Does Personalized Nudging Wear Off? A Longitudinal Study of AI Self-Modeling for Behavioral Engagement

Qing He*
Weitzman School of Design
University of Pennsylvania
Philadelphia, Pennsylvania, USA
Key Laboratory of Pervasive Computing, Ministry of Education,
Department of Computer Science and Technology
Tsinghua University
Beijing, China
qingh@upenn.edu

Zeyu Wang*
Department of Computer Science and Technology, Beijing National Research Center for Information Science and Technology (BNRist)
Tsinghua University
Beijing, China
wang-zy23@mails.tsinghua.edu.cn

Yuzhou Du
Grado Department of Industrial and Systems Engineering
Virginia Tech
Blacksburg, Virginia, USA
daviddu@vt.edu

Jiahuan Ding
Joint School of Design and Innovation
Xi'an Jiaotong University
Xi'an, China
dingjh@stu.xjtu.edu.cn

Yuanchun Shi
Key Laboratory of Pervasive Computing, Ministry of Education,
Department of Computer Science and Technology, Beijing National Research Center for Information Science and Technology (BNRist)
Tsinghua University
Beijing, China
Qinghai University
Xining, Qinghai, China
shiyc@tsinghua.edu.cn

Yuntao Wang†
Department of Computer Science and Technology
Tsinghua University
Beijing, China
School of Computer Technology and Application
Qinghai University
Xining, Qinghai, China
yuntaowang@tsinghua.edu.cn

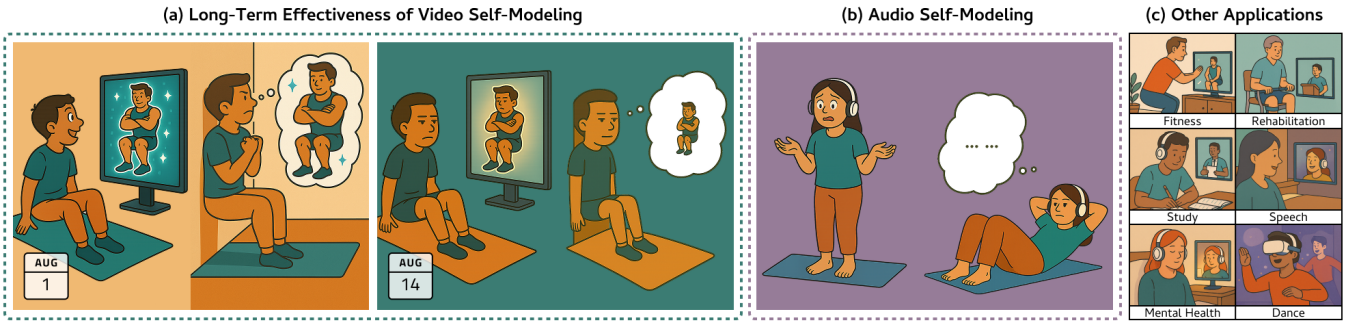


Figure 1: The nudging effect of AI self-modeling across time and modalities in fitness. (a) Change of nudging effect of video self-modeling(VSM): users are motivated by the AI self-model at first, then grow numb to the repeating VSM intervention, yet the motivated self-model has internalized as their goal. (b) non-significant benefit of audio self-modeling (ASM) for fitness; (c) broader applications in other domains such as rehabilitation or learning.

Abstract

Sustaining the effectiveness of behavior change technologies remains a key challenge. AI self-modeling, which generates personalized portrayals of one's ideal self, has shown promise for motivating

*Both authors contributed equally to this research.
†Corresponding Author.

CHI '26, Barcelona, Spain

ACM ISBN 979-8-4007-2278-3/26/04
https://doi.org/10.1145/3772318.3791777

behavior change, yet prior work largely examines short-term effects. We present one of the first longitudinal evaluations of AI self-modeling in fitness engagement through a two-stage empirical study. A 1-week, three-arm experiment (visual self-modeling (VSM), auditory self-modeling (ASM), Control; N=28) revealed that VSM drove initial performance gains, while ASM showed no significant effects. A subsequent 4-week study (VSM vs. Control; N=31) demonstrated that VSM sustained higher performance levels but exhibited diminishing improvement rates after two weeks. Interviews uncovered a catalyst effect that fostered early motivation through clear, attainable goals, followed by habituation and internalization which stabilized performance. These findings highlight the temporal dynamics of personalized nudging and inform the design of behavior change technologies for long-term engagement.

CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**; **HCI theory, concepts and models**; • **Applied computing** → *Health informatics.*

Keywords

AI Self-Modeling, Longitudinal Study, Nudging, Behavior Change Technologies

ACM Reference Format:

Qing He, Zeyu Wang, Yuzhou Du, Jiahuan Ding, Yuanchun Shi, and Yuntao Wang. 2026. Does Personalized Nudging Wear Off? A Longitudinal Study of AI Self-Modeling for Behavioral Engagement. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 20 pages. https://doi.org/10.1145/3772318.3791777

1 Introduction

Long-term behavior change remains notoriously difficult. Even with the aid of personal informatics and behavior tracking tools, many users abandon their efforts after only a short time [29, 37, 62]. In response, researchers have looked toward **nudging** [38], a behavioral science approach that subtly guides individuals towards desired behaviors without restricting their choices [10, 26]. However, traditional technology-mediated nudges are relatively inflexible, with limited adaptation to individual needs and contexts [11, 55]. With recent advances in artificial intelligence (AI), new opportunities have emerged for performing more interactive, personalized, and context-aware interventions [34, 45, 51, 66]. A promising concept is **AI self-modeling**, which uses AI technologies to generate realistic, individualized simulations of a user's future self. By making long-term consequences vivid and personally relevant, it has the potential to sustain engagement, reinforce self-efficacy, and improve performance over time.

Early studies offer encouraging evidence across modalities. FakeForward, for instance, utilizes AI to swap faces onto a peer model with better performances in specific tasks [13]. Voice-based approaches, such as Emotional Self-Voice (ESV) and Mirai, use voice cloning technology to generate personalized and context-related audio responses [23, 24]. These self-referential representations demonstrate the potential of AI self-modeling to boost motivation and performance. However, prior evaluations primarily focus on short-term effects, leaving open whether such benefits can be sustained in everyday contexts.

Psychological theories suggest that sustaining such effects over time is challenging. Habituation theory [54, 65] suggests that repeated exposure reduces novelty and diminishes impact. In practice, fitness and health behavior change technologies (BCTs) face early attrition and engagement decay, driven by novelty decay, monitoring burden, lack of personalization, variety, and ongoing support [29, 33, 63]. Affective–Reflective Theory also indicates that exercise decisions are often affect-driven in the short term, but sustained adherence requires reflective attitudes [6, 7]. For AI self-modeling, which provides deeply self-referential, identity-linked feedback, engaging mechanisms distinct from generic prompts, its capacity to sustain nudging effects over time remains an important question.

To evaluate the pattern of long-term impact of AI self-modeling, our research focuses on fitness as a meaningful testbed whose objective performance and subjective experience are both measurable. Following previous works, we examine two representative AI self-modeling modalities: Video Self-Modeling (VSM) [13] and Audio Self-Modeling (ASM) [23, 24], providing an ideal self for participants during daily exercise. These implementations reflect common ways that self-modeling has been operationalized in practice, allowing us to situate AI-generated self-representations within everyday fitness routines. To investigate their long-term effects, our study is guided by the following research questions:

- **RQ1** Does AI self-modeling sustain performance over time in everyday fitness practice?
- **RQ2** Does AI self-modeling sustain improvement rate over time in everyday fitness practice?
- **RQ3** What design factors influence the long-term impact of AI self-modeling, and how can these insights inform future behavior change technologies?

We first conducted a one-week exploratory study (N=9,10,9) to evaluate both AI self-modeling techniques. The initial results showed that ASM failed to boost fitness performance, while VSM demonstrated effectiveness for one week. Based on the results, we expand the experiment period to four weeks for the VSM and the Control group (N=31). Our final results demonstrated that VSM could successfully sustain performance for four weeks, whereas it failed to sustain the improvement rate and showed a converging trend after two weeks. Further analysis of long-term subjective feedback revealed habituation to the continuous VSM intervention. Through semi-structured interviews with the participants, we discussed reasons for the results and provided in-depth design implications for future long-term nudging systems.

We claim the following novel contributions of our work:

- We present the first 28-day long-term study of AI self-modeling with 31 participants, extending prior work [13, 23] limited to short-term or one-off effects and revealing a stage-dependent trajectory, early acceleration followed by stabilization, which offers a more nuanced pattern than simple habituation decay.

- Our studies offer design implications for long-term behavior change technologies, reflecting how modality, personalization, and subjective experiences are crucial for designing interventions with lasting impact.

2 Background and Related Works

This section establishes the theoretical foundation for our work. We begin with the challenge of long-term behavior change and motivational decline, then introduce AI self-modeling and its theoretical potential to counteract such declines. Finally, we situate these mechanisms within the fitness context, which serves as the longitudinal testbed for our study.

2.1 Long-Term Behavior Change and Motivational Decline

Prior works in HCI have repeatedly highlighted the challenge of sustaining engagement with behavior change technologies [29, 37, 62]. While mobile and wearable devices often spark strong initial interest, sustained use proves elusive: studies show that many users discontinue within weeks or months [68]. Research has shown that users often abandon devices once the novelty fades, when self-tracking becomes burdensome, or when the collected data no longer feels meaningful [14]. Likewise, longitudinal studies emphasize that while BCTs such as goal-setting, feedback, or self-monitoring can spark short-term motivation, their benefits often taper off with repeated use [29].

To further explain this decline, researchers emphasize motivational deterioration as the key driver of disengagement [9, 15, 48]. The Athlete Burnout model, for example, conceptualizes this process across three dimensions: emotional and physical exhaustion, reduced sense of accomplishment, and sport devaluation [52, 53], with the latter two being especially predictive of dropout [46]. Complementary perspectives point to mechanisms of attentional and affective decline: Habituation theory suggests that repeated exposure to the same stimulus gradually reduces its motivational salience [65], while Attention Restoration Theory argues that prolonged repetitive engagement depletes attentional resources and undermines motivation [35]. Affective–Reflective Theory emphasizes the relation between short-term affective impulses and longer-term reflective attitudes [6, 7]. In digital and exercise contexts, these processes manifest as fading novelty effects and flattened engagement curves.

These empirical findings and psychological theories reveal that declining engagement is not a superficial issue but a fundamental challenge of human motivation and attention. However, while these theories explain why motivation decays, they provide limited guidance on how technology might sustain motivation across repeated daily exposures. This motivates our examination of whether AI self-modeling, which is augmented with identity cues and novelty, can counteract such decline.

2.2 AI Self-Modeling as Personalized Nudging for Behavior Change

2.2.1 Existing Works on AI Self-Modeling. Nudging, a strategy for influencing behavior without restricting choice [10, 26, 38], has become a prominent approach in HCI. However, traditional technology-mediated nudges are often static and insensitive to individual differences [11, 55]. Generative AI now enables *AI self-modeling*, which produces individualized portrayals of one's future or ideal self to deliver identity-based nudges.

Visual approaches synthesize personalized images or videos that depict users (or lookalikes) successfully executing target behaviors, such as aged future selves [32], or increasing attainability perceptions through deepfaked peer demonstrations [13]. Audio-based approaches modified self-voices can alter emotions [16] and nudge people to achieve daily goals [36], cloned voices can deliver motivational utterances [23], and wearable "inner-voice" systems provide context-aware interventions [24]. Such systems primarily operate through verbal persuasion and affective modulation, which may influence motivation differently from visual interventions. Conversational future-self agents also combine visual and audio modalities to reduce anxiety and strengthen identification with one's future self [49]. Across these modalities, the common principle is aligning intervention content with a user's identity signals ("this is me") to strengthen intention, self-efficacy, and adherence.

Despite these technological advances, empirical evaluations have been largely confined to short-term or one-off effects. Consequently, the longitudinal durability of these AI-driven interventions in everyday contexts remains underexplored.

