framework, research on PA gesture design has aimed to mirror human gestural behavior in educational contexts. Studies suggest that integrating beat, deictic, iconic, and metaphoric gestures enhances attention, retention, and comprehension, with notable advantages for second-language learners (Beege et al., 2020; Craig et al., 2015; Davis & Vincent, 2019; Pi et al., 2022). Yet, while students intuitively comprehend iconic and metaphoric gestures that depict semantic information, this is not the case for discipline-specific terminology (Wei & Chow, 2023). The way gestures distribute within speech is also tightly linked to linguistic content; when speakers describe visual imagery, they are more likely to use iconic gestures (Masson-Carro et al., 2017). Additionally, an individual's fluency in spoken language influences gesture production (Chui, 2005). The effectiveness of iconic and metaphoric gestures, therefore, hinges on subject-specific terminology, the extent of visual descriptions, and language fluency.

By contrast, cohesive and beat gestures frequently appear in discourse, reinforcing coherence and prosody. They do not carry semantic meaning on their own but function to emphasize or add prominence to specific words or phrases, much like verbal stress. Their fundamental role in communication makes them a promising avenue for further research on gesture design. Cohesive gestures help maintain continuity across conversational turns and organize spatial references (Belhiah, 2013; McNeill & Levy, 1993; Sekine & Kita, 2015), making them a natural tool for structuring discourse. Investigations into instructional gestures in PAs, including spatially communicative gestures (Wu et al., 2021) and rhythmic hand motions (Pi et al., 2022), shed light on how fluid transitions between movements contribute to a more natural delivery of speech. Beyond their inherent communicative functions, gestures are also shaped by the instructional setting. Researchers argue that factors such as classroom layout, seating arrangement, and material placement influence both the frequency and type of gestures used by teachers (Wei & Chow, 2022). Additionally, variations in gesture frequency contribute to different learning outcomes (Davis et al., 2021). Enhanced gesture frequency significantly improved cued recall and recognition, while average frequency does not, highlighting the role of social cue strength in learning effectiveness. Thus, PA gesture design should consider gesture types, teachers's instructional intentions, frequency, and contextual factors. This study provides insight into the natural integration of instructional gesture transitions in PAs and their influence on student perceptions. Yet, much remains to be explored regarding gestures that encode semantic information, particularly cohesive and beat

gestures. Translating these findings into actionable design specifications for AI-generated agents, however, requires a methodological bridge that neither gesture synthesis research nor pedagogical literature has established on its own.

2.5. Instructional materials design

Educational material design encompasses the systematic planning, development, and evaluation of instructional resources to achieve specific learning objectives (Dick et al., 2005). Seminal frameworks provided by scholars such as Merrill (2002), Gagné (1985), and Reigeluth (1999) emphasize learner-centered approaches, advocating for the deliberate alignment of content, guided by the educator's instructional intent, with cognitive processes and the specific learning context. In the digital era, the integration of technology has received significant attention, with Mayer's Principles of Multimedia Learning (Mayer, 2005; Mayer, 2002) providing a theoretical foundation for utilizing digital media to foster dynamic engagement and support diverse learning styles.

A critical dimension of contemporary instructional design is the duration of digital content. While early empirical studies on massive open online courses (MOOCs) observed a decline in engagement for videos exceeding six minutes (Davis, 2018; Guo et al., 2014), more recent pedagogical literature suggests that this "six-minute rule" may not apply to complex, university-level curricula. Research indicates that durations ranging from 12 to 20 min are effective when supported by instructional scaffolding and high-quality production (Lagerstrom et al., 2015; Manasrah et al., 2021). From a social agency perspective, extended exposure is often necessary to establish a stable perception of an agent's persona and to allow for the cumulative effect of continuous gestural cues (Castro-Alonso et al., 2021; Schroeder et al., 2025). Furthermore, the efficacy of such materials is influenced by the pacing of information delivery. While speaking rates in educational contexts vary widely, typically ranging from 48 to 254 words per minute (wpm), an optimum of approximately 160 wpm is often recommended for clarity (Williams, 1998). A moderated pace (e.g., 140–150 wpm) is frequently employed in complex technical subjects to manage cognitive load without compromising engagement (Pastore, 2012).

Beyond structural parameters, the design of instructional materials integrates cultural responsiveness and accessibility to address diverse learning needs (Gay, 2002). This focus on inclusivity aligns with the movement toward Open Educational Resources (OER), which aims to increase collaboration and knowledge democratization (Wiley & Hilton, 2009). Within this context, Human-Computer Interaction (HCI) serves as a particularly effective instructional subject for a broad student population; its interdisciplinary nature not only aligns with foundational curricula (Hewett et al., 1992) but also provides a versatile platform for exploring the intersection of design, technology, and human behavior (Wilcox et al., 2019). By applying Mayer's principles to HCI-oriented materials, instructional designers can create visual and auditory elements tailored to the demographic and cognitive profiles of modern learners. This strategic alignment ensures that content is relevant and engaging, as HCI's inherent user-centricity mirrors the pedagogical goal of creating accessible, culturally resonant resources. Such a dual-layered approach reflects the educator's intent to ensure that instructional material is not only cognitively accessible but also professionally resonant for students across various disciplines.

As AI becomes increasingly integrated into instructional delivery through automated agents, synthetic voices, and generated gestures (Burton et al., 2020; Dai et al., 2024; Li et al., 2024), how educators' instructional intent survives translation into automated systems emerges as a central design challenge. Whether students and educators trust such systems depends in part on whether this intent remains coherent and legible in the agent's behavior (Birmingham et al., 2020; Chow, 2026; Johnson & Lester, 2016; Mehrotra et al., 2024). Lee and See (2004) address this directly, identifying performance (demonstrated capabilities), process (how the system operates), and purpose (the goal-oriented intent embedded by designers) as the three bases of trustworthy automation. PA research has made considerable progress on performance and process, including evaluating gesture naturalness, improving synthesis techniques, and refining agent appearance, yet pedagogical intent, as the purpose dimension of PA design, has received comparatively little attention. How an educator's instructional goals are encoded and preserved across content, pacing, and embodied behavior remains underexplored, and constitutes the design challenge this work addresses.

2.6. Design processes and reflective practices

In the field of HCI, the design process is widely recognized as a sequence of decision-making and reflection steps that foster iterative development. Prior work addresses the potential for embedding new knowledge within design artifacts, illustrating the role of design processes and prototypes in advancing research agendas—a concept encapsulated in Research-through-Design (RtD) (Zimmerman et al., 2007) and articulated through *Strong Concepts* that solidify abstract design principles (Höök & Löwgren, 2012). Schön's notable concept of *reflective practice* characterizes this iterative design as a continuous cycle of "Reflection-in-action," in which practitioners dynamically evaluate and adapt their methods in-situ (Schon, 1992). Designers, through reflective engagement in their activities such as sketching, prototyping, or creating representational artifacts - gain insights that inform and refine their design thinking in real time. This approach resonates with Ingold (2013)'s idea of *thinking through making*, where design thinking emerges organically from iterative making practices, progressing from broad concepts to intricate details (Wiberg, 2014). Moreover, Brown et al. (Seely Brown et al., 1989) introduces the notion of *indexing*, which emphasizes that learning is inherently contextual and illustrates how deep immersion in design activity strengthens experiential learning, analogous to language acquisition through contextual immersion.

Reflective practices extend into instructional activities, such as the gestures acted by PA instructors and the design of educational materials. Although PA gesture design might seem peripheral to traditional design discourse, multimedia learning research has shown that effective knowledge transmission requires an intentional alignment of verbal and visual elements (Mayer, 2002). Consequently, PA gesture design could benefit from design research methodologies, opening up a less-explored space of design in educational contexts. Additionally, recent studies suggest that iterative, experimental approaches involving collaborative partnerships can help navigate risks and enhance innovation outcomes (Magistretti et al., 2021). Agile design thinking methodologies emphasize prototyping and human-centered perspectives to tackle such complex educational design challenges (Carlgren et al., 2016; Cooper & Sommer, 2016; Micheli et al., 2019). By incorporating a blend of design research methodologies—conceptual framing, experimentation, and user-focused evaluation—the development of multimedia instructional materials and educational practices can achieve improved effectiveness. In the following section, we will explore how design research methodologies can guide the iterative design of educational materials and PA gestures, underscoring an exploratory design approach tailored to educational contexts.

2.7. Theoretical framing

This research synthesizes three bodies of work: Human-Centered AI methodologies that operationalize high-level principles into concrete design practices (Shneiderman, 2020; Xu & Gao, 2023), co-speech gesture synthesis that aligns gestural behavior with linguistic structure (McNeill, 2011; Wei & Chow, 2023; Zhang et al., 2024), and pedagogical research on multimodal instruction (Goldin-Meadow, 2015; Mayer, 2002). Each domain, while productive independently, addresses only part of the challenge: HCAI principles lack concrete design processes for artifacts like gestural PAs; gesture synthesis prioritizes technical fluency over instructional grounding; and pedagogical research documents gesture effectiveness without translating findings into implementable agent design specifications. By adopting HCAI as our theoretical lens, we reframe PA gesture design as a collaborative process (Schon, 1992; Zimmerman et al., 2007) that preserves educators' agency while leveraging computational tools. This perspective positions our four-stage methodology as a systematic approach to capture tacit pedagogical knowledge and translate it into implementable specifications validated through student perceptions (Nakagawa et al., 2021; Schneider et al., 2022). Situated at the intersection of these three bodies of work, the framework treats pedagogical intent as the *purpose* dimension of trustworthy PA design (Lee & See, 2004), positioning it as a traceable design input alongside the technical and evaluative processes that carry it through to the final agent.

3. Design methodology

Our methodology was developed through a Research-through-Design process. It is structured to ensure that the design of the PA is continuously grounded in the instructional needs of the educators and the

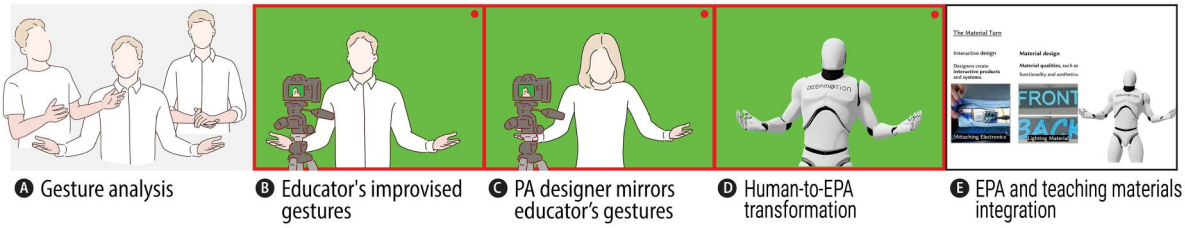


Figure 1. Key steps in our human-centered design methodology: (A) *gesture analysis*: Understanding teachers' intents by observing a senior teacher's gestures. (B) *Educator's improvised gestures*: Capturing tacit instructional knowledge through performance. (C) *PA designer mirrors educator*: Facilitating interdisciplinary collaboration and refinement. (D) *Human-to-EPA transformation*: Upper body motions extracted via DeepMotion for 3D animation. (E) *EPA and teaching materials integration*: 3D animation embedded alongside teaching materials.

perceptual needs of the learners. The process involves a close collaboration between an educator and a PA designer, spanning five key steps from gesture analysis to EPA integration (Figure 1; panels A–E are labeled in the figure).

3.1. Collaborative team and course preparation

The teaching material preparation was a collaborative effort between the PA designer and the educator, each contributing their domain expertise. *PA designer*, with over ten years of experience in animation production and design education, was responsible for gesture design and animation. The *educator*, trained as a designer and with over five years of teaching experience in design education, delivered the keynote using pre-scripted presentation notes. To ensure alignment with the desired presentation quality, the educator and PA designer engaged in iterative discussions to articulate a shared vision of the instructional delivery. These sessions led to the formulation of preliminary design criteria for performance-based teaching, which informed both the structure of a 14-minute lecture and the design of word-synchronized co-speech instructional gestures.

As part of the effort to eliminate the influence of prior knowledge earned by the participants in the field of HCI, the team performed an exclusive literature inquiry strategy on ACM's digital library. This search was conducted using the following formula [Title: "guide"] OR [Title: "tutorial"] AND [E-Publication Date: Past year], which returned 92 results. Following the removal of duplicates and the application of the "full text research article" filter, 37 entries were selected for the screening repository. the two collaborators independently screened the titles and abstracts of these entries, subsequently reviewing the full texts to include only those papers deemed fit for college-level comprehension, ensuring the inclusivity of the material for all participants regardless of the participants' study backgrounds. Although these efforts yielded four suitable entries (i.e. (Freeman & Curtis, 2023; Micheli et al., 2019; Freeman & Curtis, 2023; Ibrahim et al., 2023; Shatilov et al., 2023; Xing et al., 2023)) deemed appropriate for teaching, one particular case arose to our attention (Xing et al., 2023) recently published on ACM International Conference on Tangible, Embedded, and Embodied Interaction (TEI). It utilizes a pictorial format that centralizes visuals and diagrams to describe a novel material-centered design process (i.e., *analysis, synthesis, and detailing*) of crafting interactive materiality in a step-by-step manner. The two collaborators agreed to use (Xing et al., 2023) as the final teaching material due to its flow of articulation and comprehensibility for college students (Figure 2).