2.2.2 Theories and Principles Behind Self-modeling. Self-modeling draws from multiple psychological frameworks, but its foundational mechanism is best described as feedforward, or "learning from the future" [20, 22]. Unlike traditional feedback, which corrects past errors, feedforward creates a memory of success that has not yet occurred. AI operationalizes this by synthesizing "future-perfect" representations, allowing users to sense their own potential. This feedforward mechanism influences long-term behavior through two complementary pathways: a capability pathway that shifts self-efficacy (how capable one feels), and an identity pathway that shifts the salience and relevance of valued future selves (who one feels they are becoming).

Self-Referential Modeling Improves Self-Efficacy. Within Social Cognitive Theory, self-efficacy is shaped by multiple sources, including vicarious experience, verbal persuasion, and affective states [1]. Self-modeling can be understood as a form of *self-referential modeling*, where curated representations of one's own performance act as symbolic models that inform what one believes they can achieve [20–22].

Identity-Based Motivation and Possible Selves. Beyond perceived capability, identity-based motivation [47] highlights that people are motivated to act in ways that are congruent with "people like me," and that they are more likely to interpret effortful actions as meaningful when these actions fit a valued identity. Possible Selves Theory [40] further describes how mental representations of hoped-for and feared future selves provide a cognitive bridge between present choices and long-term outcomes. When future selves are vivid, plausible, and experientially close, they render long-term goals personally relevant and worthy of effort.

Compared to traditional methods, AI self-modeling theoretically enhances these pathways by rendering future selves with

greater vividness (strengthening self-efficacy) and dynamic coherence (preserving identity fit). At the same time, AI allows these self-representations to be flexibly updated to match a person's current progress and goals, preserving a tight fit between the modeled capabilities and the identity that feels attainable and self-relevant. In this way, AI self-modeling has the potential to counteract habituation and sustain the vividness required for the identity pathway, which is under-explored in previous studies.

2.3 Fitness and Health as Context for Behavior Change

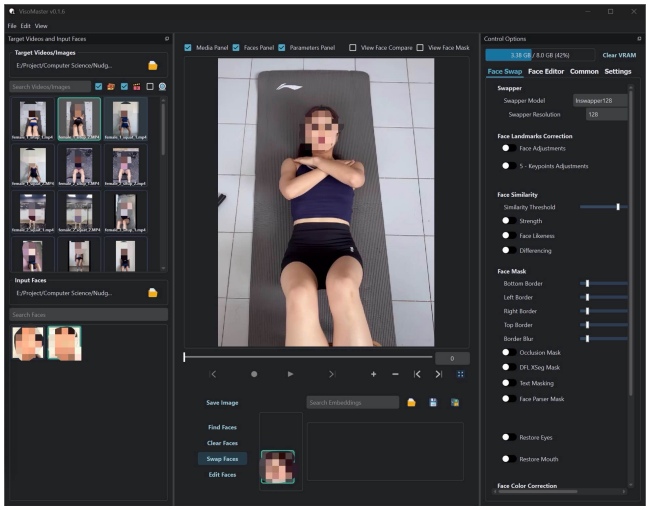
To investigate these longitudinal dynamics in a motivationally demanding environment, we situate our study in fitness and health. Fitness tasks inherently demand repeated and sustained effort, making them an ideal setting for studying long-term motivational dynamics. Prior work has leveraged fitness as a natural testbed: Gouveia et al. [29] conducted a ten-month field study of an activity tracker and found that only 38% of participants continued after the first week, with most dropping out within two weeks.

Fitness and health also offer a methodological advantage: it enables systematic evaluation of both *objective performance outcomes* (e.g., exercise completion, physical improvement) and *subjective motivational experiences*. For example, Choe et al. [12] showed how self-trackers reflect not only on numerical performance but also on motivational struggles in everyday practice. Mazeas et al. [41] used accelerometers, pedometers, and self-report questionnaires to collect subjective and objective data, measuring the outcome of gamification on physical activity.

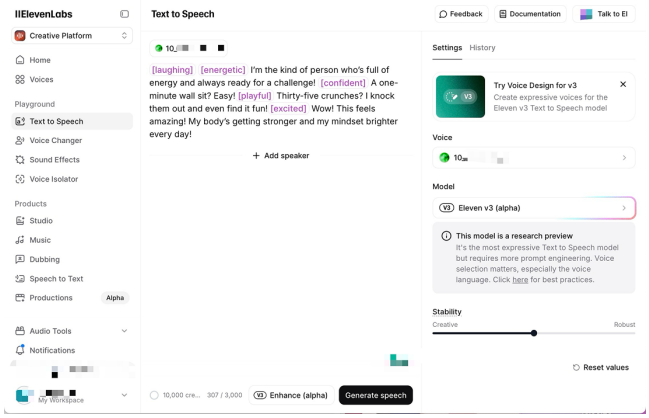
Crucially, physical training naturally accommodates the two dominant forms of AI self-modeling: visual and auditory interventions. In exercise contexts, visual cues are intrinsic to motor learning and form calibration (e.g., video feedback [43]), while auditory cues are fundamental for affective regulation and persistence (e.g., rhythm and self-talk [5, 30, 31]). Consequently, fitness provides an ecologically valid environment to implement both VSM and ASM as natural, representative instances of the AI self-modeling paradigm. This allows us to examine the longitudinal efficacy of the AI self-modeling concept broadly, rather than being limited to a single modality.

3 AI Self-Modeling Nudging System Implementation

Based on the discussion in Section 2, behavior change technologies (BCTs) suffer from habituation during long-term usage. The emergence of AI self-modeling offers greater vividness and flexibility to strengthen self-efficacy and identity-based motivation. Our work aims to investigate whether these unique properties of AI self-modeling have the potential to alleviate longitudinal motivation decay. Guided by this rationale, we implemented our system by reproducing and extending the pipelines from [13, 23], deploying them for the first time in a long-term study. The system consists of two components: **Video Self-Modeling (VSM)** and **Audio Self-Modeling (ASM)**, which are the mainstream strategies in AI self-modeling. In this section, we first describe our implementation for VSM and ASM, followed by the key modifications in our implementation compared to the original pipeline [13, 23].



(a) Video Self-Modeling



(b) Audio Self-Modeling

Figure 2: System implementation for AI self-modeling. Interfaces are adapted from existing platforms and included for illustration only.

3.1 Video Self-Modeling Pipeline

Our video self-modeling pipeline builds on the FakeForward framework [13], which produces AI-edited future-self videos by swapping a participant's face onto a better-performing peer model. The key components of our implementation involve three stages:

3.1.1 Peer Model Preparation. In this stage, we prepared high-quality peer model videos as face-swapping sources. Six participants (3 female, 3 male), each representing a different ideal body type, were recruited to demonstrate superior form and performance in target exercises. The recordings featured consistent lighting and camera angles to ensure seamless swapping, while varied clothing and backgrounds for each peer model enhanced resemblance and engagement.

3.1.2 Headshot Capture and Personalization. In this stage, the participant is guided to capture a headshot and select an ideal body

type. Participants should ensure their face is clearly visible—avoid wearing glasses, having bangs or hair covering the face, and wearing makeup, to improve the quality of face-swapping.

3.1.3 Face-swapping and Video Generation. In the final stage, the participant's face is swapped onto the peer model's body using the open-source VisoMaster framework[1], with built-in face detection and Inswapper128 as the swapper model (Figure 2a). During face-swapping, facial movements were synchronized with the peer model's actions for a seamless effect, and videos were kept concise (30–40 seconds) to sustain attention.

3.2 Audio Self-Modeling Pipeline

The audio self-modeling pipeline adapts the ESV's approach [23], which provides contextually aware interventions to support personalized goals. We adapt this method into the following steps:

3.2.1 Voice Cloning. Participants record a short audio clip by reading a script with colloquial happy and sad sentences in the local language (see Appendix A.1). Recordings were conducted in a quiet environment and then uploaded to ElevenLabs' voice cloning service[2]. Participants were instructed to speak clearly at a moderate pace to ensure both intelligibility and emotional nuance were conveyed.

3.2.2 Response Generation. To adapt to participants' daily goals, which can be influenced by cognition and emotions, this stage generates motivational responses based on the participant's current context, which is captured through a daily Audio Generation Questionnaire (namely AGQ, see Appendix B.2), which contains the same fields as ESV [23] except for content choice. Due to the lack of API support at the time of the study, we did not re-implement its UI as the original pipeline did. Instead, we leveraged Azure GPT-4o's API [3] for generation and followed the prompt structure in [23]. The prompt design (Appendix A.2) ensured that generated messages reliably incorporated two validated psychological components: (1) positive self-talk from sports psychology [30, 31], providing short directive cues known to enhance persistence (e.g., "you can hold this pace"); and (2) self-efficacy reinforcement grounded in social-cognitive theory [1], supporting participants' belief in their ability to complete the exercise.

3.2.3 Audio Generation. In this final stage, we leveraged ElevenLabs' V3 model to generate the motivational audio with emotions [4]. ElevenLabs' V3 can generate emotional audio controlled by open-vocabulary emotion description and special tokens. We generate four audio clips with emotional enhancement and choose the one that preserves the highest human-like quality and emotional expressiveness.

3.3 Adjustment to Original Pipeline

To preserve comparability, we adhered closely to the original pipelines, matching their data preparation, prompt/script templates, and content scheduling, and introduced only two pragmatic substitutions to fit our study setting. For VSM, we replaced the old face-swapping backbone with a state-of-the-art model to improve identity fidelity and temporal coherence. For ASM, we adopted a speech-synthesis pipeline with stronger support for our local language to enhance naturalness and intelligibility. To measure the effectiveness of the self-modeling methods rather than the impact of specific content, we also introduced variation: VSM peer models prepared multiple outfits during filming, and ASM participants completed the AGQ daily to generate diverse audio clips. These changes updated the tooling while keeping the experimental structure and stimulus intent intact.

[1]https://github.com/visomaster/VisoMaster
[2]https://elevenlabs.io/voice-cloning
[3]https://learn.microsoft.com/en-us/azure/ai-foundry/
[4]Eleven v3: https://elevenlabs.io/text-to-speech

Table 1: Demographic information of participants across three groups.

Group	Gender (F/M)	Total	BMI (M±SD)
VSM	5/4	9	22.88±2.82
ASM	5/5	10	22.75±3.25
Control	6/3	9	23.11±2.91
Total	16/12	28	22.91±2.90

3.4 Ethical Concerns

We treated identity editing and voice cloning as sensitive interventions and implemented safeguards addressing consent, privacy, fairness, transparency, and well-being in a unified protocol. All participants and peer models were provided explicit written consent in advance, specifying permissible uses (study-only, non-commercial), storage duration, etc. This research was reviewed and approved by the local Institutional Review Board.

4 Study 1: Exploratory Validation of AI Self-Modeling

We first conducted a one-week study to evaluate the effectiveness of video and audio AI self-modeling in our fitness context. The goal was to identify which implementation reliably produces measurable effects and could be extended to the subsequent long-term study. In addition, we examined whether any nudging effects would diminish over time, as suggested by prior work on novelty decay and habituation.

4.1 Study Design

We designed a between-subjects experiment comparing the effects of daily exposure to AI-generated self-modeling media against a control condition. Participants were assigned to one of three groups based on their original physical condition: (1) video self-modeling (VSM), (2) audio self-modeling (ASM), or (3) Control (without nudging).

4.1.1 Participants. We recruited 30 participants (10 per group) with balanced gender distribution and similar BMI across groups to minimize potential confounding effects in physical performance. Two participants (one from the video group and one from the Control group) withdrew before the first intervention due to scheduling conflicts, leaving 28 participants for analysis (see Table 1

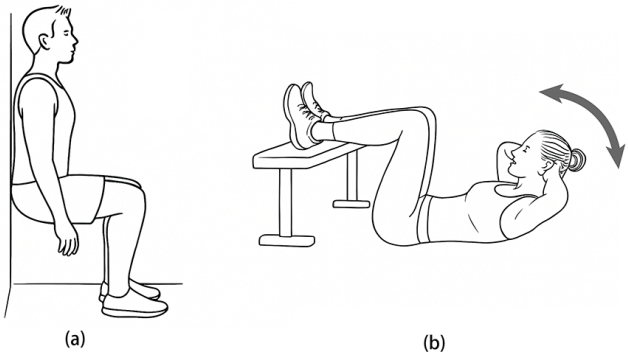


Figure 3: Two selected exercising tasks: (a) wall-sit, (b) crunch.

for demographics). Each participant received 20 USD per hour as compensation in accordance with local standards.