3.2. Developing the instructional materials

The overall structure of the prerecorded video closely mirrors the structure of the article. It begins with an introduction to the literature concerning material-centered and interactive materiality (for further details, please refer to the original paper (Xing et al., 2023)), and then provides a brief overview of the design artifact known as Puffy. Then, it outlines the analysis-synthesis-detailing (A-S-D) design and implementation process in a total of thirteen methodical steps, which are reflected in the presentation slides. The *educator* also prepared presentation notes for each slide to allow a smooth teaching

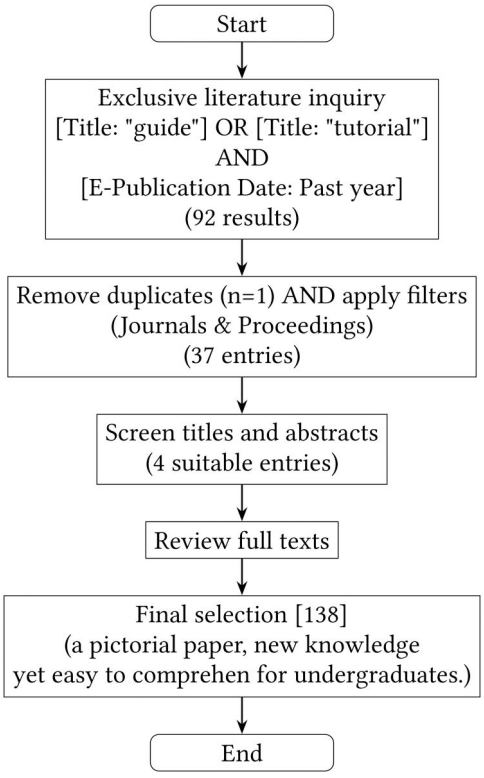


Figure 2. Flowchart for literature selection process.

experience. While developing these materials, a gap was observed as there was no clear guidance in the existing literature to support the design implementation of the PAs at an operational level. Yet, the two collaborators recognized that Puffy's design process might serve as a valuable reference. It exemplifies an exploratory procedure that encompasses 1) analyzing various shape and material explorations, 2) synthesizing suitable samples, and 3) detailing processes. the two collaborators were inspired by that and started to unfold the PA design in the following section.

3.3. Implementing the methodology: From human action to embodied agent

The core of our methodology consists of four iterative stages for implementing the PA's gestures: (1) Preparation, (2) Human PA Acting, (3) EPA Acting, and (4) EPA-assisted course delivery and evaluation (Figure 3). Each stage builds upon the last, ensuring that human requirements gathered early on are carried through to the final AI artifact.

3.3.1. Stage 1: Preparation (multimodal corpus analysis)

The initial iteration aimed to investigate the natural gestures employed in an educational context (Figure 4).

3.3.2. Preparation

The initial iteration aimed to investigate the natural gestures employed in an educational context, specifically within an undergraduate design course. As motivated by the word-synchronized gesture study (Wei & Chow, 2023), this phase was initiated by *a) observing and analyzing* (Figure 4(A)). Eleven teachers (8 male, 3 female), each with more than five years of teaching experience, were recruited via email. They taught design- and art-related courses, with video data collected from 13 one-hour sessions (7 offline, 4 online, and 2 hybrid). Their courses featured detailed case studies, increasing the likelihood of using metaphoric and iconic gestures (Cienki & Müller, 2008; McNeill, 2011). During the lectures, full-body movements and presentation slides were recorded using an iPad Pro. Researchers analyzed both gestures and verbal language from the video recordings. Gestures were manually annotated using

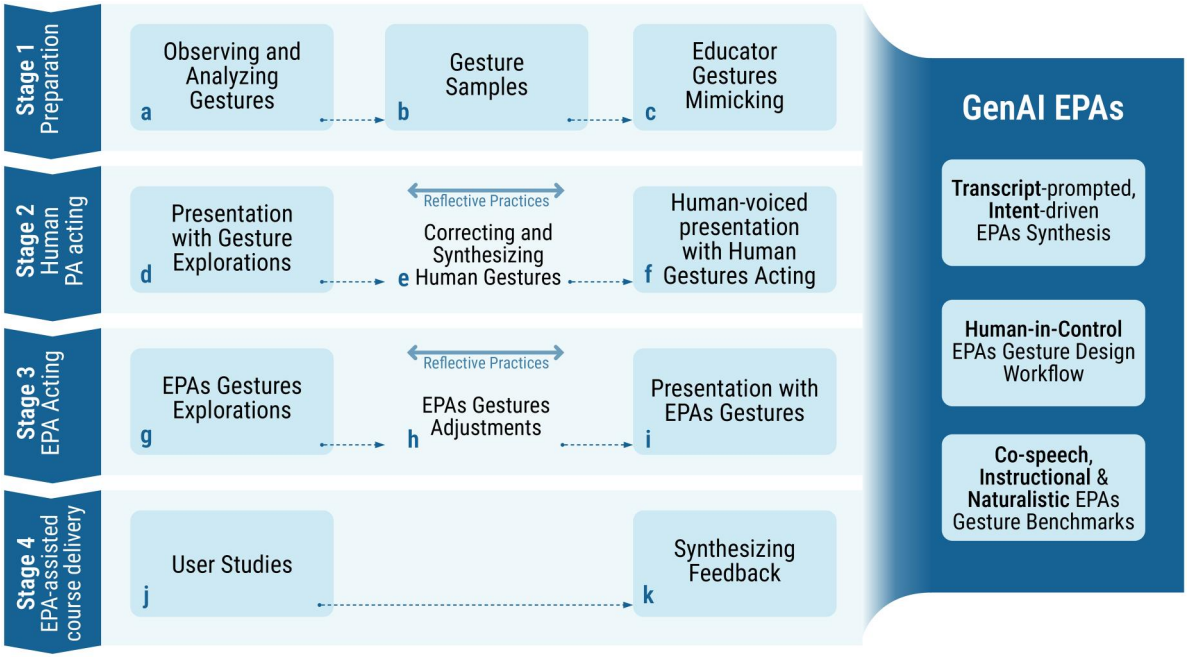
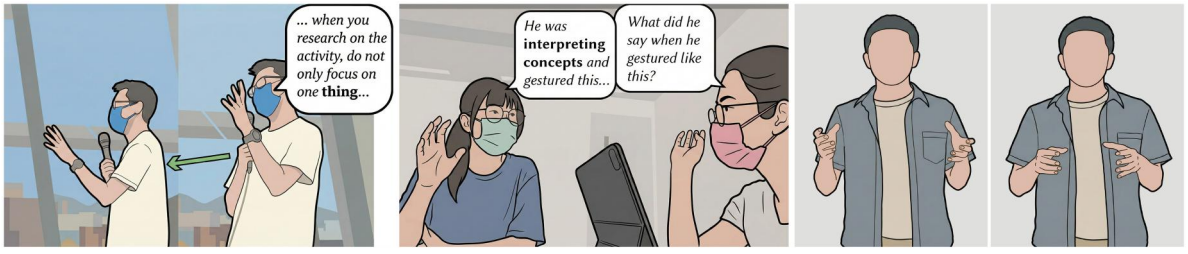


Figure 3. Visual overview of the entire pedagogical agent (PA) gestures design procedures.



(A) Observing and analyzing gestures (B) Gesture samples (C) *Educator* gestures mimicking

Figure 4. Stage 1: Preparation.

ELAN (ELAN, 2026), following McNeill's five-category gesture taxonomy (cohesive, beat, deictic, iconic, metaphoric) (McNeill, 2011). Two researchers underwent structured training prior to coding: both studied McNeill's classification criteria and jointly coded a calibration video, during which definitional boundaries for each category were discussed and formalized into a shared coding protocol. Subsequent annotations were conducted through a collaborative consensus process: where a gesture's classification was disputed, both coders discussed the case against the established criteria until agreement was reached. This negotiated agreement procedure is recognized as a valid reliability mechanism in qualitative content analysis, as it ensures that category boundaries are systematically applied and that ambiguous cases are resolved through principled deliberation rather than arbitrary assignment (Johnny Saldaña, 2021). Speech was transcribed through OtterAI[7] and subsequently reviewed for accuracy. To investigate word-synchronized gestures, a chronological coding approach was applied, grouping one to three consecutive words into microcosmic units based on the lecture timeline. Using Natural Language Processing (NLP) with Python (Bird et al., 2009), we systematically coded 71,252 samples of gestures aligned with their respective words and PoS. Results showed that cohesive gestures accounted for 44.34% (33,031 occurrences) of the total, followed by beat gestures at 38.26% (34,532 occurrences). Deictic (8.64%), metaphoric (2.98%), and iconic gestures (2.53%) were less frequently used. The distribution across PoS revealed that nouns were the most common (36.59%), followed by verbs (23.8%), pronouns (16.63%), adverbs (14.91%), adjectives (8.32%), and conjunctions (7.24%). Cohesive gestures are essential for establishing referential connections and enhancing the clarity of lectures. Following our Pearson's Chi-squared correlation analysis, we discovered that the use of nouns, adverbs, particles, and

conjunctions in gestures enhances communication coherence, effectively reinforcing the delivery of key educational concepts (Figure 4(A)). The researchers concluded that among these, cohesives and beats constituted the most frequently employed yet unobtrusive categories compared to dectics, iconics, and metaphorics. As highlighted in the related work section, cohesive and beat gestures can augment the coherence of PAs' bodily movements, thereby enhancing students' learning experiences. These types of gestures were selected to further develop PA with naturalistic instructional gestures.

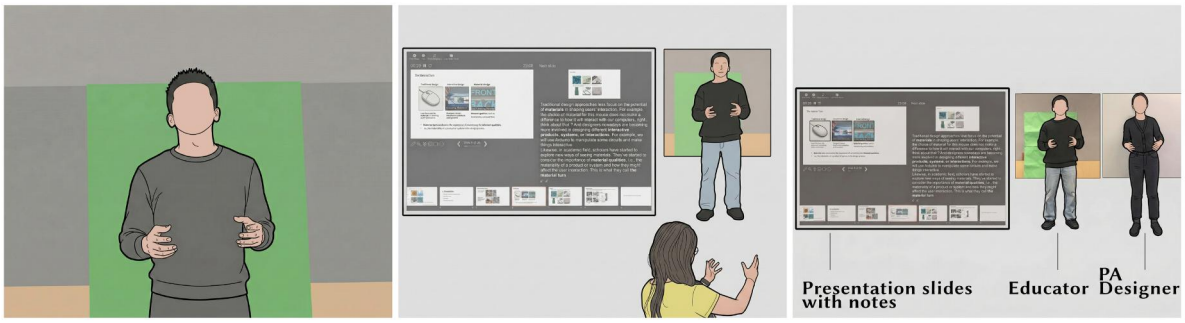
Subsequent substantial effort was devoted to *b) gesture samples*. The *PA designer* randomly selected one student from seven out of the eleven lectures to recruit a paired group of seven participants (Male = 3; Female = 4; aged 20–22 years). English is a second language for these students, who aimed to interpret selected cohesive and beat gestures extracted from the lecture videos (Figure 4(B)). When students observed the instructors' cohesive and beat gestures without any accompanying audio cues, 12 out of 13 were able to easily infer the contextual meanings. For example, when instructors repeatedly performed upward or downward hand movements, students recognized these gestures as an intentional emphasis on specific concepts—often accompanied by nouns or adverbs indicating a strong attitude toward the definition. Similarly, when instructors linked gestures in sequence, students identified this as a transition between concepts, with the likely presence of conjunctions or similar connecting terms. As informed by these insights, the next phase involved *c) educator gesture mimicking* (Figure 4(C)). The *educator* initially executed a trial featuring a sequence of both intentional and non-intentional gestures, grounded in the prepared presentation notes. The *educator* and the *PA designer* then engaged in a detailed discussion to reach a consensus on criteria for word-gesture alignment, drawing from the trial's gesture sequences. Throughout the process, the *educator* reflected on key moments and logical connections in the course content where key concepts emerged, while also identifying how gestures could reinforce these concepts and facilitate knowledge transitions to advance the course flow. Meanwhile, the *PA designer* documented these reflections. This dialogue directly influenced the design of naturalistic instructional gestures for subsequent phases.

3.3.3. Stage 2: Human PA Acting (performance-based rehearsal)

In the second iteration, the Human PA acting phase, the educator acted as a "human prototype" to externalize tacit pedagogical knowledge through performance (Figure 5).

In the second iteration, known as the Human PA acting phase, involving the *educator* as the central figure. During this stage, the educator primarily focused on professionally delivering the prepared presentation, integrating instructional gestures summarized from the previous phase. The educator seamlessly incorporated them into natural, rhythmic hand movements, designing for a smooth flow that closely mirrors everyday teaching interactions (see Figure 5(D)). This stage consisted of multiple rounds of (*a) presentation rehearsals and acting out* to ensure that gestures, speech rate, and intonation were synchronized timely and coherently with the presentation contents.

Following these rehearsals, the educator formally delivered the lecture in English, adhering to the presentation notes, visual aids, and gesture design criteria developed in the prior stage (see Figure 1(B)). The *PA designer* observed the gestures performed by the educator, and both parties reflected on how to better



(D) Presentation with gesture explorations

(E) Correcting and synthesizing human gestures

(F) Human-voiced presentation with human gestures acting

Figure 5. Stage 2: Human PA acting.

express word-synchronized gestures to (*b) correct and synthesize the human gesture samples* (see Figure 5(E)). For instance, when an image of a swelling pufferfish was displayed on the screen, the teacher made a obvious raising-hands gesture to emphasize the new concept. The entire process was recorded using an iPhone with the voice captured simultaneously using AirPods Pro as a separate audio track, ensuring high-quality audio.

Subsequently, the recording was trimmed and edited into a 14-minute presentation, serving as a key reference for the *PA designer*. However, flaws such as inconsistent transitions between gestures appeared in the draft recording. Therefore, the PA designer shadowed the educator's gestures in context, a process that was also being recorded, yielding a smoother and more well-organized (see Figure 1(C)) *c) presentation video with synthesized educator's gestures* acting out, thereby maintaining consistency in the gestures and transitions between them (see Figure 5(F)).

3.3.4. Stage 3: EPA acting. (Human-to-agent motion transfer)

The third phase, *Embodied PA Acting*, emphasized transforming the human performance into an embodied agent form through human-to-agent motion transfer via video-based pose estimation and manual joint correction (Figures 1(D) and 6(G)).

The third phase emphasized transforming the human PA form into an embodied agent form, specifically concentrating on the post-production aspects of 3D avatar design, mainly executed by the *PA designer*. Moving from human performance video to an EPA form serves a purpose beyond replication: it decouples gestural content from the identity of a specific presenter, enables systematic modification of gesture parameters across iterations, and provides a controlled research instrument for isolating gesture-specific effects on student perceptions. We opted for minimalist, robotic avatars over humanoid configurations to focus participants' attention on the instructional gestures—the primary research focus. While prior research indicates that agent appearance significantly influences student perceptions (Beege et al., 2022; Shiban et al., 2015), our faceless design choice aimed to isolate gesture-specific effects by minimizing confounding factors related to facial features, physique, and attire. We retained the original human-voiced soundtrack to mitigate the potential bias introduced by synthesized speech quality, which could impact the learning experience (Dai et al., 2022). This also ensures the provision of appropriate social cues for student engagement, while maintaining a level of abstraction to prevent unintended associations or biases related to physical appearance (see Figure 6(G)). To achieve this transformation, several AI-based motion capture and tracking techniques were explored, where a 3D animated avatar was constructed to mirror the educator's presentation, including platforms such as Plask Motion[8], MoveAI[9], and DeepMotion[10]. We settled on DeepMotion due to its robust pose estimation capabilities, particularly its accessibility and precise hand-tracking. Subsequently, the previously recorded human PA acting video was uploaded to DeepMotion and then transformed to 3D animated avatar video.

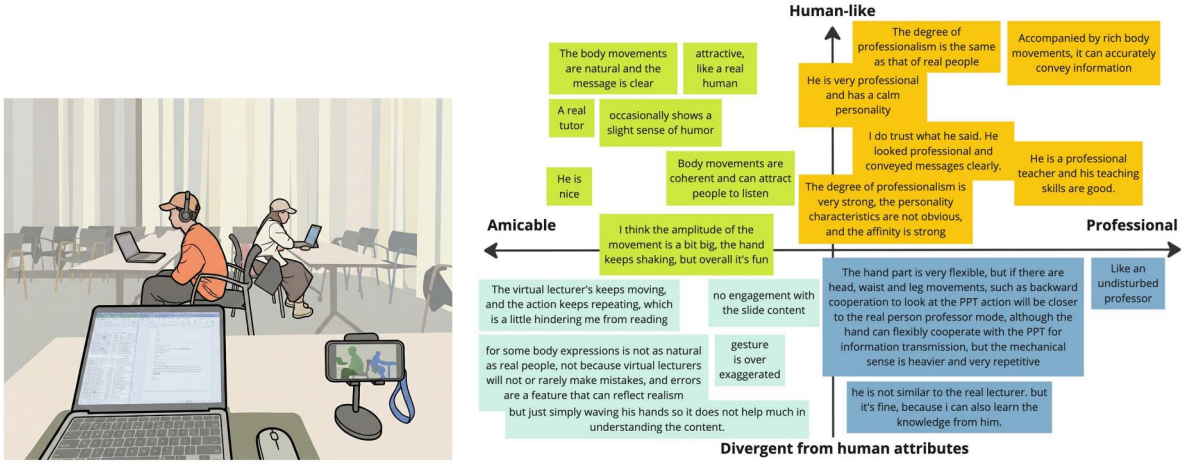
The *PAs designer* refined human PA video, excising repetitions and reorganizing the footage into seven 2-minute sequences to optimize readability and transform capability on the DeepMotion platform. Upon completion of the transition from human to 3D PA, minor adjustments were made to the avatar joints to rectify issues, especially concerning the conversion of 2D videos into 3D animations, where the depth information was frequently either imprecise or entirely absent (see Figure 6(H)). In such instances, the PA designer manually moved the joint nodes to properly show the gestures, adhering to the principle of texturing the gestures as naturally and human-like as possible. Subsequent to the rendering of the 3D animated PA and its integration into the presentation slides using Adobe AfterEffect, we yielded a *c) presentation video with rendered PA gestures* (see Figure 6(I)). The half-body size of the PA is positioned on the right-hand side of the video frame to optimize visual clarity. This intentional design configuration facilitates an unobstructed view of both the educational materials and the PA's gestural expressions, thereby enhancing the students' capacity for cognitive engagement.

3.3.5. Stage 4: EPA-assisted course delivery

The final stage involved deploying the EPA-assisted presentation (see Figure 7(J)) in an actual learning context to evaluate its effectiveness and gather user feedback (Figure 7).

(G) EPAs gestures explorations (H) EPAs gestures adjustments (I) Presentation with EPAs gestures

Figure 6. Stage 3: EPA acting.



(J) User studies (K) Synthesizing feedback

Figure 7. Stage 4: EPA-assisted course delivery.

3.3.6. EPA-assisted course delivery

To investigate the efficacy and user experience of a virtual lecture featuring a gestural PA customized to address teachers' multimodal instructional needs. We conducted *a) user study* on campus (see Figure 7(J)). The student participants were recruited through on-site poster advertising. English is their second language, and all participants possess an IELTS score of 6 or equivalent, enabling them to understand courses primarily taught in English. Eligibility criteria stipulated that candidates possess limited prior knowledge of the subject matter but exhibit a keen interest in interactive materiality. Upon arrival at a designated empty classroom, each participant was furnished with a laptop and noise-canceling headphones. The self-paced learning module had an average duration of 45 min. A researcher was on-site to facilitate the study and address any queries, ensuring an environment devoid of extraneous disruptions. Comprehensive audio-visual recordings, along with on-site notations, were secured for subsequent analysis (see Figure 7(K)).

4. User study

4.1. Method and materials

Our user study aimed to evaluate the learning experiences and perceptions of EPA-assisted courses customized according to the instructional needs of educators. The study was conducted through a structured three-phase evaluation protocol (on-campus setup, see Figure 7(J)). The initial phase consisted of a pre-questionnaire encompassing a demographic inquiry and a pre-transfer test to assess participants' baseline knowledge of the subject. The second phase featured a 14-minute video-based instructional session. This duration was selected based on pedagogical recommendations for university-level content (Lagerstrom et al., 2015), suggesting that 12–20 min provide sufficient exposure for students to form

stable perceptions of an agent's social presence and benefit from cumulative gestural cues (Castro-Alonso et al., 2021). The 14-minute format also enabled comprehensive coverage of complete instructional content while maintaining engagement, aligning with our research focus on perceptual validation rather than long-term learning outcome assessment. The final phase comprised a post-questionnaire that included a post-transfer test, seven subquestions adopted from the Agent Persona Instrument (API) (Baylor & Ryu, 2003), which covered four aspects: engagement, person-like qualities, instructor-like qualities, and credibility, along with two open-ended questions designed to elicit in-depth reflections on the educational experience. In the pre-questionnaire, we gathered data on participants' gender, age, major, and educational level. Additionally, we assessed their English proficiency, familiarity with the theme of the course we designed, and their level of interest in the course. For these assessments, we employed a 7-point scale for each question. The pre- and post-transfer tests consist of the same set of 15 questions. These questions are designed based on an average of one key concept appearing for every two slides in the lecture videos, with a total of 30 slides. The lecture script was designed at CEFR B2 level (Flesch-Kincaid Grade Level = 11.6; Flesch Reading Ease = 42.3) (Flesch, 1948; Kincaid et al., 1975), and the narration was delivered at a speaking rate of 142 words per minute (wpm)—a moderated pace suitable for complex technical subjects that balances cognitive load management with engagement (Pastore, 2012; Williams, 1998). The teacher explains related concepts using various methods, such as presenting relationship diagrams and providing examples. Transfer test scores are determined as follows: 3 points for integrating video knowledge with personal explanation and examples; 2 points for reciting video content with examples; 1 point for partial recall or somewhat relevant examples; 0 points for unrelated answers. The maximum score per lecture is 45 points. The open-ended questions were formulated as follows: "*Can you provide a detailed description of your experience with today's virtual learning process, offering specific examples where possible?*" and "*What are your impressions of the PA? Please elucidate on aspects such as professionalism, personality traits, and the efficacy of body movements and communicative gestures.*"

4.2. Subjects

The study initially screened candidates to ensure they met specific pretest criteria, resulting in a final sample of 38 students recruited from a university in Hong Kong. This sample comprised 32 females and 6 males, aged between 18 and 26 years old ($M = 21$; $SD = 2.05$). The participants come from diverse academic backgrounds including Business ($n = 17$), Engineering ($n = 9$), Health Science ($n = 7$), Computing ($n = 2$), Design ($n = 2$), and Linguistics ($n = 1$). Among them, 27 were undergraduates and 11 were postgraduates. As English is the medium of instruction at Hong Kong universities, all participants were required to demonstrate a minimum English proficiency equivalent to an IELTS score of 6 or higher, ensuring their ability to comprehend the lecture script (CEFR B2 level). While the majority of participants were local or from mainland China and Southeast Asia—speaking English as a second language—they self-rated their proficiency for academic communication as relatively high ($M = 4.89$, $SD = 0.89$). Their familiarity with the subject of our tested course was relatively low ($M = 2.95$, $SD = 1.49$), yet their interest in the course was high ($M = 4.84$, $SD = 1.41$). Following the lecture, participants rated the perceived difficulty of the content at $M = 4.16$ out of 7 ($SD = 1.35$), indicating that the material—designed at CEFR B2 level—provided a sufficient cognitive challenge without exceeding their comprehension levels, which corroborates the alignment between the script's readability metrics and students' actual learning experience.

4.3. Data analysis

Our primary research interest centered on students' qualitative perceptions and experiences of gesture-based PAs. To ensure the validity of these interpretations, we first established that participants engaged meaningfully with the instructional material and formed coherent impressions of the PA. Paired-sample t-tests on pre- and post-transfer scores verified knowledge acquisition during the session, while descriptive statistics for API dimensions (engagement, person-like qualities, instructor-like qualities, and credibility) confirmed that participants developed stable perceptions of the PA's social and instructional

presence. These quantitative checks served as essential prerequisites for interpreting the subsequent qualitative data.

The core analysis focused on open-ended interview responses, which underwent thematic content analysis (Clarke & Braun, 2017). Two independent coders reviewed the qualitative feedback, identifying recurring themes related to the PA's professionalism, human-likeness, approachability, and gesture effectiveness, guided by a shared coding protocol developed during the Stage 1 calibration phase. Where categories were disputed, both coders discussed until consensus was reached, following established consensus coding procedures (Johnny Saldaña, 2021). Representative quotes were then organized into a two-dimensional feedback categorization matrix, systematically mapping student perceptions across dimensions of human-likeness versus professionalism and human-likeness versus approachability. This approach enabled us to synthesize patterns in how specific gesture types (cohesive, beat) influenced students' subjective experiences and social impressions of the PA, thereby generating grounded insights to inform future design iterations in gesture-based embodied pedagogy.

4.4. Results

Upon completion of data collection and subsequent data cleaning, we performed a paired-sample t-test, revealing statistically significant disparities between pre- and post-transfer scores (pre-transfer $M = 3.95$, $SD = 2.99$; post-transfer $M = 7.82$, $SD = 4.37$; $t(37) = 6.02$, $p < 0.001$, Cohen's $d = 0.98$), confirming that participants engaged meaningfully with the instructional material. Additionally, Students' scores for API engagement ($M = 4.46$, $SD = 1.55$), API person-like qualities ($M = 4.39$, $SD = 1.56$), API instructor-like qualities ($M = 4.97$, $SD = 1.16$), and API credibility ($M = 4.68$, $SD = 1.58$) reflect a generally positive impression of the PA, as students perceived it to be human-like and capable of conveying important learning information, along with a high level of credibility. These results confirm that participants engaged meaningfully with the lecture and formed stable perceptions of the PA. Then we engaged in the *b) synthesized feedback* garnered from student participants. Eighteen out of thirty-eight respondents reported that the instructional materials were understandable and well-organized, while seventeen expressed a desire for opportunities to continue learning related courses. Their feedback described the course as "interesting," "satisfying," and "meaningful." Fourteen respondents noted that the PA resembled a real person, seven commented on the naturalness of its movements, and another seven described the PA as amicable. Nineteen respondents perceived the PA as professional, with remarks such as, "the degree of professionalism is the same as that of real people," along with observations about its organized speech and occasional humor. However, four respondents mentioned that the PA's hand movements were noticeably repetitive and exaggerated, using phrases like "very intelligent use of a lot of body language."

Following the analytic approach described above, the feedback categorization matrix presents our findings across four distinct dimensions of the PA's human-like characteristics: human-like and professional, human-like and amicable, professional yet divergent from human attributes, and amicable yet distinct from human attributes. Students who perceived the PA as both human-like and professional offered feedback such as, "*He is very professional and has a calm personality*" and "*accompanied by rich body movements, it can accurately convey information.*" Conversely, certain students expressed that while the PA possessed a professional demeanor, the semblance to a real human was not entirely convincing. One student commented, "*he is not similar to the real teacher. but it's fine, because I can also learn the knowledge from him.*" Meanwhile, a majority of students conveyed that the PA remarkably resembled a genuine human and projected an agreeable personality. They cited instances such as "*occasionally shows a slight sense of humor*" and "*body movements are coherent and can attract people to listen.*" However, a subset of students contended that the PA's authenticity was compromised by repetitive gestures, with one student noting, "*the PA's keeps moving, and the action continues to repeat, which is a little hindering me from reading the words on the slides,*" and by the absence of interactions with the surrounding content, as voiced by another student, "*there is no engagement between the PA and the visual elements on the slides.*"

Based on the matrix analysis, we deduced four distinct directions for further refinement. The initial direction underscores that the integration of cohesive and beat gestures into the PA design can

engender a heightened sense of professionalism among students. The synchronization between speech rhythms and beat gestures effectively punctuates crucial information, mimicking the discernible behavior of a genuine teacher. Secondly, cohesive gestures play a pivotal role in facilitating smooth transitions and connections within the PA's gestural repertoire, thereby cultivating a sense of approachability among students. Thirdly, consideration should be accorded to the integration of the PA's bodily movements with instructional materials and the audience. This strategic alignment bolsters students' engagement and engenders a heightened sense of social involvement. Lastly, the tendency for repetitive gestures to enhance visual awareness should be recognized. However, excessive use can negatively affect students' perceptions, leading to feelings of unreality and emotional detachment. Overall, incorporating pedagogically grounded gestures into PAs enhances students' social presence and perceived instructional quality, thereby enriching their learning experience and creating conditions conducive to engagement and knowledge acquisition.