4.1.2 Tasks. We selected two baseline tasks, *wall-sit* and *crunch*. Both exercises (1) offer clear, quantifiable metrics for daily longitudinal tracking, (2) were validated in prior self-modeling work [13], and (3) target different muscle groups to avoid fatigue accumulation that could confound daily adherence: wall-sit for lower body and crunch for core/upper body. Wall-sit (Figure 3a) is a static lower-body strength-endurance exercise in which participants slide down a wall until their hips and knees are flexed at a 90° angle and hold this position for as long as possible. In contrast, the crunch (Figure 3b) is a dynamic core and upper-body strength-endurance task. Participants lie supine with knees bent, cross their hands over their shoulders, and repeatedly raise their torso to a vertical position before lowering back down. Combining one static lower-body and one dynamic upper-body/core exercise distributes muscular load and minimizes fatigue, enabling sustained daily participation.

4.1.3 Measurements. To measure the effectiveness of the nudging interventions, we utilized both objective and subjective measurements.

For objective measures, we recorded wall-sit duration and crunch repetitions. Participants performed both exercises once per day for 7 consecutive days, aiming to maximize repetitions for crunches and holding time for wall-sits. The results were recorded once they burned out or could not continue. To reduce individual variance, performance was normalized relative to each participant's Day 1 value. Specifically, we analyzed percentage outcomes defined as $Pct_i = 100 \times \frac{x_t - x_1}{x_1}$, where x_t is the performance value on day t and x_1 is the Day 1 baseline. This transformation enabled us to model relative improvements over time, with Pct serving as the dependent variable in our subsequent analyses.

For subjective measures, we administered three validated questionnaires adapted from prior works:

- **Intrinsic Motivation Inventory** (IMI) [57, 59] is used to measure intrinsic motivation and self-regulation [42]. We included the Interest/Enjoyment subscale (intrinsic motivation) and the Perceived Competence subscale (self-assessed task competence), rated on a 7-point Likert scale (see Appendix B.3).
- **Exercise Self-Efficacy Scale** (ESES), originated from Bandura's self-efficiency theory [1], measures confidence in conducting activities under various conditions. ESES is an adjunct domain proposed in the Self-Efficacy Scale [2] and has been validated by several studies [56, 64]. A 4-point Likert scale was used to rate the items (see Appendix B.4).
- **Video/Audio Identification Questionnaire** (VAIQ), adapted from Player Identification Questionnaire [67] for avatars in online games. We applied Perceived Similarity, Embodied Presence, and Wishful Identification subscales to AI-generated video identification, while only Perceived Similarity for audio identification, rated on a 7-point Likert scale (see Appendix B.5).

To minimize participant fatigue and inattentive responding, the full questionnaires were required on Day 1 (baseline) and Day 7 (post-intervention). On Days 2–6, participants completed shortened versions containing only essential items, while preserving psychometric validity.

4.1.4 Procedure. The study comprised three phases: Day 1 (lab session), Days 2–6 (remote sessions), and Day 7 (lab session). Figure 4 shows an overview of the experimental procedures for task 1.

Day 1 (Lab Session). Upon arrival, participants provided informed consent, completed demographics (Appendix B.1), and filled out the Physical Activity Readiness Questionnaire (PAR-Q) to ensure they could safely perform the exercises. Eligible participants then filled out the IMI and ESES. Group-specific preparations followed: the VSM group provided a headshot and chose a peer model for AI video generation; the ASM group recorded a script and submitted AGQs for voice cloning; the Control group skipped this step. To minimize expectancy effects, participants were not informed about the existence of other groups.

While waiting for the generation of AI contents, participants in all groups were introduced to two target exercises, wall-sit (measured in duration) and leg-raise crunches (measured in repetitions). Then they viewed a brief example video demonstrating correct form and received in-person posture guidance with practice trials. Afterward, they were exposed to their assigned intervention:

- **VSM**: personalized AI-generated video (videos were watched before each exercise).
- **ASM**: personalized AI-generated audio clip (audios were played once before all tasks).
- **Control**: neutral pre-scripted verbal instructions.

Participants subsequently performed the wall-sit and crunch tests to exhaustion, with 2 minutes rest between exercises. Performance was manually counted and recorded. A shortened IMI, ESES, and VAIQ (ASM, VSM only) were then required to fulfill after finishing the tasks.

Days 2–6 (Remote Sessions). Participants attended daily remote sessions via video-conference, scheduled at consistent times when possible. Before each session, the experimenter confirmed participants' physical condition and absence of injuries. Then participants repeated the same exercise–intervention–exercise sequence as on Day 1. Objective performance was recorded remotely, and subjective measures were collected through online questionnaires.

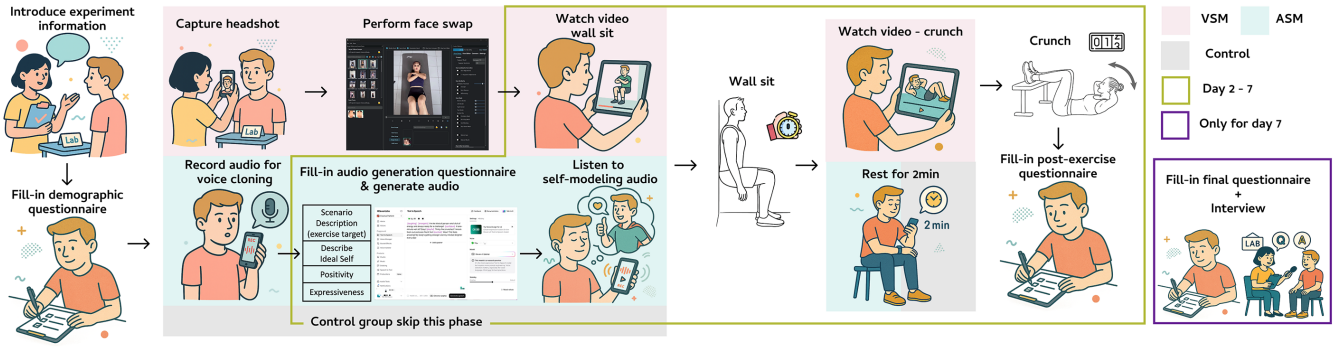


Figure 4: Overview of the experimental procedure for three groups (VSM, ASM, and Control). The diagram illustrates the sequential tasks on Day 1 and subsequent sessions on Day 2-Day 7.

To ensure the variations in nudging content, experimenters randomly selected pre-generated peer model videos for VSM. While for ASM, participants are required to complete the AGQ 30 minutes before daily remote session for generating personalized audio clips.

Day 7 (Lab Session). Participants returned to the lab to repeat the Day 1 testing protocol, including full IMI, ESES, and VAIQ (ASM, VSM only) questionnaires. Finally, participants completed an open-ended interview to share subjective impressions of the intervention and study experience.

4.2 Linear Mixed-Effects Analysis

We employed Linear Mixed Effects models (LME) for data analysis. Specifically, we modeled participant i's percentage performance Pct (wall-sit or crunch performance relative to Day 1 baseline) on day j as:

$$Pct_{ij} = \beta_0 + \beta_1 t_{c,ij} + \beta_2 t_{c,ij}^2 + Group_i \times (\beta_3 + \beta_4 t_{c,ij} + \beta_5 t_{c,ij}^2) + u_i + \epsilon_{ij} \tag{1}$$

where $t_{c,ij}$ is mean-centered day, $Group_i \in \{0, 1\}$ (0=Control, 1=VSM or ASM) denotes the intervention group, $u_i \sim \mathcal{N}(0, \sigma_u^2)$ is a participant random intercept, and $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ is residual error. We estimated the fixed-effect coefficients $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_5)$, which quantify population-average effects on Pct.

4.2.1 Test for RQ1 - sustained performance. The performance difference between the intervention group and the Control group can be formulated as:

$$\begin{aligned}\Delta(t) &= \mathbb{E}[Pct \mid Group{=}1] - \mathbb{E}[Pct \mid Group{=}0] \\ &= \beta_3 + \beta_4 t_c + \beta_5 t_c^2\end{aligned} \tag{2}$$

where $t_c = t - c$, and c is the mean study day. Choose an **early** reference time t_E (the mean of an early window) and a **late** time t_L (the mean of a late window). Define $t_{cE} = t_E - c$ and $t_{cL} = t_L - c$. We assess "sustained performance" with two linear-contrast tests from a single fit of Eq. 1:

(1) **Late advantage (level at the end).** We test whether the intervention still outperforms Control at the late time:

$$H_0 : \Delta(t_L) \le 0 \quad \text{vs.} \quad H_1 : \Delta(t_L) > 0. \tag{3}$$

Table 2: Estimated fixed-effect coefficients (β_0–β_5) from Linear Mixed Effects models for VSM and ASM conditions.

	Group	β_0	β_1	β_2	β_3	β_4	β_5
Wall-sit	VSM	13.58	13.06	2.63	39.80	4.97	−2.61
	ASM				−10.91	−7.10	−1.63
Crunch	VSM	20.45	6.75	−0.08	30.75	9.79	0.26
	ASM				7.12	3.69	0.64

(2) **Sustainment (late ≥ early).** We test whether the gap at the end is no smaller than early on:

$$H_0 : \Delta(t_L) - \Delta(t_E) < 0 \quad \text{vs.} \quad H_1 : \Delta(t_L) - \Delta(t_E) \ge 0. \tag{4}$$

We claim "sustained performance" if $\Delta(t_L) > 0$ and $\Delta(t_L) - \Delta(t_E) \ge 0$ (CIs exclude 0 in the one-sided direction).

4.2.2 Test for RQ2 - sustained improvement rate. We operationalize "sustained improvement rate" as a non-decreasing change in slope difference between the intervention and Control groups over time. In contrast, we define "**convergence**" as a decreasing slope gap once the nudging effect has peaked. Drawing on theories like habituation, we hypothesize that long-term participants may become desensitized to the intervention, leading to convergence in the improvement trajectory. In Eq. 1, this is captured by the quadratic term β_5 ($Group \times t_c^2$), which governs how the slope differences evolve across days. We test convergence with the following equation:

$$H_0 : \beta_5 \ge 0 \quad \text{vs.} \quad H_1 : \beta_5 < 0 \tag{5}$$

4.3 Results and Analysis

The overall performance of all groups is illustrated in Figure 5, with model fitting results reported in Table 2. Within the 7-day intervention, the VSM showed faster early improvements, though the rate of gain tapered over time, providing only partial support for "sustained performance". In contrast, the ASM group demonstrated less consistent improvements. These patterns suggest that different self-modeling modalities may shape performance trajectories in distinct ways.

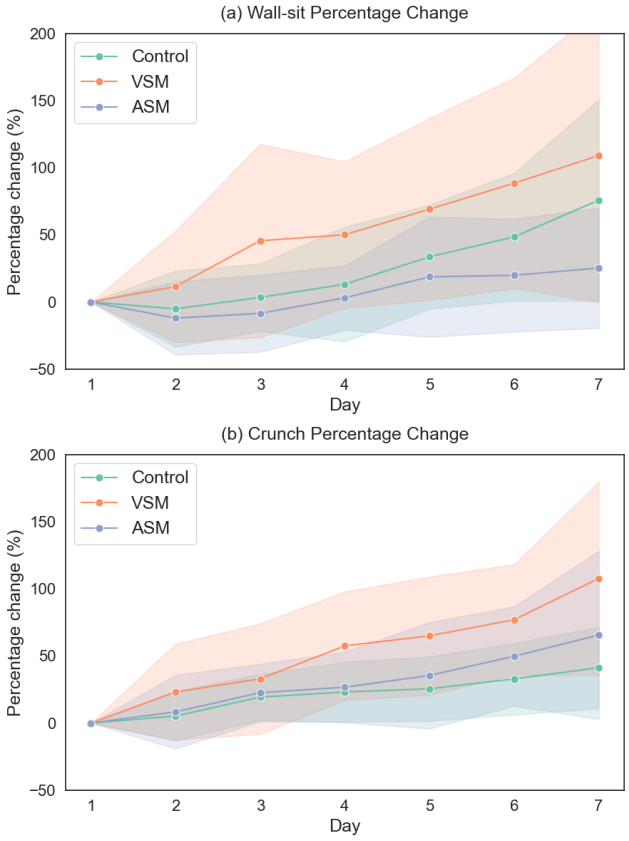


Figure 5: Average performance change for VSM, ASM, and Control groups over 7 days.