5. Insights for the design of instructional gesture generation

Building on the student feedback patterns identified above, this section synthesizes insights from our four-stage methodology—from multimodal analysis of authentic teaching (Stage 1) through collaborative prototyping (Stage 2), technical implementation (Stage 3), to student evaluation (Stage 4)—into actionable design principles. Our approach offers a replicable process for developing pedagogically grounded gestures: analyzing gesture patterns in authentic instruction, collaboratively externalizing educators' tacit knowledge through rehearsal, navigating technical constraints during implementation, and validating designs through student perceptions. Our focus on cohesive and beat gestures is grounded in both empirical prevalence (these types dominated observed instructional discourse) and theoretical transferability: unlike iconic or metaphoric gestures tied to disciplinary terminology (Masson-Carro et al., 2017; Wei & Chow, 2023), cohesive and beat gestures serve fundamental communicative functions—linking concepts, marking emphasis, regulating rhythm—applicable across domains (Belhiah, 2013; McNeill, 2011). What varies across contexts is the linguistic anchors (which words trigger emphasis, where transitions occur) and timing parameters, not the gesture types themselves. The five instructional functions we identify (emphasis, flow, emotion, rhythm, connection) and the methodology for operationalizing them show potential for application across educational domains.

5.1. Teaching objectives and multimodal instruction design

The corpus analysis of 71,252 gesture samples from 13 design and art lectures (Stage 1) revealed that cohesive (44.34%) and beat gestures (38.26%) dominated instructional discourse. This distribution extends prior findings on word-synchronized gestures (Wei & Chow, 2023), providing further evidence that these gesture types frequently accompany specific linguistic structures in teaching contexts. This foundation informed our gesture selection for the PA design case. One of the first steps in the collaboration was to clarify the teaching objectives, as these directly influence the type of gestures required. Educators would articulate their instructional goals, such as explaining and emphasizing key concepts or guiding students through a thought process. From a multimodal instruction design perspective, instructional gestures structure information flow, regulate cognitive load (Mayer, 2002), and amplify verbal content (Goldin-Meadow, 2015). Based on these goals, PA designers would translate them into indexes and labels for instructional intents. The educator's instructional intents can be categorized into five key types. These include:

1. Emphasis on Key Concepts: This is achieved by varying the pace, size, and amplitude of gestures, which helps draw attention to important knowledge or transitions.
2. Natural Flow of Communication: Changes in the direction of the gestures' movement contribute to a smooth and natural flow of interaction, facilitating better communication with the students.
3. Emotional Expression: The contrast between forward and backward gestures, as well as the directional movement of the gestures, helps convey emotional tone and intent, aiding in emotional engagement and emphasis.

4. Adjustment of Course Rhythm: Pauses or breaks in the gesture flow help regulate the pace of the lesson, providing students with the opportunity to process information.
5. Connection Between Concepts: Smooth transition gestures are employed to link key points of the lesson, maintaining continuity and coherence in the instructional flow.

For instance, when the educator needed to explain an abstract concept, the instructional gesture could be labeled as an "emphasis gesture" by the LLM-based generative retrieval system, manifested through an increase in amplitude and described as "hands moving vertically." This gesture would signal important knowledge points to the students, highlighting shifts in the educator's communication and helping students identify key concepts or transitions in the material, thus enhancing focus and comprehension. In another example, when the educator's script indicated a semantic shift, such as "Now let's move on to the next step in the design process," the LLM-based generative retrieval system could label the gesture as a "cohesive gesture," guiding students through the transition and helping them follow the change in focus.

5.2. Translating student feedback into design specifications

Based on the feedback categorization matrix, we identified four main needs from students that can guide the refinement of PA gestures. These needs span across gesture variability, engagement with content, professional demeanor, and human-likeness:

1. Gesture Variability: Some students noted that repetitive gestures hindered their engagement. Therefore, there is a need for more varied gestures to maintain students' focus and to enhance the authenticity of the PA.
2. Engagement with Instructional Content: Students expressed a desire for the PA to interact more dynamically with the instructional materials, such as slides. This would improve student engagement and create a more interactive learning experience.
3. Professional Demeanor: Students who perceived the PA as professional yet lacking in human-like qualities indicated the need for gestures that align more closely with human communication patterns, enhancing both the professionalism and relatability of the PA.
4. Human-Likeness: While many students appreciated the PA's friendly demeanor, they suggested that more social and emotional cues, such as humor or varied gestures, would improve the overall learning experience.

By addressing these needs—enhancing gesture variability, improving interaction with instructional content, and balancing professional and human-like qualities—PA designs can be refined to better engage students and support their learning process. For example, when students provide feedback such as, "The gesture is unclear or confusing," the LLM-based generative retrieval system can generate prompts based on the semantic context to adjust the direction of the gestures. This ensures that the gestures' aiming direction is clearer and aligns with the visual content, enhancing the overall clarity and effectiveness of the communication.

5.3. Synchronization with speech

Synchronizing gestures with speech was a key aspect of the collaborative design process. This involved aligning gestures not only with the semantic content and instructional intent but also with temporally appropriate moments in speech. Building on research correlating gesture types with parts of speech (Wei & Chow, 2023), we developed operational heuristics through Stage 2 rehearsals that guided gesture-speech integration. While the specific timing parameters emerged from English prosody and our case context, the underlying principle—that gestures should synchronize with linguistic structure—applies across languages, though local calibration may be needed for different rhythmic patterns. To facilitate this, the gesture generation system should incorporate gesture timing annotations relative to

PoS, enabling precise temporal alignment that allows gestures to integrate seamlessly with verbal communication rather than appear disjointed.

5.4. Iterative refinement

Throughout the process, iterative testing and refinement were central to the collaboration. The PA designer and the educator continuously worked together, conducting mock sessions and gathering feedback from students. This iterative approach allowed for fine-tuning both the gestures and their synchronization with the lecture content. The PA designer's shadowing process (Stage 2c) exemplifies this iterative refinement: while the educator's initial performance featured rich semantic content, some transitions appeared abrupt when isolated for animation. By mirroring the educator's movements and experimenting with interpolated transitions, the PA designer smoothed the gestural flow, ensuring each movement had clear onset, stroke, and retraction phases—a principle rooted in gesture phase theory (McNeill, 2011) that transcends specific instructional contexts. Feedback loops helped ensure that the gestures were not only pedagogically effective but also culturally sensitive and contextually relevant to the students' learning environment.

5.5. Exploration of GenAI tools

We generated instructional gesture-based teaching videos using commercial text-to-video generation tools (i.e., Sora and KlingAI) to explore their pedagogical applicability. In our iterative testing, we observed that Sora produced highly realistic visuals with minimal body deformation and executed simple gestures like directional movements effectively, while KlingAI offered motion trajectory brushes that enabled finer control over gesture sequences. Through testing various educator-intent prompts and researcher discussions, we found that prompt structure plays a critical role in shaping gesture rhythm, transitions, and emphasis. Prompts specifying temporal phases, "*performs an emphasizing gesture, briefly pauses, slightly turns*" - yielded significantly more coherent output than semantic labels alone. For instance, "cohesive gesture" produced ambiguous results, but "a lateral hand sweep from left to right, synchronized with a specific timepoint" generated recognizable transitions. This extends insights from Stages 1–2: linguistic anchoring and temporal structuring improve GenAI output quality. As an example, we produced a 5-second PA video in which the teacher explains an abstract concept using an emphasis gesture followed by a directional body shift, enhancing communicative flow (see Figure 8). Without such explicit structuring, generated videos often lack temporal coherence, reducing communicative clarity. These findings suggest that future text-to-video systems should incorporate timing, transition, and emphasis cues as explicit prompt parameters to support pedagogically grounded generation.

To quantitatively evaluate these generation approaches and compare the effectiveness of different prompt strategies, we conducted a controlled comparison of four generation methods: (1) trajectory-augmented textual prompts (KlingAI 1.5 with motion paths), (2) structured textual gesture prompts

Figure 8. KlingAI Generated PA in a sequence of gestures animated by using the prompt, "*a male teacher is enthusiastically describing an academic definition. He performs an emphasizing gesture, briefly pauses, and slightly turns his upper body to look toward the right side of the students. The shot only includes the teacher, with the camera fixed at a medium-close range, and the teacher remaining stationary.*" the first gesture highlights emphasis, and the second demonstrates intent, as the PA slightly turns their upper body toward the right side, as if engaging with students on that side.

(Sora), (3) structured textual gesture prompts (KlingAI 2.6), and (4) baseline with speech script only (KlingAI 2.6, no gesture instructions). Through online recruitment, we screened respondents based on having at least one year of experience in educational practice and one year of familiarity with AI tools, ultimately selecting 12 qualified experts (aged 29.67 ± 4.50 years; 4 male, 8 female). Each expert rated the four videos on nine rating items using 7-point Likert scales; items were grouped into six dimensions: gesture execution quality, prompt adherence, temporal coherence, visual realism, instructor authenticity, and instructional suitability.

Repeated measures ANOVAs revealed significant main effects for all six dimensions (all $F > 3.99$, $p < 0.05$, partial $\eta^2 = 0.27$), with prompt adherence showing the largest effect ($F(3, 33) = 15.26$, $p < 0.01$, partial $\eta^2 = 0.58$). Post-hoc pairwise comparisons with Bonferroni correction ($\alpha = 0.008$) indicated that trajectory-augmented guidance (condition 1) significantly outperformed other methods on multiple dimensions (all $p < 0.008$, $d = 0.40$–2.04). Critically, all gesture-enhanced conditions (1–3) significantly surpassed the baseline (all $p < 0.008$, $d = 0.66$–2.93). These findings are consistent with our exploratory observations, providing quantitative evidence that structured gesture prompts (temporal structuring, linguistic anchors, and motion trajectories) substantially improve instructional quality and enable a human-in-control workflow for pedagogically grounded video generation.

5.6. Methodological contributions and scope

Advancing existing knowledge. First, while prior research documents gesture importance (Li et al., 2022; Schneider et al., 2022), systematic frameworks for translating pedagogical intent into implementable specifications remain limited—our framework operationalizes this through a structured methodology that encompasses corpus analysis, collaborative rehearsal, technical implementation, and student validation. Second, existing synthesis research evaluates naturalness (Wolfert et al., 2022; Zhang et al., 2024), yet few studies empirically link teaching behaviors to student perceptions of pedagogical dimensions—our feedback matrix shows API ratings and qualitative evidence validate word-synchronized gesture design. Third, our GenAI exploration reveals structured prompts (temporal phases and linguistic anchors) produce more instructionally coherent output than semantic labels alone.

Contextual grounding. This methodology was developed and validated through one instructional case: a 14-minute lecture on material-centered design processes, delivered in English to university students (primarily undergraduate level) in a Hong Kong educational context where English serves as the medium of instruction yet most students are second-language speakers. The educator explained a procedural workflow through step-by-step demonstration, using cohesive gestures to connect sequential phases and beat gestures to emphasize key concepts. This linguistic context shaped both the lecture's moderate speaking rate (142 wpm) and the design emphasis on multimodal cues beyond verbal content. The case reflects our human-centered approach: we began by observing authentic teaching to identify prevalent patterns, captured an educator's tacit pedagogical knowledge through collaborative rehearsal, and validated designs through student perceptions of instructional quality. This case demonstrates how the framework operationalizes HCAI principles—preserving educator agency while leveraging computational tools—and the design heuristics derived from this process inform future applications where similar collaborative, reflective practices can translate instructional intentions into embodied agent behavior.

6. Discussion

Our work, centered on designing and evaluating pedagogical agent (PA) gestures, is presented as a methodological case study for the broader field of Human-Centered AI (HCAI). This discussion reflects on our process and findings in relation to the core challenges of HCAI, particularly the need to bridge the gap between high-level principles and the concrete, practical work of designing and implementing intelligent systems. We structure our contributions into three main areas: the human-centered design framework we developed, the empirical results of its application, and a critical assessment of how our process informs the future of Generative AI.

6.1. Novel aspects of the pedagogical agent design process

6.1.1. Collaborative reflective practices

A distinctive feature of our process was the integration of collaborative *reflection-in-action* (Schon, 1992) during the *Human PA Acting* phase. While reflective practice traditionally centers on individual cognition, our approach extended this to a dyadic context where the *educator* and *PA designer* engaged in the iterative refinement of gestures, resembling an ongoing decision-making process. Within this collaborative setting, gestures were continuously explored, evaluated, and adjusted through real-time feedback and shared reflection. Although the *educator* worked from a pre-scripted presentation, multiple rehearsal rounds proved essential for this reflective cycle. This process revealed a key benefit of collaborative reflection: the *educator*'s propensity for subconscious or extraneous gesturing was identified by the *PA designer*, who, serving as an external observer, provided immediate feedback. This intervention was crucial for mitigating unintended movements and suggesting enhancements for clarity. Through this dynamic of observation and adjustment - a process reminiscent of collaborative human-AI interaction—the *PA designer*'s role was instrumental in ensuring that gestures were precisely aligned with pedagogical goals, thereby enhancing their overall communicative efficacy.

6.1.2. Improvizational practices in gesture design

A core characteristic of our design process was the use of structured improvization for gesture creation. In contrast to conventional teaching where educators' gestures are often spontaneous and may lack consistent alignment with verbal content, our methodology employed iterative rehearsals across the *Human PA Acting* and *EPA Acting* stages. This was done to situate gestures meaningfully within the instructional narrative. During the *Human PA Acting* phase, the rehearsal process afforded the *educator* the flexibility to adapt and refine gestures in real time, fostering an alignment with the teaching content that was both natural and intentional.

This improvizational flexibility ensured that gestures were not only synchronized with speech but also contextually tailored to the educational material, thereby enhancing their pedagogical relevance. This method established a robust foundation for the *EPA Acting* stage, generating a repository of refined human movements that the *PA designer* could translate into a cohesive gestural performance for the agent. While this approach ensures high fidelity between the human-performed gestures and their EPAs' embodiment, it introduces a practical tradeoff: the process is time-intensive, requiring multiple iterations to achieve the desired level of refinement in both gestures and posture.