4.3.1 Understanding Control group's performance. The Control group displayed a steady upward trend throughout the week (Figure 5). The result of Mixed Linear Model regression confirmed significant positive linear effects of Time for the Control group on both wall-sit ($\beta_1 = 13.06, SE = 4.67, p = .003^{**}$) and crunch ($\beta_1 = 6.75, SE = 2.78, p = .008^{**}$). The findings indicate that participants in the Control group exhibit a gradual improvement in performance, reflecting natural practice and learning effects without any personalized intervention.

4.3.2 RQ1: Does AI self-modeling sustain performance over time for both experiment groups? As shown in Figure 5, all groups improved on both wall-sit and crunch over the intervention period; the VSM group appeared to improve more rapidly, especially during the early days, while the ASM group exhibited less consistent gains. To evaluate the significance of VSM and ASM sustained performance gains, we tested both late advantages (end-of-week vs. Control) and sustainment (late vs. early advantage).

For wall-sit, neither VSM nor ASM demonstrated significant advantages. VSM's late advantage was positive but not reliable ($\Delta_L = 31.18, SE = 42.56, p = 0.232$), and sustainment was also not significant ($\Delta_L - \Delta_E = 29.81, SE = 44.48, p = 0.251$). ASM, by contrast, trended negatively, showing no late advantage ($\Delta_L = -47.25, SE = 41.46, p = 0.873$) and no sustainment ($\Delta_L - \Delta_E = -42.60, SE = 43.35, p = 0.837$). For crunch, VSM participants achieved a significant late advantage ($\Delta_L = 62.46, SE = 23.24, p = 0.004^{**}$) and sustainment across the week ($\Delta_L - \Delta_E = 58.71, SE = 25.05, p = 0.010^{**}$). ASM also showed a positive but weaker late advantage ($\Delta_L = 23.93, SE = 22.65, p = 0.145$) and non-significant sustainment ($\Delta_L - \Delta_E = 22.15, SE = 24.41, p = 0.182$) on performance.

The results indicate that AI self-modeling can support performance gains. At the same time, the weaker improvements observed for ASM may partly reflect the nature of the selected exercises, which rely heavily on posture and visual form. Because these tasks offer clearer visual than auditory information, different implementations of self-modeling may interact with the task context in distinct ways. We discuss this task-related limitation further in Section 7.3.

Key Findings: VSM sustained performance gains for crunch but not for wall-sit, while ASM showed no reliable benefit compared to Control.

4.3.3 RQ2: Does AI self-modeling sustain improvement rate over time for the VSM group? To test whether VSM affected improvement rate (slope) rather than only level, we examined the quadratic interaction term (β_5) for $Group \times t_c^2$. A significant negative β_5 indicates convergence, reflecting an unsustained improvement rate relative to the Control as the nudging effect diminishes.

For crunch, $\beta_5 = 0.26$ ($SE = 0.97$, $p = 0.606$), providing no evidence of convergence under Eq. 5, and likewise non-significant when testing the alternative ($\beta_5 > 0$). In contrast, wall-sit ($\beta_5 = -2.61$, $SE = 1.32$, $p = 0.023^{*}$) showed significant convergence. These results indicate that VSM did not sustain an accelerating improvement rate; instead, performance gains decelerated and converged toward Control levels, particularly in wall-sit. This pattern suggests that while self-modeling can boost early improvement, its rate effects diminish over time, reinforcing the importance of testing long-term trajectories in Study 2.

Key Findings: VSM intervention demonstrates a decreased improvement rate for wall-sit during one week of exercise, but not for crunch.

4.4 Conclusion and Implication for Study 2

Overall, the results provide partial evidence addressing our research questions. VSM improved crunch performance and sustained outcomes, but wall-sit effects were inconsistent, showing convergence by Day 6. In contrast, ASM provided no reliable benefits in this context.

4.4.1 Interpretation of ASM effects. Compared to VSM, the method of audio intervention was far less effective in this scenario. Several explanations emerged from both the performance data and follow-up interviews. **First**, many participants reported that the generated audios sounded funny or even distracting, which made it difficult for them to take the feedback seriously or recall it during exercise. **Second**, participants often perceived the audio feedback as a simple emphasis on their goals rather than as a meaningful representation of their progress. As a result, the audio cues provided only limited encouragement. **Third**, unlike video feedback, which directly visualizes one's body movements and progress, audio is relatively abstract and less embodied, reducing the potential to enhance their

Table 3: Summary of participants in each cohort by experimental condition.

Cohort	Control	VSM	Total
Cohort 1	9	9	18
Cohort 2	6	7	13

self-efficacy and engagement in training contexts. Taken together, these factors may explain why ASM produced little improvement beyond natural practice effects.

4.4.2 Implication for Study 2. For **RQ1**, we can partially conclude that VSM nudging can sustain performance for crunches over one week. For **RQ2**, VSM nudging did not sustain an improvement rate and instead demonstrated signs of convergence on wall-sit. However, as only one task reached statistical significance, the evidence is not strong. Moreover, one week of exposure might not be sufficient for a long-term study [18]. Therefore, we recruited more participants and expanded the experiment cycle to 4 weeks in Study 2. Additionally, through ASM participants' feedback and performance, we discarded the ASM group in Study 2.

5 Study 2: Longitudinal Effects of Video Self-Modeling

Building on the findings from Study 1, we designed a one-month longitudinal study to examine the durability of AI self-modeling and focused exclusively on VSM. Extending the duration allowed us to capture performance and improvement rates trajectories over time, and to test whether the nudging effect of AI self-modeling can persist in everyday fitness practice.

5.1 Study Design

This subsection details the participant recruitment and study procedure.

5.1.1 Participants. This study was composed of 31 participants in total. Among these 18 participants were from Study 1, of whom we referred to as Cohort 1; we also included 13 new participants, whom we referred to as Cohort 2 (see Table 3). Study 2 aims to track performance over 4 weeks; thus Cohort 1 was expected to complete three additional weeks, and Cohort 2 was for four weeks, both following the same procedure. Each participant was compensated 20 USD per week. Figure 6 shows the engagement gradually declined from 31 to 15 participants by the end, a typical attrition pattern in long-term interventions.

To ensure comparability, all core elements, including *peer models* (Section 3.1.1), *tasks* (Section 4.1.2), and *outcome measures* (Section 4.1.3), were kept identical to Study 1. The only difference was a short gap for Cohort 1 between studies, whereas Cohort 2 began without interruption. As shown in gap validation analysis (see Section5.2.1), this gap did not introduce systematic bias, allowing the two cohorts to be pooled for further analysis.

5.1.2 Procedure. Informed by participants' feedback from Study 1, we conducted Study 2 entirely through remote sessions. Participants followed the structured rules to ensure consistency and reliability:

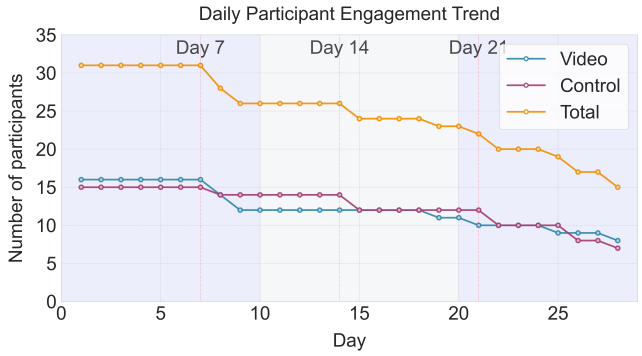


Figure 6: Daily active participants during the 28-day intervention.

- **Voluntary participation:** Unlike Study 1, participants could withdraw at any point during the one-month period.
- **Weekly unit:** Compensation was calculated by full weeks only; partial weeks were not compensated.
- **Autonomy:** Exercises were completed independently at home without direct monitoring from researchers.
- **Recording protocol:** Participants submitted daily video recordings in a standardized format: began with a phone screen showing the current local time, followed by continuous footage of the session. For VSM participants, the video-watching phase was also included.

Onboarding session. For cohort 2, we conducted an online onboarding session to introduce the study protocol, including the nudging intervention (VSM only), tasks, and recording requirements. Participants also completed the baseline questionnaires (IMI, ESES, and demographics) and a remote Physical Activity Readiness questionnaire to confirm eligibility.

VSM participants selected a preferred peer model and submitted a high-resolution headshot, which was used to generate four personalized videos (two wall-sits, two crunches) with varied contextual settings. The Control group relied solely on instructions from researchers without AI-generated content.

Daily routine. Each day, participants completed the following procedures independently at home: **First**, perform *wall-sits* to exhaustion (preceded by watching a wall-sit video for VSM, or following researcher instructions for the Control); **Second**, rest for two minutes; again perform *crunches* to exhaustion following the same intervention for each group; **Finally**, upload the recording to the Cloud Drive and complete a daily questionnaire on objective performance and subjective motivation.

Post-study session. At the end of the period, participants completed the full IMI, ESES, and VAIQ (where applicable) questionnaires again, followed by an open-ended survey about their experiences and perceptions of the intervention.

5.2 Data Integrity and Attrition Analysis

Before the main analysis, we conducted three checks to validate the effectiveness of our data. First, we examined whether merging the two cohorts (continuing vs. new participants) was statistically justified. Second, we rigorously assessed attrition patterns and

potential survivor bias to ensure that dropout did not confound the comparison between intervention conditions. Third, we assessed the reliability of the subjective scales on the pooled data, confirming internal consistency across items.

5.2.1 Gap Validation for Cohorts. To examine whether the gap between Study 1 and Study 2 introduced systematic bias in Cohort 1, we compared early improvement slopes between Cohort 1 and Cohort 2 during the initial phase of Study 2 by including Cohort in Equation 1 as a variable:

$$\begin{aligned} Pct_{ij} = \gamma_0 + \gamma_1 t_{c,ij} + \gamma_2 t_{c,ij}^2 + Cohort_i \times (\gamma_3 + \gamma_4 t_{c,ij} + \gamma_5 t_{c,ij}^2) \\ + Group_i \times (\gamma_6 + \gamma_7 t_{c,ij} + \gamma_8 t_{c,ij}^2) + u_i + \epsilon_{ij} \end{aligned} \tag{6}$$

where $Cohort_i \in [0, 1]$ denotes cohort 1 or not. We fitted LME with performance data from Days 1-14, centering time at Day 7 so that the $Cohort \times t_c$ interaction represented the instantaneous slope difference at Day 7. For crunches, the Day 7 slope difference between cohorts was small and non-significant ($\gamma_4 = -0.70, SE = 0.53, p = 0.182$); for wall-sit, it was also not significant ($\gamma_4 = -1.30, SE = 1.88, p = 0.491$). Thus, there was no evidence of systematic cohort bias, justifying the pooling of Cohort 1 (continued) with Cohort 2 (new recruited) for the subsequent longitudinal analyses.

5.2.2 Attrition and Survivor Bias Analysis. We first examined whether attrition differed across cohorts or intervention groups, as differential dropout could bias longitudinal estimates. We compared (1) dropout proportions across $Cohort \times Group$ combinations using a Fisher exact test, and (2) attrition timing across the 28-day period using a Kaplan–Meier log-rank test. Both analyses showed no significant differences (Fisher exact $p = 1.00$; log-rank $p = 0.820$), indicating that neither cohort nor intervention condition exhibited distinct survival patterns.

To further rule out survivor bias (i.e., whether only highly motivated participants remained), we conducted a split analysis comparing the baseline characteristics (intrinsic motivation and Week-1 average performance) of *Completers* (who finished the 4-week study) versus *Dropouts* separately for each condition. For the VSM group, independent t-tests revealed no significant differences between Completers and Dropouts across all baseline metrics ($p > 0.050$), indicating that attrition was non-systematic and the intervention was effective across a broad range of users.

In the Control group, while baseline physical performance was similar ($p > 0.050$), Completers had significantly higher IMI-Interest /Enjoyment ($p < 0.001^{**}$) and Competence ($p = 0.042^{*}$) than Dropouts. This result implies a conservative comparison for our main findings: the retained Control participants represented a subset of highly motivated individuals. Consequently, any performance advantage observed in the VSM condition would be achieved against a "highly motivated" Control baseline, thereby reinforcing the robustness of the intervention's effect.