6.1.3. Re-contextualizing role-playing as a design method

In our PA design process, we re-contextualized the established design research method of role-playing—also known as "acting out" (Druin, 2002)—specifically for the design of PA gestures. Originally employed to give designers an embodied understanding of the user experience, we adapted this method to generate and refine gestural interactions. During the *Human PA Acting* phase (see Figure 3), the *educator* engaged in an immersive enactment of the script, concentrating on the coherent and timely synchronization of gesture with speech. This performance was augmented by multiple explorative rehearsals aimed at discovering the most intuitive gestural expressions.

Concurrently, the *PA designer* adopted the dual roles of observer and proxy audience. In this capacity, they provided critical feedback to calibrate the *educator*'s gestures for optimal precision and expressiveness, preventing them from being either exaggerated or overly subtle. This interaction also engendered a sense of social presence for the *Educator*, reinforcing the awareness that they were addressing students and thereby sharpening their focus on effective gestural communication (Goldin-Meadow, 2015). Ultimately, this application of role-playing fostered greater empathy with the prospective students' experience and provided invaluable first-hand feedback integral to the design process.

6.2. Toward human-centered AI for gesture generation

A key contribution of this work is the presentation of a structured, four-stage methodology that guides the design of PA with instructional gestures. This framework responds to the challenge identified in

previous research (Xu & Gao, 2023) regarding the difficulty of translating abstract human-centered AI principles—such as amplifying human capabilities (Ozmen Garibay et al., 2023; Shneiderman, 2020)—into concrete and replicable design practices. By grounding the process in educators' instructional intents and fostering iterative collaboration, the approach provides a practical and flexible methodology to operationalize these principles, offering valuable guidance for the design and development of embodied AI systems. In doing so, it contributes to advancing human-centered AI by informing the creation of PAs that have the potential to enhance teaching and learning experiences.

Our methodology is distinguished by two key features. First, it centers human expertise. In contrast to purely data-driven approaches that train models on generic data, our method begins with and iteratively returns to the embodied, tacit knowledge of a human expert. The *Human PA Acting* stage is a critical methodological step for eliciting and capturing nuanced performative requirements that are otherwise difficult to articulate; Second, it provides a clear pathway from requirements to implementation. It establishes a transparent and traceable link from the initial analysis of human gestures (Stage 1) to the final EPA (Stage 3). This ensures that the nuanced requirements identified during the ideation and acting phases are faithfully carried into the downstream stages of design and development.

This four-stage process may offer a structured blueprint for developing future HCAI systems where human expertise actively guides and refines AI-driven outputs. Here, AI serves not as a replacement for the designer but as a powerful collaborative partner. Much of contemporary AI development, even when labeled "human-in-the-loop," remains data-centric, focusing on collecting vast amounts of data where the human role is often reduced to that of a data producer. In contrast, our workflow advocates for an intent-centric perspective. The primary input is not raw data, but the structured, embodied, and articulated pedagogical intent of an expert. In Lee and See's (Lee & See, 2004) terms, pedagogical intent constitutes the purpose dimension of trustworthy PA design—the educator's goals inherited by the system through deliberate design. The framework makes this intent explicit and traceable: surfacing tacit knowledge (Stage 1) through collaborative rehearsal (Stage 2), embedding it through motion transfer (Stage 3), and validating it through student perceptions (Stage 4). The *Human PA Acting* stage is specifically designed to capture a high-fidelity representation of this intent. For HCAI research, this marks a crucial shift: it provides an epistemological basis for building AI systems that are not merely statistically accurate relative to a dataset, but are semantically aligned with the goals and values of their human collaborators.

The framework is structured as a human-AI collaboration in which educators and designers retain authorship over instructional intent, while AI tools handle motion capture and gesture generation in service of that intent. The human investment across stages is the means by which pedagogical intent is extracted, formalized, and made available as a standard for automated systems to build upon. As intent-aware generation tools mature, individual stages of this process become natural candidates for progressive automation, guided by the pedagogical standards established through the human-centered process.

The development of gestural PAs raises important questions regarding pedagogical authority and student trust in AI instructors. Our study found positive ratings on credibility and professionalism, yet responsible deployment requires transparency about AI involvement (Shneiderman, 2020), clear communication of system limitations, and institutional frameworks that position PAs as supplementary tools preserving human educators' essential roles in mentorship and adaptive support. Research indicates that trust in algorithmic systems depends critically on users' understanding of system boundaries and institutional context (Burton et al., 2020), highlighting the need for thoughtful implementation policies.

6.3. Empirical validation and design insights

6.3.1. Impact of gesture types on student perceptions

Our findings affirm the potential impact of cohesive and beat gestures on students' perceptions, suggesting that these gestures can improve professionalism and approachability in PAs. This aligns with previous research (Nakagawa et al., 2021; Schneider et al., 2022; Sinatra et al., 2021), which suggests that gestures could enhance social engagement and learning experiences among students in educational

settings. The findings also provide empirical evidence that extends the existing understanding of how gestures contribute to enhancing the effectiveness of educational interactions (Beege et al., 2020; Davis, 2018; Schneider et al., 2022). In particular, the study's identification of cohesive and beat gestures as enhancing students' sense of professionalism and a pproachability on PA expands upon the previous research's focus on the multifaceted functions of gestures in instructional communication.

6.3.2. Context as the core of human-centered AI gesture design

A key insight from our study is the necessity of designing for context, a dimension frequently overlooked by current generative AI. While many AI models can generate realistic human motion, they often lack an understanding of the situational factors that make a gesture meaningful. Our work demonstrates that the synergy between a presenter's bodily movements, the specific educational materials, and the engagement of the audience is what fosters a vital sense of social involvement. This aligns with previous findings that the educational environment itself shapes gestural communication (Wei & Chow, 2022).

Ultimately, both our study and prior work argue that for gestures to enhance the situated learning experience, they cannot be generic; they must be contextually appropriate. This has profound implications for the design of human-centered AI, particularly for PAs in immersive settings like VR (Petersen et al., 2021). The future goal must be to move beyond mere kinematic mimicry and toward creating agents whose gestures are semantically rich and socially aware, thereby transforming an immersive environment into a truly collaborative and intelligent educational space.

6.3.3. Balancing gesture frequency and variety

The study highlights the potential pitfalls of excessive repetitive gestures and the need for a balanced approach to gesture design. This recognition of balance is reminiscent of studies that have examined the optimal use of gestures to augment learning. For instance, Pi and her colleagues (Pi et al., 2022) discussed the complexity and frequency of gestures in relation to their impact on student engagement and perception. This contributes to the broader discussion on how gesture frequency and variety intersect with learning experiences. Furthermore, the call for a balanced approach aligns with previous investigations into gesture complexity and variability. Studies have shown that the use of gestures, including both iconic and deictic gestures, can enhance students' engagement with different concepts (So et al., 2012). Similarly, the negative implications of excessive repetition resonate with concerns raised in previous research regarding the potential cognitive overload caused by constant or monotonous gestures. For further development of gestural PA design, PA designers should consider utilizing various gestures and carefully balancing their frequency to enhance the sense of realism and meaningfulness while also avoiding cognitive overload for students.

6.4. Generative AI and future directions

6.4.1. Fields of application

The design and application of gestures in PAs have broad implications for various fields within and beyond educational technology. Building upon the foundational insights provided by Petersen et al. (Petersen et al., 2021), regarding the potential of PAs to improve educational experiences within VR environments, we investigated and contributed to the intricacies of gesture design for PAs, specifically focusing on word-synchronized gestures. For online education in particular, gesture-enabled PAs could serve as virtual teacher clones designed to supplement educators during fatigue or technical difficulties. This opens up the possibility for more engaging and authentic AI-driven presentations and addresses practical problems such as teacher absenteeism. Another field is that of *AI Spokespersons* (e.g., AI Humans (DeepBrain[11], Synthesia[12])) for multimedia content, particularly in virtual platforms where the agent serves as a stand-in for human presenters. Previous research in AI spokespersons has focused primarily on vocal intonation and facial expressions, and gesture design could be an essential addition. Another promising field is *e-commerce*, which engages in product promotions. However, the heavy reliance on human streamers for constant endorsements can be problematic. This dependence creates a significant manpower workload, misleading consumers with human streamers' emotional cues and

subtle behaviors. To address these issues, gesture-rich virtual agents could be leveraged as a substitute for the real streamer. Given that behavioral cues can significantly impact online user engagement (Ahmed et al., 2022), a gesture-rich agent could substantially improve customer interaction metrics. This alleviates the workload on human presenters and provides a level of consistency and engagement that can be programmed and fine-tuned for maximum impact.

6.4.2. Implications for generative AI teacher design

Building on our exploration of collaborative and reflective practices in PA gesture design, the future of embodied AI teacher development holds promising opportunities for incorporating advanced technologies from HCI, such as Computer Vision (CV), AI-Generated Content (AIGC), and Large Language Models (LLMs). These technologies offer untapped potential to enhance efficiency and flexibility in the PA design process by automating or augmenting aspects of gesture creation and content delivery.

6.4.2.1. Toward transcript-prompted AI teacher synthesis. Tools that leverage AIGC in creative media, as demonstrated by He et al. (2024), where an AI generative model can process the educator's speech transcripts and instructional intents as input, generating multiple options for AI-generated gesture synthesis. Moreover, drawing inspiration from the combination of rule-based and learning-based techniques to select naturalistic instructional gestures (Sadoughi & Busso, 2019; Zhang et al., 2024; Zhi et al., 2023), this research offers valuable insights into crafting prompts for semantic motion generation. This approach can augment the PA's alignment with teaching intents, ultimately enhancing AI's efficiency in delivering multimedia instructions within a virtual learning environment.

Furthermore, our preliminary findings indicate that cohesive and beat gestures contribute positively to students' perceptions of engagement and enjoyment. This insight makes a strong case for additional quantitative studies. One possible approach would be to use computer vision techniques to analyze the relationship between gesture types and parts of speech within large video datasets such as TED Talks. Insights from such analysis could inform the development of Generative Adversarial Network models that optimize gesture and speech coordination, resulting in more authentic and engaging pedagogical agents. AI-generated content technology also provides flexibility to design highly customizable and diverse virtual agent appearances, surpassing the limitations of traditional 3D animation platforms. While AI-generated content tools can handle complex visual customization, large language models could generate responsive instructional content and support real time interaction, which would create a more immersive and dynamic learning experience. Looking ahead in gesture design, a longer term trajectory involves shifting from pre-scripted content to agents capable of adapting gestures in real time. Future systems could utilize multimodal learner sensing, including gaze patterns, facial expressions, or affective signals, to adjust gesture intensity, pacing, and type according to real time engagement states. This would mark a transition from the human in control workflow featured here to a synchronous and adaptive instructional system in which pedagogical intent is not simply encoded at the outset but is continually negotiated between the agent and learner.

6.4.2.2. Toward highly controllable and versatile PA gestures. In this study, we primarily focused on cohesive and beat gestures as foundational elements for PAs in interactive environments. However, these gestures represent only the starting point of a broader exploration. Future research could expand this foundation by incorporating a diverse range of gesture types, each providing unique functionality depending on context and application. One promising direction involves deictic gestures, which can be used to precisely identify, point to, or highlight specific elements within the user interface or learning environment. These gestures are particularly useful in educational or interactive settings where users need to focus attention on particular details. Another area of growth is the inclusion of iconic gestures, which visually represent the content being manipulated or discussed. Such gestures could have immense utility in domains that rely on visualization, such as science and engineering, where complex structures and phenomena need to be depicted dynamically through PAs' multimodal expressions. Additionally, metaphoric gestures can enable abstract and symbolic representations, which would be valuable in fields such as literature, social sciences, or philosophy, where concepts are often intangible and need a more

creative and expressive mode of interaction. The future evolution of PA gestures can therefore offer users a high degree of control and versatility, enhancing the richness and expressiveness of human-computer interactions.

6.4.2.3. Enriching and diversifying training datasets. The use of diffusion-based models for generating co-speech gestures has gained considerable attention. While existing studies predominantly rely on TED talks or interview-style videos, these datasets primarily feature conversational content where gestures are often spontaneous and context-dependent. This focus on informal or conversational settings raises several concerns regarding the applicability and generalizability of the model in instructional or scripted settings. To enhance the model's performance and applicability, we propose expanding and diversifying the training datasets, particularly by focusing on instructional videos that exhibit a broader range of gesture types. In these instructional settings, gestures are typically more deliberate and serve distinct communicative purposes—whether to emphasize key points, clarify instructions, or convey emotions. Videos of educators, coaches, and instructors should be included in the dataset, as these settings naturally offer a greater variety of gestures intended for clear communication. Moreover, including content that specifically focuses on the role of gestures in conveying meaning could significantly enhance the model's understanding and generation of context-appropriate gestures.

6.5. Limitations

6.5.1. Experimental design and scope

This study employed a within-subjects design focused on perceptual validation, consistent with Research-through-Design (RtD) methodologies that prioritize understanding user experience and informing design principles (Zimmerman et al., 2007). We did not include a control condition (e.g., lecture without gestures or static slides), as our primary objective was to validate whether the collaboratively designed PA elicited coherent and positive student perceptions. The observed pre-post transfer gains served as manipulation checks to confirm meaningful engagement, rather than to isolate gesture-specific learning effects. Future controlled experiments with counterbalanced conditions (e.g., gesture-based PA vs. static PA vs. slides-only) are necessary to isolate the specific causal contributions of instructional gestures to learning effectiveness. Additionally, the 14-minute instructional duration, while aligned with pedagogical recommendations for establishing stable agent persona perceptions (Castro-Alonso et al., 2021; Lagerstrom et al., 2015), imposed practical limitations. The brief session restricted the complexity and depth of content covered and may have prevented students from fully experiencing sustained instructional interaction typical of authentic classroom settings. Future research should examine student perceptions and learning experiences in authentic classroom settings with extended instructional sessions. Additionally, our evaluation relied on self-reported perceptions and transfer test scores. Complementing these with objective physiological measures, such as gaze-tracking or EEG-based cognitive load assessment—would provide more direct empirical evidence of gestural contributions to learning processes. We recognize this as a meaningful direction for future work.

6.5.2. Sample population

Another limitation concerns the characteristics of the sample population. The sample exhibited gender imbalance, which may influence perceptions of agent characteristics. The study findings, derived from a specific age group and second language speakers in a Hong Kong educational context where English instruction begins early, may not be generalizable to a broader or different demographic. Future research should employ balanced gender sampling and recruit participants from diverse linguistic backgrounds (e.g., native speakers of Japanese, Russian, or other languages) and educational systems. It is crucial to explore how PAs interact with diverse age groups, cultural backgrounds, and learning styles to gain more comprehensive insights.

6.5.3. Instructional material selection

The teaching material was selected for its pictorial format and procedural structure to ensure cross-disciplinary accessibility. However, we did not explicitly analyze how disciplinary background might mediate

comprehension across the diverse majors represented in our sample. Future research should systematically assess materials' disciplinary adaptability and test the framework with content from multiple knowledge domains.

6.5.4. Pedagogical agent appearance design

The design of the PAs used in the study poses a limitation. We employed a minimalist virtual avatar lacking facial features to isolate gesture-specific effects. However, this design choice was not comparatively evaluated against more realistic agent appearances. Given that agent appearance significantly influences student perceptions (Beege et al., 2022; Shiban et al., 2015), the faceless design may limit the external validity and generalizability of findings to real-world applications where more realistic agents are common. Future research should examine how varying levels of agent realism interact with gesture design across different instructional contexts.

6.5.5. Generative AI tool exploration and evaluation

Our exploration focused on two commercial platforms (Sora and KlingAI), representing a subset of available technologies. Other tools and open-source models may exhibit different capabilities in instructional gesture generation. The quantitative evaluation compared specific model versions, providing initial evidence but limited by small sample size and tool-specific focus. Given rapid technological evolution, findings reflect a particular moment. Future research should expand to diverse platforms, larger samples, and longitudinal assessments to establish broader design principles.

6.5.6. Interactivity of pedagogical agents

The limitation on PA interactivity goes beyond technical issues to include methodological challenges in gesture analysis. A key constraint is the limited gesture repository, which restricts PA diversity and user experience. We addressed this by analyzing gestures from experienced teachers in real scenarios, but future work should expand the gesture database and include more subjects.

7. Conclusion

In conclusion, our work addresses the critical gap between high-level principles and the practical design of co-speech gestures for pedagogical agents (PAs). We introduced and validated a human-centered design framework that empowers educators and designers to systematically translate instructional intentions into meaningful agent gestures. Our evaluation, involving a 14-minute course module assessed by 38 university students, demonstrated the framework's effectiveness. The findings confirmed that an EPA with human-designed gestures was perceived as more professional and approachable, with students describing enriched engagement and positive learning experiences. This research offers three main contributions to the HCI community: (1) We provide a replicable four-stage design framework that enables a reflective and collaborative process for embodying pedagogical goals in an agent. (2) We offer empirical evidence that well-designed cohesive and beat gestures improve student perceptions and support meaningful learning. (3) Through exploratory testing and quantitative evaluation of text-to-video generation tools, we demonstrate that structured prompts with temporal phases and linguistic anchors improve instructional coherence, providing concrete design principles for future AI-generated instructional content.

Notes

1. Throughout this paper, we use PA (Pedagogical Agent) as the broader category that includes both non-embodied and embodied instructional agents, and EPA (Embodied Pedagogical Agent) to refer specifically to agents with a three dimensional physical or virtual form capable of producing body movements. When describing the visual style of our specific design artifact, a minimalist and faceless 3D avatar, we use 3D animated avatar as an appearance descriptor rather than a category designation.
2. DeepMotion. Available at: https://www.deepmotion.com/.
3. HeyGen. Available at: https://www.heygen.com/.
4. Synthesia. Available at: https://www.synthesia.io/.

5. iFlyTek. Available at: https://iflytek.com.
6. Parts of Speech refer to the grammatical categories used to describe the function of a word in a sentence, including nouns, verbs, adjectives, and adverbs.
7. OtterAI. Available at: https://otter.ai/.
8. Plask. Available at: https://plask.ai/en-US.
9. MoveAI. Available at: https://move.ai/.
10. See Note 2.
11. DeepBrain. Available at: https://docs.aistudios.com/.
12. See Note 4.

Author contributions

CRediT: **Lai Wei**: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Validation, Writing – original draft, Writing – review & editing; **Sark Pangrui Xing**: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Visualization, Writing – original draft, Writing – review & editing; **Kenny K. N. Chow**: Funding acquisition, Project administration, Resources, Supervision; **Stephen Jia Wang**: Funding acquisition, Project administration, Resources, Supervision, Writing – review & editing.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This project is supported by the University Grants Committee (UGC) Funding Scheme [RHCE & G.73.xx.R006] from The Hong Kong Polytechnic University.

ORCID

Lai Wei http://orcid.org/0000-0002-5476-5450
Sark Pangrui Xing http://orcid.org/0000-0002-2273-4772
Kenny K. N. Chow http://orcid.org/0000-0002-8368-0157
Stephen Jia Wang http://orcid.org/0000-0001-9835-9932

References

Ahmed, B., Zada, S., Zhang, L., Sidiki, S. N., Contreras-Barraza, N., Vega-Muñoz, A., & Salazar-Sepúlveda, G. (2022). The impact of customer experience and customer engagement on behavioral intentions: Does competitive choices matters? *Frontiers in Psychology*, *13*(May 2022), 864841. https://doi.org/10.3389/fpsyg.2022.864841

Ali, G., Lee, M., & Hwang, J.-I. (2020). Automatic text-to-gesture rule generation for embodied conversational agents. *Computer Animation and Virtual Worlds*, *31*(4–5), e1944. https://doi.org/10.1002/cav.1944

Baylor, A. L., & Kim, S. (2009). Designing nonverbal communication for pedagogical agents: When less is more. *Computers in Human Behavior. Including the Special Issue: State of the Art Research into Cognitive Load Theory*, *25*(2), 450–457. https://doi.org/10.1016/j.chb.2008.10.008

Baylor, A. L., & Ryu, J. (2003). The effects of image and animation in enhancing pedagogical agent persona. *Journal of Educational Computing Research*, *28*(4), 373–394. https://doi.org/10.2190/V0WQ-NWGN-JB54-FAT4

Beege, M., Krieglstein, F., & Arnold, C. (2022). How instructors influence learning with instructional videos - the importance of professional appearance and communication. *Computers & Education*, *185*, 104531. https://doi.org/10.1016/j.compedu.2022.104531

Beege, M., Ninaus, M., Schneider, S., Nebel, S., Schlemmel, J., Weidenmüller, J., Moeller, K., & Rey, G. D. (2020). Investigating the effects of beat and deictic gestures of a lecturer in educational videos. *Computers & Education*, *156*, 103955. https://doi.org/10.1016/j.compedu.2020.103955

Belhiah, H. (2013). Using the hand to choreograph instruction: On the functional role of gesture in definition talk. *The Modern Language Journal*, *97*(2), 417–434. https://doi.org/10.1111/j.1540-4781.2013.12012.x

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: Analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Birmingham, C., Hu, Z., Mahajan, K., Reber, E., & Matarić, M. J. (2020). Can I trust you? A user study of robot mediation of a support group. In *2020 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 8019–8026). IEEE.

Bonfert, M., Zargham, N., Saade, F., Porzel, R., & Malaka, R. (2021). An evaluation of visual embodiment for voice assistants on smart displays. In *Proceedings of the 3rd Conference on Conversational User Interfaces*, New York, NY, USA (Article 16, pp. 1–11). Association for Computing Machinery. https://doi.org/10.1145/3469595.3469611

Brock, A. (2018). Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv: 1809.11096*

Burton, J. W., Stein, M.-K., & Jensen, T. B. (2020). A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making*, *33*(2), 220–239. https://doi.org/10.1002/bdm.2155

Carlgren, L., Rauth, I., & Elmquist, M. (2016). Framing design thinking: The concept in idea and enactment. *Creativity and Innovation Management*, *25*(1), 38–57. https://doi.org/10.1111/caim.12153

Cassell, J. (1998). A framework for gesture generation and interpretation. *Computer vision in human-machine interaction*., 191–215.

Castillo, S., Hahn, P., Legde, K., & Cunningham, D. W. (2018, November). *Personality analysis of embodied conversational agents*. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents, IVA 2018* (Vol. 2018, pp. 227–232). ACM. https://doi.org/10.1145/3267851.3267853

Castro-Alonso, J. C., Wong, R. M., Adesope, O. O., & Paas, F. (2021). Effectiveness of multimedia pedagogical agents predicted by diverse theories: A meta-analysis. *Educational Psychology Review*, *33*(3), 989–1015. https://doi.org/10.1007/s10648-020-09587-1

Ceha, J., & Law, E. (2022, April). Expressive auditory gestures in a voice-based pedagogical agent. In *Conference on Human Factors in Computing Systems - Proceedings*. ACM. https://doi.org/10.1145/3491102.3517599

Chen, B., Li, Y., Zheng, Y., Ding, Y.-X., & Zhou, K. (2025). Motion-example-controlled co-speech gesture generation leveraging large language models. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers (SIGGRAPH Conference Papers '25)*, New York, NY, USA (Article 55, pp. 1–12). Association for Computing Machinery. https://doi.org/10.1145/3721238.3730611

Cheng, Q., Li, X., & Fu, X. (2024). Siggesture: Generalized co-speech gesture synthesis via semantic injection with large-scale pre-training diffusion models. In *SIGGRAPH Asia 2024 Conference Papers (SA '24)*, New York, NY, USA (Article 133, pp. 1–11). https://doi.org/10.1145/3680528.3687677

Chow, K. K. N. (2026). What AI pretends to be? Design rhetoric and trust in the appearances of AI applications. *Pragmatics and Society*, Forthcoming.

Chui, K. (2005). Temporal patterning of speech and iconic gestures in conversational discourse. *Journal of Pragmatics*, *37*(6), 871–887. https://doi.org/10.1016/j.pragma.2004.10.016

Cienki, A., & Müller, C. (2008). Metaphor, gesture, and thought. *The Cambridge handbook of metaphor and thought* (483–501). Cambridge University Press.

Clarke, V., & Braun, V. (2017). Thematic analysis. *The Journal of Positive Psychology*, *12*(3), 297–298. https://doi.org/10.1080/17439760.2016.1262613

Cooper, R. G., & Sommer, A. F. (2016). The agile–stage-gate hybrid model: A promising new approach and a new research opportunity. *Journal of Product Innovation Management*, *33*(5), 513–526. https://doi.org/10.1111/jpim.12314

Craig, S. D., Twyford, J., Irigoyen, N., & Zipp, S. A. (2015). A test of spatial contiguity for virtual human's gestures in multimedia learning environments. *Journal of Educational Computing Research*, *53*(1), 3–14. doi:. https://doi.org/10.1177/0735633115585927

Dai, C.-P., Ke, F., Pan, Y., Moon, J., & Liu, Z. (2024). Effects of artificial intelligence-powered virtual agents on learning outcomes in computer-based simulations: A meta-analysis. *Educational Psychology Review*, *36*(1), 31. https://doi.org/10.1007/s10648-024-09855-4

Dai, L., Jung, M. M., Postma, M., & Louwerse, M. M. (2022). A systematic review of pedagogical agent research: Similarities, differences and unexplored aspects. *Computers & Education*, *190*, 104607. https://doi.org/10.1016/j.compedu.2022.104607

Davis, R. (2018). The impact of pedagogical agent gesturing in multimedia learning environments: A meta-analysis. *Educational Research Review*, *24*, 193–209. https://doi.org/10.1016/j.edurev.2018.05.002

Davis, R. O., & Vincent, J. (2019). Sometimes more is better: Agent gestures, procedural knowledge and the foreign language learner. *British Journal of Educational Technology*, *50*(6), 3252–3263. https://doi.org/10.1111/bjet.12732

Davis, R. O., Vincent, J., & Wan, L. (2021). Does a pedagogical agent's gesture frequency assist advanced foreign language users with learning declarative knowledge? *International Journal of Educational Technology in Higher Education*, *18*(1), 19. https://doi.org/10.1186/s41239-021-00256-z

Dhariwal, P., & Nichol, A. (2021). Diffusion models beat gans on image synthesis. In *Proceedings of the 35th International Conference on Neural Information Processing Systems (NIPS '21)*, Red Hook, NY, USA (Article 672, pp. 8780–8794). Curran Associates Inc.

Dick, W., Carey, L., & Carey, J. O. (2005). *The systematic design of instruction*. Pearson.

Druin, A. (2002). The role of children in the design of new technology. *Behaviour and Information Technology*, *21*(1), 1–25. https://doi.org/10.1080/01449290110108659

ELAN (2026). *Max Planck Institute for Psycholinguistics, The Language Archive* (Version 7.1) [Computer software]. Retrieved from https://archive.mpi.nl/tla/elan

Flesch, R. (1948). A new readability yardstick. *The Journal of Applied Psychology*, *32*(3), 221–233. https://doi.org/10.1037/h0057532

Freeman, J. L., & Curtis, A. N. (2023, April). Putting the self in self-tracking: The value of a co-designed 'how might you' self-tracking guide for teenagers. In *CHI '23. Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery. https://doi.org/10.1145/3544548.3580938

Gagne, R. (1985). *The conditions of learning and theory of instruction*. Holt, Rinehart & Winston.

Gao, W., Mei, Y., Duh, H., & Zhou, Z. (2025). Envisioning the incorporation of generative artificial intelligence into future product design education: Insights from practitioners, educators, and students. *The Design Journal*, *28*(2), 346–366. https://doi.org/10.1080/14606925.2024.2435703

Gay, G. (2002). Preparing for culturally responsive teaching. *Journal of Teacher Education*, *53*(2), 106–116. https://doi.org/10.1177/0022487102053002003

Ginosar, S., Bar, A., Kohavi, G., Chan, C., Owens, A., Malik, J. (2019). Learning individual styles of conversational gesture. https://arxiv.org/abs/1906.04160 arXiv: 1906.04160 [cs.CV].