Detailed statistical tables and survival curves are provided in Appendix D.

5.2.3 Subjective Data Consistency Validation. We evaluated the reliability of subjective scores for IMI, ESES, and VAIQ using Cronbach's α, a standard index of internal consistency within a scale [17]. Values above 0.70 are generally considered acceptable, with higher values indicating stronger reliability. IMI reached excellent internal consistency for both Interest ($\alpha = 0.83$) and Competence ($\alpha = 0.86$) subscales; ESES also showed strong reliability ($\alpha = 0.92$). For VAIQ, all dimensions demonstrated extremely high internal consistency with α values approaching 1.0.

5.3 RQ1: Does AI self-modeling (VSM) sustain performance over time?

We applied the same LME analysis methodology as described in Study 1 (Section 4.2) to assess performance in this one-month study, with the fitted curves shown in Figure 8. Given the study length, we defined early and late stages as the average over 7 consecutive days (Days 1–7 and Days 22–28, respectively).

For wall-sit, results indicated a significant late advantage ($\Delta_L = 31.34$, $SE = 12.70$, $p = 0.007^{**}$), showing that the VSM group clearly outperformed the Control group in the final stage of the program. However, the sustainment contrast was not significant ($\Delta_L - \Delta_E = 13.22$, $SE = 15.18$, $p = 0.192$), suggesting that the observed late-stage gap may not represent a statistically significant increase compared to the early-stage difference. For crunch, we again observed a robust late advantage ($\Delta_L = 27.98$, $SE = 7.94$, $p < 0.001^{**}$). The sustainment test showed a positive trend ($\Delta_L - \Delta_E = 13.37$, $SE = 9.48$, $p = 0.079$), but did not reach conventional significance thresholds. Across both tasks, AI self-modeling (VSM) reliably produced superior performance at the late stage, confirming that its benefits did not vanish over time. Nevertheless, the evidence for sustaining or increasing relative gains is weaker: although trends point toward maintenance or slight growth of the advantage, the sustainment contrasts did not achieve strong statistical support. These findings suggest that VSM nudging secures lasting late-stage improvements, but further work is needed to verify whether such advantages consistently amplify with time.

Key Findings: In a four-week duration, VSM reliably leads to better performance at the late stage, whereas the sustainment of the relative gain is suggestive but not significant for both tasks.

5.4 RQ2: Does AI self-modeling (VSM) sustain improvement rate over time?

Study 1 showed that VSM did not sustain an accelerating rate of improvement; instead, performance gains tended to converge. To verify and characterize this convergence in a longer time frame, we combined LME analysis (Eq. 5) with subjective reports.

5.4.1 Stepwise Convergence Analysis. As illustrated in Figure 7, two distinct phases were suggested: an early rapid-growth phase for VSM, followed by convergence toward the Control group. To avoid obscuring such phase-specific dynamics and identify when the VSM growth rate began to decelerate, we adopted a stepwise, windowed strategy: we segment the series into progressively larger contiguous windows (7–28 days) and fit Eq. 1 within each window. The evolving pattern of β_5 (quadratic $Group \times t_c^2$ term) and its p-value is shown in Figure 9, revealing the two-stage convergence pattern.

In the early to mid phase (Days 7–15), β_5 values for both exercises were negative and accompanied by gradually significant

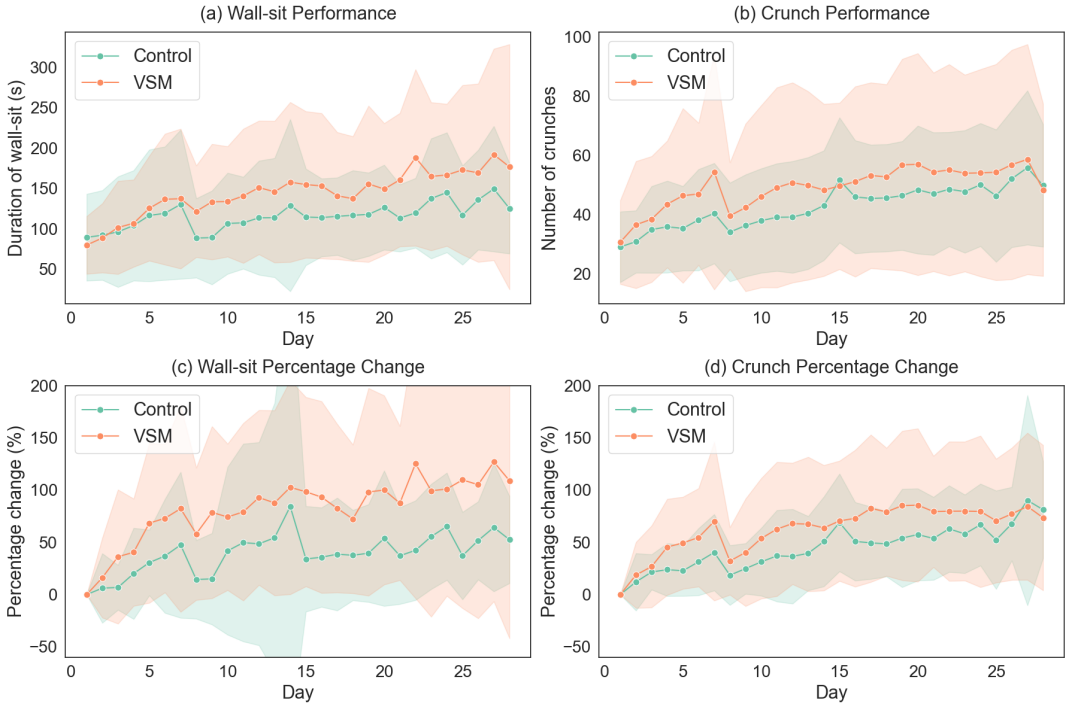


Figure 7: Average objective performance and relative performance change for VSM and Control groups across four weeks.

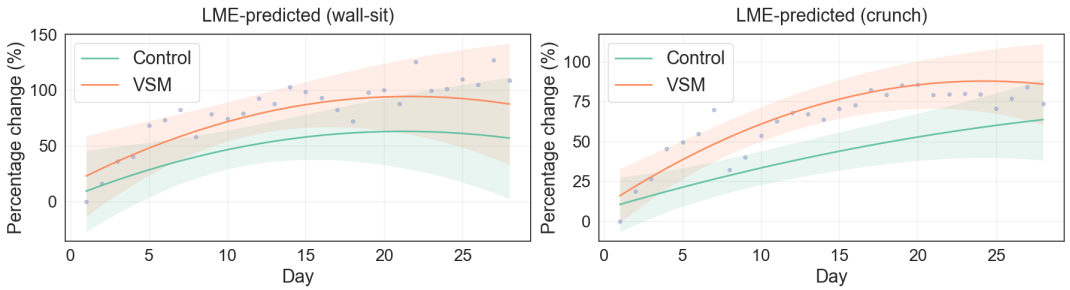


Figure 8: Fitted performance curves over the one-month study based on the LME model.

p-values ($p < 0.050$ around Day 15 for crunch, Day 11-21 for wall-sit), indicating the existence of a decreasing improvement rate. This suggests that the strong initial acceleration of the VSM group's improvement began to weaken, reducing the improvement rate gap with Control. However, in the later phase (Days 21–28), β_5 continued to increase toward zero without significance. This fact implies that the VSM group's performance does not demonstrate a consistent parabolic trend; instead, their performance change possibly converged to the same trend as the Control group in the later phase. This shift suggests that while VSM sustained its early advantage throughout Study 2 (Section 5.3), the marginal contribution of the intervention diminished over time.

Key Findings: The trajectory of the VSM group's performance is staged: an early rapid-growth burst when participants are first exposed to VSM nudging; a mid-phase convergence in which improvement rate decelerates (reflected by negative β_5); and a late phase (Days 21–28) where the VSM and Control groups' performances run roughly in parallel, implying attenuation of the VSM benefit.

5.4.2 Subjective Motivation Analysis. Subjective measures provide additional insights into the mechanisms underlying the performance patterns observed in RQ1 and RQ2. As shown in Figure 10, participants' psychological experiences followed a two-stage pattern: an initial novelty-driven boost, followed by attenuation and partial stabilization.

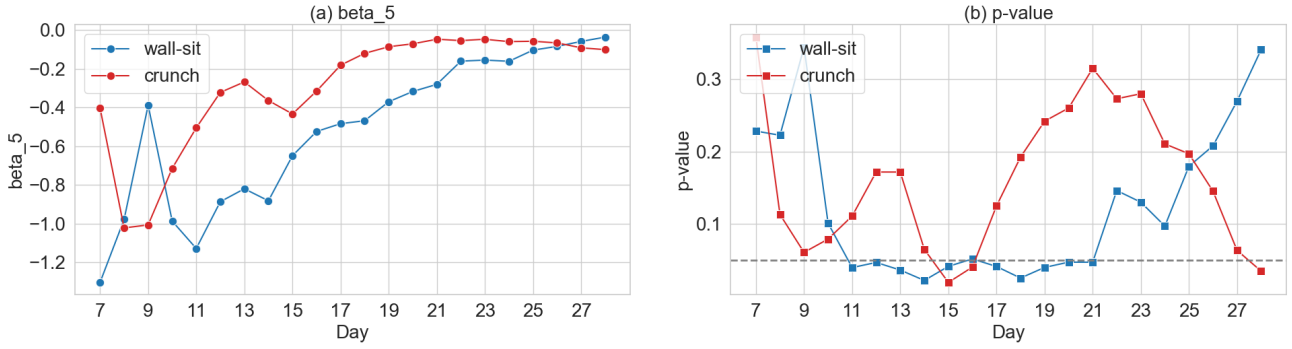


Figure 9: Stepwise analysis of convergence: quadratic coefficients (β_5) and p-values across study days.

Figure 10: Subjective measures of motivation and identification: changes in IMI, ESES, and VAIQ subscales over time.

For the VSM group, both IMI-Interest/Enjoyment (Figure 10a) and IMI-Perceived Competence (Figure 10b) exhibited a sharp increase within the first week, peaking around Days 5–7 before declining toward control levels. This trend reflects the objective performance, suggesting that AI-generated videos initially stimulated excitement and competitiveness, but the motivational salience diminished with repeated daily exposure. Similarly, measures of identification experienced an early rise. VAIQ-Perceived Similarity (Figure 10d) and Embodied Presence (Figure 10e) increased markedly during the first week, then stabilized or fluctuated. These results indicate that novelty facilitated strong short-term self-identification, but its impact leveled off over time.

For the ESES (Figure 10c), the VSM group demonstrated an overall upward trend relative to Control, suggesting that although the direct nudging effect of video self-modeling attenuated over time, the intervention fostered a more durable sense of exercise self-efficacy. The slight decline observed in the final week may reflect attrition effects, as the number of participants had dropped to 15. Similarly, VAIQ-Wishful Identification (Figure 10f) showed a steady increase, indicating that participants' pursuit of their ideal self and identification with the peer model became stronger over time. This sustained confidence and strengthened aspiration toward the ideal body explain why VSM participants maintained superior performance levels compared to the Control group even after novelty effects had diminished. In Section 7.1, we further examine these

mechanisms through interviews and describe the process as an early catalyst followed by internalization.

5.4.3 Conclusion. In this study, we extend the short-term findings of Study 1 and demonstrate that VSM can sustain higher performance levels across a four-week intervention. For RQ1, VSM participants maintained superior outcomes relative to Control, even after novelty effects diminished. RQ2, however, the improvement rate did not sustain; instead, performance gains followed a two-stage trajectory of early acceleration and later convergence. Subjective measures further reinforced this pattern, showing initial jumps of interest and identification that later stabilized into more durable self-efficacy and wishful identification.

6 Design Implications

Based on our post-study interviews and qualitative analysis of participants' reflections, we distill three design implications for future behavior change technologies (BCTs). Interviews were fully transcribed and analyzed using Braun and Clarke's six-phase reflexive thematic analysis [8]. Two researchers independently reviewed and coded the data before collaboratively refining a shared codebook. Detailed interview questions and our coding protocol are provided in Appendix C. The resulting themes informed three implications: (1) modality choices illustrated by the contrast between VSM and ASM; (2) personalized feedback informed by reflections on one's performance/subjective data; (3) participants' direct design suggestions, which emphasize desired system features and interactive functions.

6.1 VSM vs ASM: Embodiment as Design Priority

Study 1 revealed a significant contrast in effectiveness between VSM and ASM: VSM improved and sustained performance, while ASM failed to outperform the control condition. We posit that this difference stems from VSM's superior ability to facilitate embodiment, which is crucial for fitness tasks. This contrast highlights three critical dimensions for future design:

Design for embodiment: enhancing motivation for physical tasks. Embodiment is a powerful mechanism for motor learning and motivation. As established in embodied cognition and model-based learning research [27, 61] and prior HCI work on visual guidance for movement [39, 60], physical tasks rely on perceptual–motor mapping that is effectively supported through visual cues. VSM's effectiveness likely stems from its embodied visual presentation, providing a concrete, inimitable "visual anchor" that helps participants map their own form, perceive gaps, and identify improvements. As P-V10 stated, *"it feels like this is a body that I can achieve through training"*. In contrast, the auditory cues of ASM are disconnected from the user's physical experience, creating a cognitive and somatic gap. Bridging this gap may require moving beyond verbal cues to interactive audio systems, which have been shown to effectively foster embodiment [4].

Design for "becoming" instead of "improving". Both ASM and VSM employed idealized self-models, while only VSM successfully fostered a sense of "becoming", achieving the internalization of the self-model as a plausible future identity that supports sustained motivation. VSM's visual avatar naturally shifts users toward an aspirational identity, consistent with Possible Selves [40] and Identity-Based Motivation [47]. In contrast, ASM did not provide the necessary conditions for identity adoption. During the goal-setting phase, ASM participants frequently articulated process-oriented goals (e.g., "be persistent," "work harder") rather than future-self outcomes, placing them in an "improving" mindset rather than a "becoming" one. As P-A2 explained, auditory cues had a limited lasting impact: *"Once you start exercising, you basically won't think about it anymore."* Similarly, P-A6 viewed the cues as transient support: *"I prefer it gives me some encouragement words during the exercise."* This suggests that the key design challenge for ASM is how to guide users toward a more embodied and identity-aligned experience, rather than merely offering momentary performance prompts.

Fidelity must surpass a threshold to create immersion. Our study reveals the existence of a "fidelity threshold": while VSM's face-swapping was imperfect, its fidelity was sufficient for immersion. For example, although participants P-V5, P-V6, P-V7, and P-V13 noted that the *"facial expressions were unnatural compared to daily life,"* they consistently reported still being able to recognize themselves; as P-V5 stated, *"I can still recognize myself in the videos."* Conversely, ASM's voice-cloning, perceived as "not similar", caused participants to feel alienated, turning motivation into distraction. P-A4 described the audio as *"slightly deliberate, a little unnatural... with interjections and long pauses,"* which disrupted immersion. Thus, systems need only surpass this threshold to sustain immersion, rather than achieve perfection. This may reflect a perceptual asymmetry challenge in ASM: we hear ourselves through a mix of air- and bone-conduction that makes the "internal" voice sound warmer and lower than purely air-conducted playback [50].

6.2 Personal Data Reflection: Tailored feedback, Self-referencing, and Dynamic Intervention

Participant's reflection on their personal data reveal that a one-size-fits-all intervention is insufficient due to individual differences. The design of BCTs should shift towards personalization: understand and adapt to the unique exercise pace, psychological states, and motivations of users.

Provide tailored feedback according to user archetypes. Through interviews, we identified three typical user archetypes: (1) "progressives", (2) "maintainers", and (3) "context responders". Progressives such as P-C2, P-V5, and P-V8 described a consistent desire to "do a little more each day." P-C2 shared, *"it became a habit that I added one or two more each day... and near the end I felt happy."* P-V5 echoed this pattern: *"I always set a small goal, just a few more crunches than yesterday."* Similarly, P-V8 described: *"Every day I wanted to push a bit further—six minutes, then seven, then eight... I kept setting higher goals, and it felt good to accomplish them."* Maintainers, by contrast, preferred stable, repeatable targets; P-V9 noted setting a "generally similar goal for each day." Meanwhile, context responders showed highly variable patterns driven by external factors. P-C7, for instance, tied his strongest performance to feeling unusually energetic on a particular day: *"I remember feeling very energetic the next day, so I did a lot."* These archetypes suggest that systems should provide differentiated goals and feedback, especially for quantifiable exercises like crunches.

Take a step further, inferring a user's strategic archetype from their exercise trends could achieve a smoother interaction experience.

Emphasize self-referencing instead of peer comparison. AI self-modeling enhances motivation by providing a self-referencing role model instead of peer comparison. Self-confidence is closely associated with perceived success, and self-referencing can help build up a positive feedback loop. *"I can easily achieve the improvement goal (set by myself), feeling more confident"* (P-V5), *"at first I could barely hold (wall-sit), then I could stay much longer, I felt my improvement and my confidence went up too"* (P-V11). Conversely, peer comparison can bring anxiety and harm to self-confidence in the long run. P-V8 lowered his competence score because *"I sometimes wondered if others were doing better than me... my competence score would be lower"*. Peer comparison may work in the short term, but it introduces risk in the long run.

Dynamic intervention for shifting motivational states. Participants point out two motivational states: (1) a positive feedback loop of "effort-progress-satisfaction", (2) a negative feedback loop of "boredom-stagnant-burden" when exercise becomes a task. P-V12 described his positive feedback loop as *"a kind of push that made me exercise... it was more like a stepwise improvement... getting closer to the video gave me a real sense of achievement"*. Oppositely, P-V6 fell into a negative feedback loop after novelty decayed - *"my interest definitely went down and I felt the effect wasn't noticeable anymore...treating it like a task totally dampened my enthusiasm"*. This suggests the need for a dynamic intervention strategy that pulls users out of the negative feedback loop.

6.3 Participant Feedback: Interactivity, Diversity, and Adaptive Goals

Specified suggestions from participants offer valuable insights to polish AI self-modeling into a user-centered design. Their feedback highlights the need for an interactive, adaptive, and motivating companion, beyond static demonstrations.

From static demonstration to interactive guidance. Participants reported a sense of distance and a mechanical feel from the VSM system. As P-V11 suggested, *"If the AI video could provide personalized feedback on my form, it would increase my interest and attention when watching it"*, requiring the system to move beyond simple video playback. Future designs could integrate real-time pose analysis to provide immediate corrections, upgrading the experience from watching to being guided. Introducing a follow-along mode could also create a sense of virtual companionship.

Combating habituation with variety and progression. Content repetitiveness was a key factor in the decline of user interest over time. P-V12 explicitly stated that he *"lost interest in the later stages due to the high repetitiveness of the videos"* and suggested *"minor updates to the exercises every week"*. Beyond varying content, highlighting users' improvement can re-energize engagement and motivate continued effort. P-V13 expressed a desire to *"see a quantitative growth curve of my physical metrics"*.

Adapting goals to user progress. An ideal AI self-model presents a near-future self that modestly outperforms the user's current state. Conversely, a static, perfect AI self-model can lead to frustration due to the perceivable gap in performance. P-V10 noted that *"I kind of doubled my crunch...but the standard in the video still felt out of reach"*. A more effective strategy is dynamic avatars that synchronize user progress with adaptive goals to reinforce a sense of improvement.

7 Discussion

7.1 From Nudging to Internalized Motivation

Our findings suggest that the motivational nudging process of VSM unfolds in three stages, beginning with an early catalyst effect, followed by novelty decrease, and eventually stabilizing through internalization of an "ideal self". The process reflects a broader understanding of human behavior change: instead of being passively nudged by environmental cues, people can transform external signals into enduring motivation [58], allowing interventions to persist beyond habituation.

While both VSM and ASM were designed as identity-based nudges, only VSM exhibited this three-stage nudging process. ASM showed limited early gains but did not transition toward internalized motivation.

Early catalyst: clear visual goal and attainability. At the beginning, VSM provided an actionable movement model and a tangible aspirational target, elevating the starting baseline relative to Control. Unlike normal fitness influencer content that is hard to achieve, the AI-generated videos struck a balance between role-modeling and attainability. By presenting exemplars that felt achievable, VSM enhanced self-efficacy and perceived control [28, 44], thereby accelerating early performance gains. In this stage, participants benefited from the clarity of a concrete visual target, which helped translate abstract goals into an actionable effort.

Novelty decrease: fading stimulation and lack of feedback. As the exposure continued, the impact of VSM nudges diminished. The repetitive nudging made the intervention more predictable, and the stimulation was no longer as significant. As habituation theory explained, the repeated exposure to the same stimulus reduces its motivational salience over time [54, 65]. Beyond fading novelty, the absence of feedback, critical for achieving both effectiveness and persistence in behavior change [3, 19, 25], made the participant reinterpret the goal as distant or discouraging. This explains why the rate of improvement slowed and trended toward convergence.

Internalization of the "ideal self": a durable self-standard outlives novelty decay. Even as attention to external stimulation declined, VSM continued to shape behavior by shifting from an external prompt to an internalized self-standard. Unlike traditional BCTs that depend on extrinsic cues such as reminders or rewards—and thus rarely sustain long-term impact unless personally meaningful [58]—VSM fostered a transition from responding to visual stimulation to interpreting the AI self-model as a durable self-standard. This finding reinforces the idea of identity-based motivation theory that people are more likely to persist when goals align with their sense of "who I am" or "who I want to become" [47]. Rather than merely fading as habituation theory predicts [54, 65], our findings suggest a more nuanced understanding: external intervention can be reframed into self-referential standards that sustain motivation.

7.2 Generalization to Diverse Application Domains

Our study demonstrates that VSM can effectively enhance motivation and performance, as evidenced by sustained improvements in wall-sit and crunch, showing its potential for supporting fitness behaviors. In contrast, ASM exhibited more variable effects: while it did not uniformly underperform, there were cases in which it interfered with performance. These findings suggest that the effectiveness of AI self-modeling is both modality-dependent and task-dependent, motivating a closer examination of how modality and task characteristics shape whether self-modeling supports or hinders sustained engagement.

From wall-sit and crunch to diverse fitness tasks. Wall-sit and crunch represent only two examples of physical exercise, but the underlying principle of VSM is not limited to these tasks. Many fitness goals are inherently tied to visual representation, such as posture, body shape, or movement quality. VSM can present an attainable "ideal self" in the most intuitive way, suggesting the potential to support motivations for other daily exercise routines. However, in more specialized and demanding performance contexts, such as athletic training or rehabilitation, where improvements may be less visible or require technical precision, the advantage of VSM remains uncertain.

Modality and tasks across broader domains. AI self-modeling can extend to tasks beyond fitness to other domains requiring sustained effort with a meaningfully modeled "ideal self". For instance, in skill learning, learners may benefit from seeing and hearing themselves playing an instrument or speaking other languages. In mental health, a calmer or more confident future self could counter negative emotions. In lifestyle changes such as diet or study habits, self-modeling may strengthen identity-linked goals.

A key question for future application is how to match modality to domain. Our findings suggest that effectiveness depends on the representation of activities. Visually demonstrable tasks like fitness, sports, and diet are more likely to benefit from VSM, whereas ASM may have more potential in acoustic-related domains such as music, language learning, or emotional support. Exploring these applications can broaden the utility of AI self-modeling and clarify modality-specific boundaries.

7.3 Limitation and Future Work

Task selection and task–modality alignment. We selected wall-sit and crunch for their clear, quantifiable performance metrics, which strengthened the reliability of daily longitudinal measurement. However, both tasks rely heavily on posture and visual form. This task choice may have limited the robustness of our evaluation and may have interacted differently with the two implementations of self-modeling. As noted in Study 1, posture-dependent exercises offer stronger visual affordances than auditory ones, which may partly explain the weaker and more variable improvements seen with ASM.

To fully understand whether audio-based self-modeling exhibits motivational patterns comparable to video-based self-modeling, future studies should examine task types that align more closely with auditory cues (e.g., endurance pacing, rhythm-based training, or internal-state tasks). Such tasks would allow a more balanced evaluation of modality–task fit and help determine whether the observed differences stem from the modality itself or from the characteristics of the selected exercises.