Goldin-Meadow, S. (2015). Gesture as a window onto communicative abilities: Implications for diagnosis and intervention. *Perspectives on Language Learning and Education*, *22*(2), 50–60. pmid: 26366247. https://doi.org/10.1044/lle22.2.50

Guo, P. J., Kim, J., & Rubin, R. (2014). How video production affects student engagement: An empirical study of mooc videos. In *Proceedings of the First ACM Conference on Learning@ Scale Conference* (pp. 41–50). ACM.

Habibie, I., Xu, W., Mehta, D., Liu, L., Seidel, H.-P., Pons-Moll, G., Elgharib, M., & Theobalt, C. (2021). Learning speech-driven 3D conversational gestures from video. *In In Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents* (pp. 101–108). ACM. https://doi.org/10.1145/3472306.3478335

Hahn, P., Castillo, S., & Cunningham, D. W. (2018, November). Look me in the lines: The impact of stylization on the recognition of expressions and perceived personality. In *Proceedings of the 18th. In International Conference on Intelligent Virtual Agents, IVA 2018* (pp. 339–340). ACM. https://doi.org/10.1145/3267851.3267881

He, R., Wei, H., & Cao, Y. (2024). An interactive system for supporting creative exploration of cinematic composition designs. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology* (pp. 1–15). ACM.

Hewett, T. T., Baecker, R., Card, S., Carey, T., Gasen, J., Mantei, M., Perlman, G., Strong, G., & Verplank, W. (1992). *ACM SIGCHI curricula for human-computer interaction* [Paper presentation]. ACM.

Hollmén Larsen, A., & Zhu, J. (2024). *Ideary: Facilitating electronic music creation with generative AI*. In *Companion Publication of the 2024 ACM Designing Interactive Systems Conference* (pp. 275–278). ACM.

Höök, K., & Löwgren, J. (2012). Strong concepts: Intermediate-level knowledge in interaction design research. *ACM Transactions on Computer-Human Interaction*, *19*(3), 1–18. https://doi.org/10.1145/2362364.2362371

Ibrahim, M., Sweetser, P., & Ozdowska, A. (2023, April). Tutorial level design guidelines for 2D fighting games. In *FDG '23. Proceedings of the 18th International Conference on the Foundations of Digital Games*. Association for Computing Machinery. https://doi.org/10.1145/3582437.3582470

Ingold, T. (2013). *Making: Anthropology, archaeology, art and architecture*. Routledge. 176 pp. https://doi.org/10.4324/9780203559055

Saldaña, J. (2021). *The coding manual for qualitative researchers*. Sage.

Johnson, W. L., & Lester, J. C. (2016). Face-to-face interaction with pedagogical agents, twenty years later. *International Journal of Artificial Intelligence in Education*, *26*(1), 25–36. https://doi.org/10.1007/s40593-015-0065-9

Kersey, A. J., Carrazza, C., Novack, M. A., Congdon, E. L., Wakefield, E. M., Hemani-Lopez, N., & Goldin-Meadow, S. (2024). The effects of gesture and action training on the retention of math equivalence. *Frontiers in Psychology*, *15*, 1386187. https://doi.org/10.3389/fpsyg.2024.1386187

Kincaid, J. P., Fishburne, R. P., Jr., Rogers, R. L., & Chissom, B. S. (1975). Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Tech. rep.

Kizilkaya, G., & Askar, P. (2008). The effect of an embedded pedagogical agent on the students' science achievement. *Interactive Technology and Smart Education*, *5*(4), 208–216. https://doi.org/10.1108/17415650810930893

Kucherenko, T., Jonell, P., Van Waveren, S., Henter, G. E., Alexandersson, S., Leite, I., & Kjellström, H. (2020, October). *Gesticulator: A framework for semantically-aware speech-driven gesture generation*. In Proceedings of the 2020 International Conference on Multimodal Interaction (pp. 242–250). ACM. https://doi.org/10.1145/3382507.3418815

Kun, P., Freiberger, M. A., Sundnes Løvlie, A., & Risi, S. (2024). Genframe–embedding generative ai into interactive artifacts. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference* (pp. 714–727). ACM.

Lagerstrom, L., Johanes, P., & Ponsukcharoen, U. (2015). The myth of the six-minute rule: Student engagement with online videos. In 2015 ASEE Annual Conference & Exposition (pp. 26–1558).

Lawson, A. P., Mayer, R. E., Adamo-Villani, N., Benes, B., Lei, X., & Cheng, J. (2021). Recognizing the emotional state of human and virtual instructors. *Computers in Human Behavior*, *114*, 106554. https://doi.org/10.1016/j.chb.2020.106554

Lee, G., Deng, Z., Ma, S., Shiratori, T., Srinivasa, S. S., & Sheikh, Y. (2019). Talking with hands 16.2 m: A large-scale dataset of synchronized body-finger motion and audio for conversational motion analysis and synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 763–772). IEEE.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Lee, S., Cho, M., & Lee, S. (2020). What if conversational agents became invisible? comparing users' mental models according to physical entity of AI speaker. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *4*(3), 1–24. https://doi.org/10.1145/3411840

Li, W., Kuang, Z., Leng, X., Mayer, R. E., & Wang, F. (2024). Role of gesturing onscreen instructors in video lectures: A set of three-level meta-analyses on the embodiment effect. *Educational Psychology Review*, *36*(3), 67. https://doi.org/10.1007/s10648-024-09910-0

Li, W., Wang, F., Mayer, R. E., & Liu, H. (2019). Getting the point: Which kinds of gestures by pedagogical agents improve multimedia learning? *Journal of Educational Psychology*, *111*(8), 1382–1395. https://doi.org/10.1037/edu0000352

Li, W., Wang, F., Mayer, R. E., & Liu, T. (2022). Animated pedagogical agents enhance learning outcomes and brain activity during learning. *Journal of Computer Assisted Learning*, *38*(3), 621–637. https://doi.org/10.1111/jcal.12634

Lin, L., Ginns, P., Wang, T., & Zhang, P. (2020). Using a pedagogical agent to deliver conversational style instruction: What benefits can you obtain? *Computers & Education*, *143*, 103658. https://doi.org/10.1016/j.compedu.2019.103658

Liu, H., Zhu, Z., Iwamoto, N., Peng, Y., Li, Z., Zhou, Y., Bozkurt, E., & Zheng, B. (2022). *Beat: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis*. In European Conference on Computer Vision, In. Springer, 612–630.

Liu, S., Luo, Z., & Fu, W. (2025). Fcdnet: Fuzzy cognition-based dynamic fusion network for multimodal sentiment analysis. *IEEE Transactions on Fuzzy Systems*, *33*(1), 3–14. https://doi.org/10.1109/TFUZZ.2024.3407739

Liu, X., Wu, Q., Zhou, H., Du, Y., Wu, W., Lin, D., & Liu, Z. (2022). Audio-driven co-speech gesture video generation. In *Advances in neural information processing systems* (Vol. 35, pp. 21386–21399). Curran Associates, Inc. https://doi.org/10.48550/arxiv.2212.02350

Liu, Y., Zhang, K., Yuan Li, Z., Yan, C., Gao, R., Chen, Z. Y., Huang, Y., Sun, H., Gao, J., et al. (2024). Sora: A review on background, technology, limitations, and opportunities of large vision models. *arXiv preprint arXiv: 2402.17177*

Magistretti, S., Ardito, L., & Messeni Petruzzelli, A. (2021). Framing the microfoundations of design thinking as a dynamic capability for innovation: Reconciling theory and practice. *Journal of Product Innovation Management*, *38*(6), 645–667. https://doi.org/10.1111/jpim.12586

Manasrah, A., Masoud, M., & Jaradat, Y. (2021). Short videos, or long videos? A study on the ideal video length in online learning. In 2021 International Conference on Information Technology (ICIT) (pp. 366–370). IEEE.

Manesh, S. A., Zhang, T., Onishi, Y., Hara, K., Bateman, S., Li, J., & Tang, A. (2024). How people prompt generative AI to create interactive VR scenes. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference* (pp. 2319–2340). ACM.

Masson-Carro, I., Goudbeek, M., & Krahmer, E. (2017). How what we see and what we know influence iconic gesture production. *Journal of Nonverbal Behavior*, *41*(4), 367–394. https://doi.org/10.1007/s10919-017-0261-4

Mayer, R. E. (2002). Multimedia learning. In *Psychology of learning and motivation* (Vol. 41, 85–139). Academic Press. https://doi.org/10.1016/S0079-7421(02)80005-6

Mayer, R. E. (2005). Principles of multimedia learning based on social cues: Personalization, voice, and image principles. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 201–212). Cambridge University Press. https://doi.org/10.1017/CBO9780511816819.014

Mayer, R. E., & DaPra, C. S. (2012). An embodiment effect in computer-based learning with animated pedagogical agents. *Journal of Experimental Psychology. Applied*, *18*(3), 239–252. https://doi.org/10.1037/a0028616

McNeill, D. (2011). Hand and mind. In *Hand and mind* (pp. 351–374). De Gruyter Mouton. https://doi.org/10.1515/9783110874259.351

McNeill, D., & Levy, E. T. (1993). Cohesion and gesture. *Discourse Processes*, *16*(4), 363–386. https://doi.org/10.1080/01638539309544845

Mehrotra, S., Jorge, C. C., Jonker, C. M., & Tielman, M. L. (2024). Integrity-based explanations for fostering appropriate trust in ai agents. *ACM Transactions on Interactive Intelligent Systems*, *14*(1), 1–36. https://doi.org/10.1145/3610578

Merrill, M. D. (2002). First principles of instruction. *Educational Technology Research and Development*, *50*(3), 43–59. https://doi.org/10.1007/BF02505024

Micheli, P., Wilner, S. J. S., Bhatti, S. H., Mura, M., & Beverland, M. B. (2019). Doing design thinking: Conceptual review, synthesis, and research agenda. *Journal of Product Innovation Management*, *36*(2), 124–148. https://doi.org/10.1111/jpim.12466

Mildner, T., Cooney, O., Meck, A.-M., Bartl, M., Savino, G.-L., Doyle, P. R., Garaialde, D., Clark, L., Sloan, J., Wenig, N., et al. (2024). Listening to the voices: Describing ethical caveats of conversational user interfaces according to experts and frequent users. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–18.

Moon, J., & Ryu, J. (2021). The effects of social and cognitive cues on learning comprehension, eye-gaze pattern, and cognitive load in video instruction. *Journal of Computing in Higher Education*, *33*(1), 39–63. https://doi.org/10.1007/s12528-020-09255-x

Nakagawa, E., Sumiya, M., Koike, T., & Sadato, N. (2021). The neural network underpinning social feedback contingent upon one's action: An FMRI study. *NeuroImage*, *225*, 117476. https://doi.org/10.1016/j.neuroimage.2020.117476

Nyatsanga, S., Kucherenko, T., Ahuja, C., Henter, G. E., & Neff, M. (2023). A comprehensive review of data-driven co-speech gesture generation. *In Computer Graphics Forum*, *42*(2), 569–596. Wiley Online Library. https://doi.org/10.1111/cgf.14776

Ozmen Garibay, O., Winslow, B., Andolina, S., Antona, M., Bodenschatz, A., Coursaris, C., Falco, G., Fiore, S. M., Garibay, I., Grieman, K., Havens, J. C., Jirotka, M., Kacorri, H., Karwowski, W., Kider, J., Konstan, J., Koon, S., Lopez-Gonzalez, M., Maifeld-Carucci, I., … Xu, W.,. (2023). Six human-centered artificial intelligence grand challenges. *International Journal of Human–Computer Interaction*, *39*(3), 391–437. https://doi.org/10.1080/10447318.2022.2153320

Pang, H., Ding, T., He, L., Tao, M., Zhang, L., & Gan, Q. (2025). LLM gesticulator: Leveraging large language models for scalable and controllable co-speech gesture synthesis. In *Eighth International Conference on Computer Graphics and Virtuality (ICCGV 2025)* (Vol. 13557, 1355702). SPIE.

Park, G. W., Panda, P., Tankelevitch, L., & Rintel, S. (2024). The coexplorer technology probe: A generative AI-powered adaptive interface to support intentionality in planning and running video meetings. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference*, 1638–1657. ACM.

Pastore, R. (2012). The effects of time-compressed instruction and redundancy on learning and learners' perceptions of cognitive load. *Computers & Education*, *58*(1), 641–651. https://doi.org/10.1016/j.compedu.2011.09.018

Peng, X. B., Abbeel, P., Levine, S., & Van de Panne, M. (2018). Deepmimic: Example-guided deep reinforcement learning of physics-based character skills. *ACM Transactions on Graphics*, *37*(4), 1–14. https://doi.org/10.1145/3197517.3201311

Peng, X., Koch, J., & Mackay, W. E. (2024). Designprompt: Using multimodal interaction for design exploration with generative AI. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference*, 804–818. ACM.

Petersen, G. B., Mottelson, A., & Makransky, G. (2021, May). *Pedagogical agents in educational VR: An in the wild study*. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21) (pp. 1–12). Association for Computing Machinery. isbn: 978-1-4503-8096-6. https://doi.org/10.1145/3411764.3445760

Petridis, S., Terry, M., & Cai, C. J. (2024). Promptinfuser: How tightly coupling AI and UI design impacts designers' workflows. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference*, 743–756. ACM.

Pi, Z., Zhu, F., Zhang, Y., Chen, L., & Yang, J. (2022). Complexity of visual learning material moderates the effects of instructor's beat gestures and head nods in video lectures. *Learning and Instruction*, *77*, 101520. https://doi.org/10.1016/j.learninstruc.2021.101520

Reicherts, L., Rogers, Y., Capra, L., Wood, E., Duong, T. D., & Sebire, N. (2022). It's good to talk: A comparison of using voice versus screen-based interactions for agent-assisted tasks. *ACM Transactions on Computer-Human Interaction*, *29*(3), 1–41. https://doi.org/10.1145/3484221

Reigeluth, C. M. (1999). What is instructional-design theory and how is it changing. In *Instructional-design theories and models: A new paradigm of instructional theory* (Vol. 2, pp. 5–29). Routledge.