Ambiguous effects of ASM: support or hinder? ASM underperformed the Control group in wall-sit and outperformed in crunch, without significance. As several participants noted, auditory cues quickly faded from attention once exercising began. Other participants also reported that the audio clip sounds funny to them. It is unclear whether ASM supports or hinders exercise tasks, and for what reason. Combined with the previous limitation, future work should investigate deeper into the properties and effects of ASM, as well as explore the reasons behind it.

Participant motivation and monetary compensation: Although financial compensation is common in long-term HCI studies, it is part of the reason that contributed to sustaining participant engagement. Several participants reported that the target exercises were not aligned with their personal fitness goals, suggesting that intrinsic motivation may not have been fully activated. Consequently, monetary incentives may have overshadowed the motivational contribution of self-modeling, especially in later stages of the intervention.

This suggests a more realistic scenario. The next step of the research should evaluate self-modeling in voluntary product settings without monetary rewards. For instance, embedding VSM into community fitness challenges, goal-tracking apps, or wearable-based health programs where users already have an intrinsic interest. Long-term deployments in such naturalistic environments would allow a clearer understanding of self-modeling's genuine motivational impact with fewer interventions of other factors.

Limitations in AI self-modeling technology. The underlying technology for both VSM and ASM had practical imperfections. For VSM, some of the generated synthesized faces were occasionally unsettling, leading to a uncanny valley effect, while some were too unrealistic to be fully motivating. These reactions point to a well-recognized challenge in embodied AI design: self-representations that are too accurate can appear eerie, while those that are too stylized can weaken identification. Determining the optimal representational fidelity that is recognizable, motivational, and non-threatening remains an open design problem for video-based self-modeling.

For ASM, fidelity issues arose from both technological and perceptual sources. Voice similarity varied across individuals, and also, the voice one hears internally differs from the voice captured externally due to bone conduction and resonance effects in human hearing. As a result, even accurately cloned voices sometimes felt "not like me," reducing the sense of identification with the audio self-modeling. Addressing this gap may require voice transformation methods that approximate users' self-perceived voice, or interactive calibration that allows users to adjust vocal timbre toward a more self-congruent representation.

These technological and perceptual constraints likely influenced how strongly participants engaged with the self-modeling content. Future systems should explore balanced and customizable visual realism, as well as voice cloning calibrated to users' self-perceived identity. Beyond fidelity improvements, expanding from static interventions to interactive and adaptive systems, such as multimodal self-modeling, context-aware feedback, or dynamically generated

goals, may better sustain engagement over extended periods and support the transition from externally prompted nudges to internalized, self-directed motivation.

8 Conclusion

This work conducted one of the first long-term empirical evaluations of AI self-modeling in the fitness domain. The first one-week study with 28 participants confirmed video self-modeling's (VSM) effectiveness, while audio self-modeling (ASM) failed to provide benefits for our setting, confirming the experiment setup for the next longitudinal study. Then, a 4-week study with 31 participants demonstrated that VSM can sustain higher performance levels but failed to maintain an accelerating improvement rate. By triangulating objective performance data with subjective measures and post-study interviews, we show that the nudging effect of AI self-modeling can be diminishing yet internalized into lasting motivation, leading to design implications for future behavior change technologies.

Acknowledgments

This paper is supported by the National Key R&D Program under Grant No. 2024YFB4505500 & 2024YFB4505503, National Natural Science Foundation of China under Grant No. 62132010, 62472244, Qinghai University Research Ability Enhancement Project under Grant No. 2025KTSA05, CCF-Lenovo Blue Ocean Research Fund, Tsinghua University Initiative Scientific Research Program, and Undergraduate Education Innovation Grants, Tsinghua University.

References

[1] Albert Bandura. 1977. Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review* 84, 2 (1977), 191–215. doi:10.1037/0033-295X.84.2.191

[2] Albert Bandura. 2006. Guide for constructing self-efficacy scales. In *Self-efficacy beliefs of adolescents*, Frank Pajares and Timothy C. Urdan (Eds.). Vol. 5. Information Age Publishing, Greenwich, CT, 307–337.

[3] Frank D Belschak and Deanne N Den Hartog. 2009. Consequences of positive and negative feedback: The impact on emotions and extra-role behaviors. *Applied Psychology* 58, 2 (2009), 274–303.

[4] David Birchfield, Harvey Thornburg, M Colleen Megowan-Romanowicz, Sarah Hatton, Brandon Mechtley, Igor Dolgov, and Winslow Burleson. 2008. Embodiment, Multimodality, and Composition: Convergent Themes across HCI and Education for Mixed-Reality Learning Environments. *Advances in Human-Computer Interaction* 2008, 1 (2008), 874563.

[5] Robert Jan Bood, Marijn Nijssen, John van der Kamp, and Melvyn Roerdink. 2013. The Power of Auditory-Motor Synchronization in Sports: Enhancing Running Performance by Coupling Cadence with the Right Beats. *PLOS ONE* 8, 8 (08 2013), 1–8. doi:10.1371/journal.pone.0070758

[6] Ralf Brand and Franziska Antoniewicz. 2016. Affective evaluations of exercising: the role of automatic–reflective evaluation discrepancy. *Journal of Sport and Exercise Psychology* 38, 6 (2016), 631–638.

[7] Ralf Brand and Panteleimon Ekkekakis. 2018. Affective–Reflective Theory of physical inactivity and exercise: Foundations and preliminary evidence. *German Journal of exercise and sport research* 48, 1 (2018), 48–58.

[8] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[9] Krzysztof Buczkowski, Ludmila Marcinowicz, Slawomir Czachowski, and Elwira Piszczek. 2014. Motivations toward smoking cessation, reasons for relapse, and modes of quitting: results from a qualitative study among former and current smokers. *Patient Preference and Adherence* 8 (2014), 1353–1363. arXiv:https://www.tandfonline.com/doi/pdf/10.2147/PPA.S67767 doi:10.2147/PPA.S67767 PMID: 25336926.

[10] Ana Caraban, Evangelos Karapanos, Daniel Gonçalves, and Pedro Campos. 2019. 23 ways to nudge: A review of technology-mediated nudging in human-computer interaction. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–15.

[11] Yu-Peng Chen, Julia Woodward, Dinank Bista, Xuanpu Zhang, Ishvina Singh, Oluwatomisin Obajemu, Meena N Shankar, Kathryn M Ross, Jaime Ruiz, and Lisa Anthony. 2024. Investigating contextual notifications to drive self-monitoring in mHealth apps for weight maintenance. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–21.

[12] Eun Kyoung Choe, Nicole B. Lee, Bongshin Lee, Wanda Pratt, and Julie A. Kientz. 2014. Understanding quantified-selfers' practices in collecting and exploring personal data. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 1143–1152. doi:10.1145/2556288.2557372

[13] Christopher Clarke, Jingnan Xu, Ye Zhu, Karan Dharamshi, Harry McGill, Stephen Black, and Christof Lutteroth. 2023. FakeForward: Using Deepfake Technology for Feedforward Learning. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 715, 17 pages. doi:10.1145/3544548.3581100

[14] James Clawson, Jessica A. Pater, Andrew D. Miller, Elizabeth D. Mynatt, and Lena Mamykina. 2015. No longer wearing: investigating the abandonment of personal health-tracking technologies on craigslist. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Osaka, Japan) *(UbiComp '15)*. Association for Computing Machinery, New York, NY, USA, 647–658. doi:10.1145/2750858.2807554

[15] Sunny Consolvo, David W. McDonald, and James A. Landay. 2009. Theory-driven design strategies for technologies that support behavior change in everyday life. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Boston, MA, USA) *(CHI '09)*. Association for Computing Machinery, New York, NY, USA, 405–414. doi:10.1145/1518701.1518766

[16] Jean Costa, Malte F Jung, Mary Czerwinski, François Guimbretière, Trinh Le, and Tanzeem Choudhury. 2018. Regulating feelings during interpersonal conflicts by changing voice self-perception. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–13.

[17] Lee J Cronbach. 1951. Coefficient alpha and the internal structure of tests. *psychometrika* 16, 3 (1951), 297–334.

[18] Carolina Del-Valle-Soto, Juan Carlos López-Pimentel, Javier Vázquez-Castillo, Juan Arturo Nolazco-Flores, Ramiro Velázquez, José Varela-Aldás, and Paolo Visconti. 2024. A comprehensive review of behavior change techniques in wearables and IoT: implications for health and well-being. *Sensors* 24, 8 (2024), 2429.

[19] Carlo C DiClemente, Angela S Marinilli, Manu Singh, and Lori E Bellino. 2001. The role of feedback in the process of health behavior change. *American journal of health behavior* 25, 3 (2001), 217–227.

[20] Peter W Dowrick. 1999. A review of self modeling and related interventions. *Applied and preventive psychology* 8, 1 (1999), 23–39.

[21] Peter W Dowrick. 2012. Self model theory: Learning from the future. *Wiley Interdisciplinary Reviews: Cognitive Science* 3, 2 (2012), 215–230.

[22] Peter W Dowrick. 2012. Self modeling: Expanding the theories of learning. *Psychology in the Schools* 49, 1 (2012), 30–41.

[23] Cathy Mengying Fang, Phoebe Chua, Samantha W. T. Chan, Joanne Leong, Andria Bao, and Pattie Maes. 2025. Leveraging AI-Generated Emotional Self-Voice to Nudge People towards their Ideal Selves. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 58, 20 pages. doi:10.1145/3706598.3713359

[24] Cathy Mengying Fang, Yasith Samaradivakara, Pattie Maes, and Suranga Nanayakkara. 2025. Mirai: A Wearable Proactive AI "Inner-Voice" for Contextual Nudging. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Association for Computing Machinery, New York, NY, USA, Article 399, 9 pages. doi:10.1145/3706599.3719881

[25] Ayelet Fishbach and Stacey R Finkelstein. 2012. How feedback influences persistence, disengagement, and change in goal pursuit. In *Goal-directed behavior*. Psychology Press, 203–230.

[26] Brian J Fogg. 2009. A behavior model for persuasive design. In *Proceedings of the 4th international Conference on Persuasive Technology*. 1–7.

[27] Lucia Foglia and Robert A Wilson. 2013. Embodied cognition. *Wiley Interdisciplinary Reviews: Cognitive Science* 4, 3 (2013), 319–325.

[28] Leire Gartzia, Thekla Morgenroth, Michelle K Ryan, and Kim Peters. 2021. Testing the motivational effects of attainable role models: Field and experimental evidence. *Journal of Theoretical Social Psychology* 5, 4 (2021), 591–602.

[29] Rúben Gouveia, Evangelos Karapanos, and Marc Hassenzahl. 2015. How do we engage with activity trackers? a longitudinal study of Habito. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '15)*. Association for Computing Machinery, New York, NY, USA, 1305–1316. doi:10.1145/2750858.2804290

[30] James Hardy. 2006. Speaking clearly: A critical review of the self-talk literature. *Psychology of sport and exercise* 7, 1 (2006), 81–97.

[31] Antonis Hatzigeorgiadis, Nikos Zourbanos, Evangelos Galanis, and Yiannis Theodorakis. 2011. Self-talk and sports performance: A meta-analysis. *Perspectives on Psychological Science* 6, 4 (2011), 348–356.