Sadoughi, N., & Busso, C. (2019). Speech-driven animation with meaningful behaviors. *Speech Communication*, *110*, 90–100. https://doi.org/10.1016/j.specom.2019.04.005

Savin-Baden, M., Tombs, G., & Bhakta, R. (2015). Beyond robotic wastelands of time: Abandoned pedagogical agents and new pedalled pedagogies. *E-Learning and Digital Media*, *12*(3–4), 295–314. https://doi.org/10.1177/2042753015571835

Schneider, S., Beege, M., Nebel, S., Schnaubert, L., & Rey, G. D. (2022). The cognitive-affective-social theory of learning in digital environments (CASTLE). *Educational Psychology Review*, *34*(1), 1–38. https://doi.org/10.1007/s10648-021-09626-5

Schneider, S., Krieglstein, F., Beege, M., & Rey, G. D. (2022). The impact of video lecturers' nonverbal communication on learning – an experiment on gestures and facial expressions of pedagogical agents. *Computers & Education*, *176*, 104350. https://doi.org/10.1016/j.compedu.2021.104350

Schon, D. A. (1992). Designing as reflective conversation with the materials of a design situation. *Research in Engineering Design*, *3*(3), 131–147. https://doi.org/10.1007/BF01580516

Schroeder, N. L., Davis, R. O., & Yang, E. (2025). Designing and learning with pedagogical agents: An umbrella review. *Journal of Educational Computing Research*, *62*(8), 1907–1936. https://doi.org/10.1177/07356331241288476

Seely Brown, J., Collins, A., & Duguid, P. (1989). *Situated Cognition and the Culture of Learning*, *18*(1), 32–42. https://doi.org/10.3102/0013189X018001032

Sekine, K., & Kita, S. (2015). Development of multimodal discourse comprehension: Cohesive use of space by gestures. *Language, Cognition and Neuroscience*, *30*(10), 1245–1258. https://doi.org/10.1080/23273798.2015.1053814

Serras Pereira, M., de Lange, J., Shahid, S., & Swerts, M. (2018). A perceptual and behavioral analysis of facial cues to deception in interactions between children and a virtual agent. *International Journal of Child-Computer Interaction*, *15*, 1–12. https://doi.org/10.1016/j.ijcci.2017.10.003

Shatilov, K., Alhilal, A., Braud, T., Lee, L.-H., Zhou, P., & Hui, P. (2023, June). Players are not ready 101: A tutorial on organising mixed-mode events in the metaverse. In *MetaSys '23. Proceedings of the First Workshop on Metaverse Systems and Applications* (pp. 14–20). Association for Computing Machinery. https://doi.org/10.1145/3597063.3597360

Shaw, K., Bahl, S., Sivakumar, A., Kannan, A., & Pathak, D. (2024). Learning dexterity from human hand motion in internet videos. *The International Journal of Robotics Research*, *43*(4), 513–532. https://doi.org/10.1177/02783649241227559

Shiban, Y., Schelhorn, I., Jobst, V., Hörnlein, A., Puppe, F., Pauli, P., & Mühlberger, A. (2015). The appearance effect: Influences of virtual agent features on performance and motivation. *Computers in Human Behavior*, *49*, 5–11. https://doi.org/10.1016/j.chb.2015.01.077

Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction*, *36*(6), 495–504. https://doi.org/10.1080/10447318.2020.1741118

Sinatra, A. M., Pollard, K. A., Files, B. T., Oiknine, A. H., Ericson, M., & Khooshabeh, P. (2021). Social fidelity in virtual agents: Impacts on presence and learning. *Computers in Human Behavior*, *114*, 106562. https://doi.org/10.1016/j.chb.2020.106562

So, W. C., Sim Chen-Hui, C., & Low Wei-Shan, J. (2012). Mnemonic effect of iconic gesture and beat gesture in adults and children: Is meaning in gesture important for memory recall? *Language and Cognitive Processes*, *27*(5), 665–681. https://doi.org/10.1080/01690965.2011.573220

Tholander, J., & Jonsson, M. (2023). Design ideation with AI-sketching, thinking and talking with generative machine learning models. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference* (pp. 1930–1940). ACM.

Thomas, S., Ferstl, Y., McDonnell, R., & Ennis, C. (2022, March). *Investigating how speech and animation realism influence the perceived personality of virtual characters and agents*. In Proceedings - 2022 IEEE Conference on Virtual Reality and 3D User Interfaces, VR 2022 (pp. 11–20). IEEE. https://doi.org/10.1109/VR51125.2022.00018

Tilekbay, B., Yang, S., Lewkowicz, M. A., Suryapranata, A., & Kim, J. (2024). Expressedit: Video editing with natural language and sketching. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, 515–536.

Tsai, W.-L., Su, L-w., Ko, T.-Y., Yang, C.-T., & Hu, M.-C. (2019, March). *Improve the decision-making skill of basketball players by an action-aware VR training system*. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)* (pp. 1193–1194). IEEE. https://doi.org/10.1109/VR.2019.8798309

Uusitalo, S., Salovaara, A., Jokela, T., & Salmimaa, M. (2024). "Clay to play with": Generative AI tools in UX and industrial design practice. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference* (pp. 1566–1578). ACM.

Wang, I., & Ruiz, J. (2021). Examining the use of nonverbal communication in virtual agents. *International Journal of Human–Computer Interaction*, *37*(17), 1648–1673. https://doi.org/10.1080/10447318.2021.1898851

Wei, L., & Chow K. K. N. (2023). When gestures and words synchronize: Exploring a human lecturer's multimodal interaction for the design of embodied pedagogical agents. In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing (CSCW '23)*. Association for Computing Machinery.

Wei, L., & Chow, K. K. N. (2022). Who shapes the network of a pedagogical space? Clues from the movements in the physical places. In *Advances in Mobile Computing and Multimedia Intelligence: 20th International Conference, MoMM 2022, Virtual Event, November 28–30, 2022, Proceedings* (pp. 143–153). Springer-Verlag. https://doi.org/10.1007/978-3-031-20436-4_14

Wei, L., & Chow, K. K. N. (2023, October 9–13). How students perceive lecturers' gestures? An exploration in gesture-meaning matching toward embodied pedagogical agent design. In D. De Sainz Molestina, L. Galluzzo, F. Rizzo, D. Spallazzo (Eds.), *IASDR 2023: Life-Changing Design*, Milan, Italy. https://doi.org/10.21606/iasdr.2023.115

Weisz, J. D., He, J., Muller, M., Hoefer, G., Miles, R., & Geyer, W. (2024). Design principles for generative ai applications. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, New York, NY, USA (pp. 1–22). Association for Computing Machinery. https://doi.org/10.1145/3613904.3642466

Wiberg, M. (2014). Methodology for materiality: Interaction design research through a material lens. *Personal and Ubiquitous Computing*, *18*(3), 625–636. https://doi.org/10.1007/s00779-013-0686-7

Wilcox, L., DiSalvo, B., Henneman, D., & Wang, Q. (2019). Design in the HCI classroom: Setting a research agenda. In *Proceedings of the 2019 on Designing Interactive Systems Conference*, New York, NY, USA (pp. 871–883). Association for Computing Machinery. https://doi.org/10.1145/3322276.3322381

Wiley, D., & Hilton, J. III. (2009). Openness, dynamic specialization, and the disaggregated future of higher education. *The International Review of Research in Open and Distributed Learning*, *10*(5). https://doi.org/10.19173/irrodl.v10i5.768

Williams, J. R. (1998). Guidelines for the use of multimedia in instruction. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *42*(20), 1447–1451. https://doi.org/10.1177/154193129804202019

Wolfert, P., Robinson, N., & Belpaeme, T. (2022). A review of evaluation practices of gesture generation in embodied conversational agents. *IEEE Transactions on Human-Machine Systems*, *52*(3), 379–389. https://doi.org/10.1109/THMS.2022.3149173

Wu, C., Yin, S., Qi, W., Wang, X., Tang, Z., & Duan, N. (2023). Visual chatgpt: talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*

Wu, Q., Wu, C.-J., Zhu, Y., & Joo, J. (2021, September). Communicative learning with natural gestures for embodied navigation agents with human-in-the-scene. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 4095–4102). IEEE. https://doi.org/10.1109/IROS51168.2021.9636208

Xing, S. P., Van Dijk, B., An, P., Bruns, M., Chuang, Y., & Wang, S. J. (2023, February). Puffy: A step-by-step guide to craft bio-inspired artifacts with interactive materiality. In *TEI '23. Proceedings of the Seventeenth International Conference on Tangible, Embedded, and Embodied Interaction*. Association for Computing Machinery. https://doi.org/10.1145/3569009.3572800

Xu, W. (2019). Toward human-centered ai: A perspective from human-computer interaction. *Interactions*, *26*(4), 42–46. https://doi.org/10.1145/3328485

Xu, W., & Gao, Z. (2023). Enabling human-centered ai: a methodological perspective. *arXiv preprint arXiv: 2311.06703*

Xu, Z., Zhou, Y., Kalogerakis, E., Landreth, C., & Singh, K. (2020). Rignet: neural rigging for articulated characters. *arXiv preprint arXiv:2005.00559*

Yang, S., Wu, Z., Li, M., Zhang, Z., Hao, L., Bao, W., Cheng, M., & Xiao, L. (2023). Diffusestylegesture: stylized audio-driven co-speech gesture generation with diffusion models. *arXiv preprint arXiv:2305.04919*

Yi, H., Liang, H., Liu, Y., Cao, Q., Wen, Y., Bolkart, T., Tao, D., & Black, M. J. (2023). Generating holistic 3D human motion from speech. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 469–480).

Yoon, Y., Cha, B., Lee, J.-H., Jang, M., Lee, J., Kim, J., & Lee, G. (2020). Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Transactions on Graphics*, *39*(6), 1–16. https://doi.org/10.1145/3414685.3417838

Yoon, Y., Ko, W.-R., Jang, M., Lee, J., Kim, J., & Lee, G. (2019, May). *Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots*. In 2019 International Conference on Robotics and Automation (ICRA) (pp. 4303–4309). IEEE. https://doi.org/10.1109/ICRA.2019.8793720

Yoon, Y., Park, K., Jang, M., Kim, J., & Lee, G. (2021, October). Sgtoolkit: An interactive gesture authoring toolkit for embodied conversational agents. In The 34th Annual ACM Symposium on User Interface Software and Technology (pp. 826–840). Association for Computing Machinery. https://doi.org/10.1145/3472749.3474789

Yoon, Y., Wolfert, P., Kucherenko, T., Viegas, C., Nikolov, T., Tsakov, M., & Henter, G. E. (2022). The GENEA challenge 2022: A large evaluation of data-driven co-speech gesture generation. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, 736–747. Association for Computing Machinery.

Zhang, H., Chen, P., Xie, X., Lin, C., Liu, L., Li, Z., You, W., & Sun, L. (2024). Protodreamer: A mixed-prototype tool combining physical model and generative ai to support conceptual design. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, 1–18. Association for Computing Machinery.

Zhang, H., Liu, Y., Jiang, M., Chen, J., Wang, M., & Paas, F. (2025). Emotional artificial intelligence in education: A systematic review and meta-analysis. *Educational Psychology Review*, *37*(4), 106. https://doi.org/10.1007/s10648-025-10086-4

Zhang, S., Li, Y., Gan, G., Pang, S., & Kim, J. H. 2025. Effects of social cues of artificial intelligence-powered pedagogical agents: A multilevel meta-analysis. *Educational Research Review*, *49*, 100746. https://doi.org/10.1016/j.edurev.2025.100746

Zhang, T., Zhang, Y., Vineet, V., Joshi, N., & Wang, X. (2023). Controllable text-to-image generation with GPT-4. *arXiv preprint arXiv:2305.18583*

Zhang, Z., Ao, T., Zhang, Y., Gao, Q., Lin, C., Chen, B., & Liu, L. (2024). Semantic gesticulator: Semantics-aware co-speech gesture synthesis. *ACM Transactions on Graphics*, *43*(4), 1–17. https://doi.org/10.1145/3658134

Zhi, Y., Cun, X., Chen, X., Shen, X., Guo, W., Huang, S., & Gao, S. (2023). Livelyspeaker: Towards semantic-aware co-speech gesture generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 20807–20817.

Zhou, J., Li, R., Tang, J., Tang, T., Li, H., Cui, W., & Wu, Y. (2024). Understanding nonlinear collaboration between human and ai agents: A co-design framework for creative design. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, New York, NY, USA (pp. 1–16). Association for Computing Machinery. https://doi.org/10.1145/3613904.3642812

Zhu, D., Chen, J., Shen, X., Li, X., & Elhoseiny, M. (2023). Minigpt-4: enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*

Zhu, Z., Wang, X., Zhao, W., Min, C., Li, B., Deng, N., Dou, M., Wang, Y., Shi, B., Wang, K., et al. (2024). Is sora a world simulator? a comprehensive survey on general world models and beyond. *arXiv preprint arXiv: 2405.03520*

Zimmerman, J., Forlizzi, J., & Evenson, S. (2007, April). *Research through design as a method for interaction design research in HCI*. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 493–502). ACM. https://doi.org/10.1145/1240624.1240704

About the authors

Lai Wei is an animation designer and postdoctoral researcher at the Brain, Language, and Computation Lab, Department of Language Science and Technology, PolyU. Her research focuses on gesture design, multimodal interaction, and virtual learning environments for education. She holds BA and MA degrees in Animation and a PhD in Design.

Sark Pangrui Xing is an interaction designer and doctoral researcher at The Hong Kong Polytechnic University. His expertise lies in manipulating material properties to craft novel interactive experiences for daily activities in varied everyday contexts. He holds a Bachelor of Engineering and a Master of Science degree in Industrial Design.

Kenny K. N. Chow is currently Head and Associate Professor in the Department of Interactive Media, Hong Kong Baptist University. His research covers interaction design, games, and animation. His books include Animation, Embodiment, and Digital Media: Human Experience of Technological Liveliness (Palgrave Macmillan) and Expressive Iteration: Designing for Meaningful Routines (Routledge).

Stephen Jia Wang is Full Professor in UX Design and Design Intelligence at PolyU's School of Design, Scheme Leader of the Master of Design, and Distinguished Visiting Professor at Tsinghua University's Future Lab. His work advances innovative design practice through research in interaction design, industrial design, and human-centred intelligent systems.