[32] Hal E Hershfield, Daniel G Goldstein, William F Sharpe, Jesse Fox, Leo Yeykelis, Laura L Carstensen, and Jeremy N Bailenson. 2011. Increasing saving behavior

through age-progressed renderings of the future self. *Journal of marketing research* 48, SPL (2011), S23–S37.
[33] Hayeon Jeong, Heepyung Kim, Rihun Kim, Uichin Lee, and Yong Jeong. 2017. Smartwatch wearing behavior analysis: a longitudinal study. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 1, 3 (2017), 1–31.
[34] Brennan Jones, Yan Xu, Qisheng Li, and Stefan Scherer. 2024. Designing a Proactive Context-Aware AI Chatbot for People's Long-Term Goals. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–7.
[35] Stephen Kaplan. 1995. The restorative benefits of nature: Toward an integrative framework. *Journal of environmental psychology* 15, 3 (1995), 169–182.
[36] Jieun Kim and Hayeon Song. 2024. My voice as a daily reminder: self-voice alarm for daily goal achievement. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–16.
[37] Amanda Lazar, Christian Koehler, Theresa Jean Tanenbaum, and David H. Nguyen. 2015. Why we use and abandon smart devices. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Osaka, Japan) *(UbiComp '15)*. Association for Computing Machinery, New York, NY, USA, 635–646. doi:10.1145/2750858.2804288
[38] Thomas C. Leonard. 2008. Richard H. Thaler, Cass R. Sunstein, Nudge: Improving decisions about health, wealth, and happiness. *Constitutional Political Economy* 19, 4 (Aug. 2008), 356–360. doi:10.1007/s10602-008-9056-2 Publisher: Springer.
[39] Tica Lin, Rishi Singh, Yalong Yang, Carolina Nobre, Johanna Beyer, Maurice A Smith, and Hanspeter Pfister. 2021. Towards an understanding of situated ar visualization for basketball free-throw training. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.
[40] Hazel Markus and Paula Nurius. 1986. Possible selves. *American psychologist* 41, 9 (1986), 954.
[41] Alexandre Mazeas, Martine Duclos, Bruno Pereira, and Aïna Chalabaev. 2022. Evaluating the effectiveness of gamification on physical activity: systematic review and meta-analysis of randomized controlled trials. *Journal of medical Internet research* 24, 1 (2022), e26779.
[42] E. McAuley, T. Duncan, and V. V. Tammen. 1989. Psychometric properties of the Intrinsic Motivation Inventory in a competitive sport setting: a confirmatory factor analysis. *Research Quarterly for Exercise and Sport* 60, 1 (March 1989), 48–58. doi:10.1080/02701367.1989.10607413
[43] Moritz Mödinger, Alexander Woll, and Ingo Wagner. 2022. Video-based visual feedback to enhance motor learning in physical education—a systematic review. *German journal of exercise and sport research* 52, 3 (2022), 447–460.
[44] Thekla Morgenroth, Michelle K Ryan, and Kim Peters. 2015. The motivational theory of role modeling: How role models influence role aspirants' goals. *Review of general psychology* 19, 4 (2015), 465–483.
[45] Yugo Nakamura, Rei Nakaoka, Yuki Matsuda, and Keiichi Yasumoto. 2023. Eat2pic: an eating-painting interactive system to nudge users into making healthier diet choices. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 1 (2023), 1–23.
[46] Luke F Olsson, Hanna L Glandorf, James F Black, Rebecca EK Jeggo, Joseph R Stanford, Karla L Drew, and Daniel J Madigan. 2025. A multi-sample examination of the relationship between athlete burnout and sport performance. *Psychology of Sport and Exercise* 76 (2025), 102747.
[47] Daphna Oyserman. 2015. *Pathways to success through identity-based motivation*. OUP Us.
[48] Jeni Paay, Jesper Kjeldskov, Mikael B Skov, Lars Lichon, and Stephan Rasmussen. 2015. Understanding individual differences for tailored smoking cessation apps. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 1699–1708.
[49] Pat Pataranutaporn, Kavin Winson, Peggy Yin, Auttasak Lapapirojn, Pichayoot Ouppaphan, Monchai Lertsutthiwong, Pattie Maes, and Hal E Hershfield. 2024. Future you: a conversation with an AI-generated future self reduces anxiety, negative emotions, and increases future self-continuity. In *2024 IEEE Frontiers in Education Conference (FIE)*. IEEE, 1–10.
[50] Christoph Pörschmann. 2000. Influences of bone conduction and air conduction on the sound of one's own voice. *Acta Acustica united with Acustica* 86, 6 (2000), 1038–1045.
[51] Mashfiqui Rabbi, Min Hane Aung, Mi Zhang, and Tanzeem Choudhury. 2015. MyBehavior: automatic personalized health feedback from user behaviors and preferences using smartphones. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*. 707–718.
[52] Thomas D Raedeke. 1997. Is athlete burnout more than just stress? A sport commitment perspective. *Journal of sport and exercise psychology* 19, 4 (1997), 396–417.
[53] Thomas D Raedeke and Alan L Smith. 2001. Development and preliminary validation of an athlete burnout measure. *Journal of sport and exercise psychology* 23, 4 (2001), 281–306.
[54] Catharine H Rankin, Thomas Abrams, Robert J Barry, Seema Bhatnagar, David F Clayton, John Colombo, Gianluca Coppola, Mark A Geyer, David L Glanzman, Stephen Marsland, et al. 2009. Habituation revisited: an updated and revised description of the behavioral characteristics of habituation. *Neurobiology of learning and memory* 92, 2 (2009), 135–138.
[55] Amon Rapp and Arianna Boldi. 2024. Open issues in persuasive technologies: six HCI challenges for the design of behavior change systems. In *International Conference on Human-Computer Interaction*. Springer, 99–116.
[56] B. Resnick and L. S. Jenkins. 2000. Testing the reliability and validity of the Self-Efficacy for Exercise scale. *Nursing Research* 49, 3 (2000), 154–159. doi:10.1097/00006199-200005000-00007
[57] Richard M. Ryan. 1982. Control and information in the intrapersonal sphere: An extension of cognitive evaluation theory. *Journal of Personality and Social Psychology* 43, 3 (1982), 450–461. doi:10.1037/0022-3514.43.3.450 Place: US Publisher: American Psychological Association.
[58] Richard M Ryan and Edward L Deci. 2000. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American psychologist* 55, 1 (2000), 68.
[59] Richard M. Ryan, Valerie Mims, and Richard Koestner. 1983. Relation of reward contingency and interpersonal context to intrinsic motivation: A review and test using cognitive evaluation theory. *Journal of Personality and Social Psychology* 45, 4 (1983), 736–750. doi:10.1037/0022-3514.45.4.736 Place: US Publisher: American Psychological Association.
[60] Alessandra Semeraro and Laia Turmo Vidal. 2022. Visualizing instructions for physical training: Exploring visual cues to support movement learning from instructional videos. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–16.
[61] Lawrence Shapiro. 2019. *Embodied cognition*. Routledge.
[62] Patrick C. Shih, Kyungsik Han, Erika Shehan Poole, Mary Beth Rosson, and Jackson Carroll. 2015. Use and Adoption Challenges of Wearable Activity Trackers. In *Proceedings of the 2015 iConference* (Newport Beach, CA). iSchools. https://api.semanticscholar.org/CorpusID:40410597
[63] Grace Shin, Yuanyuan Feng, Mohammad Hossein Jarrahi, and Nicci Gafinowitz. 2019. Beyond novelty effect: a mixed-methods exploration into the motivation for long-term activity tracker use. *JAMIA open* 2, 1 (2019), 62–72.
[64] Young Hee Shin, Hee Jin Jang, and Nola J. Pender. 2001. Psychometric evaluation of the exercise self-efficacy scale among Korean adults with chronic diseases. *Research in Nursing & Health* 24, 1 (Feb 2001), 68–76. doi:10.1002/1098-240X(200102)24:1<68::AID-NUR1008>3.0.CO;2-C
[65] Richard F Thompson and William A Spencer. 1966. Habituation: a model phenomenon for the study of neuronal substrates of behavior. *Psychological review* 73, 1 (1966), 16.
[66] Xiaoyi Tian, Zak Risha, Ishrat Ahmed, Arun Balajiee Lekshmi Narayanan, and Jacob Biehl. 2021. Let's talk it out: A chatbot for effective study habit behavioral change. *Proceedings of the ACM on human-computer interaction* 5, CSCW1 (2021), 1–32.
[67] Jan Van Looy, Cédric Courtois, Melanie De Vocht, and Lieven De Marez. 2012. Player Identification in Online Games: Validation of a Scale for Measuring Identification in MMOGs. *Media Psychology* 15, 2 (2012), 197–221. doi:10.1080/15213269.2012.674917
[68] Xiaofan Yang, Ding Pan, et al. 2020. In the face of negative data, the effects of goal type and feedback type on the willingness to continue to participate quantified-self. *American Journal of Industrial and Business Management* 10, 02 (2020), 327.

A System Implementation

A.1 Pre-recorded Scripts for Voice Cloning

Following prior work [23], we adapted the scripts to the local context. To ensure natural expression and emotional authenticity, participants read the corresponding version translated into their native language.

Read the following sentences with a happy tone:

- Hi there! Isn't the weather nice today? The sunshine is so bright, and I feel completely energized!
- I've got some great news! I received the job offer I've been dreaming of, and I'm so happy I can't stop smiling!
- Wow, this cake is just too delicious! You have to try it. It might be the best I've ever had!
- Guess what? We're going to the beach this weekend! I just can't wait any longer!

Read the following sentences with a sad tone:

- What should I do? I can't believe it… my pet ran away. I've checked everywhere but still couldn't find it.

- I've been feeling down all day today. I really wish things would get better soon.
- This morning I got a call saying I didn't pass the exam. I've felt awful the whole day.
- Why every time I try so hard, the results are still disappointing? Sometimes it's really difficult to stay optimistic.

A.2 Prompt for Audio Content Generation

Following prior work [23], we adapted the scripts to the local context. To ensure natural expression and emotional authenticity, participants read the corresponding version translated into their native language. The prompt template used in our study is provided below:

```
SYSTEM_PROMPT = (
   Your task is to imagine yourself as the person
   with these trait personalities would say to
   themselves in the given scenario that would
   encourage them to keep up with the habit. You
   should try to embody the person when their
   habit has become their identity. Use the template
   of "I am a xxx person".
   You will also be given some settings to finetune
   the emotional affect of sentence: positivity,
   emotional expressivity. The default value is
   0 and the range is -3 to +3.
   The personalities have more priority than the
   settings.
   Requirements: You must express the attitudes
   and emotions saliently. You can add vocal bursts,
   natural vocal inflections and discourse markers.
   Never use markdowns or emojis. Use first-person.
   Keep the response short within four sentences.)
USER_PROMPT = (
   Scenario: SCENARIO,
   Ideal self (5 words): IDEAL_SELF,
   Positivity: POSITIVITY;
   Emotional expressivity: EXPRESSIVITY)
```

B User Questionnaires

B.1 Demographics

- Age
- Gender
- Weekly exercise frequency
- Main type of exercise
- Average duration per workout
- Habit of exercising at home
- Prior experience with AI-synthesized content (Question: "Have you ever watched or used AI-synthesized video or voice content?")
- Self-rated understanding of AI-synthesized content (Scale: −5 = unfamiliar, 5 = familiar)

B.2 Daily Audio Generation Questionnaire

- Please describe what kind of goal you would like to achieve in your exercise today.
- Please use five words to describe your ideal self.
- How positive would you like today's generated audio to be? (Scale: −3 = negative, 3 = positive)
- How intense would you like the emotional expression in today's generated audio to be? (Scale: −3 = low intensity, 3 = high intensity)

B.3 Intrinsic Motivation Inventory (IMI)

All questions were rated on a 7-point Likert scale, with items marked † included in the shortened daily questionnaire.

Interest/Enjoyment:

- I enjoy this activity very much.†
- This activity did not hold my attention at all.
- I would describe this activity as very interesting.†
- I found this activity quite enjoyable.
- While I was doing this activity, I was thinking about how much I enjoyed it.

Perceived Competence:

- I think I did pretty well at this activity.†
- After doing this activity for a while, I felt quite competent.†
- I am satisfied with my performance in this task.
- I was quite skilled at this activity.
- This is an activity that I am not very good at.

B.4 Exercise Self-Efficacy Scale (ESES)

All questions were rated on a 4-point Likert scale, with items marked † included in the shortened daily questionnaire.

- If I try hard enough, I can overcome the barriers and challenges I encounter in physical activity and exercise.†
- I am able to find ways and means to engage in physical activity and exercise.
- I can achieve the physical activity and exercise goals I set for myself.†
- When I face barriers to physical activity or exercise, I can find multiple solutions to overcome them.
- I can engage in physical activity or exercise even when I feel tired.†
- I can engage in physical activity or exercise even when I feel down.†
- I can engage in physical activity or exercise even without the support of my family or friends.
- I can engage in physical activity or exercise even without the help of a therapist or a coach.†
- I can motivate myself to restart physical activity or exercise even after taking a break for a while.†
- I can engage in physical activity or exercise even if I don't have access to a gym, exercise facilities, or rehabilitation centers.

B.5 Video/Audio Identification Questionnaire (VAIQ)

All questions were rated on a 7-point Likert scale, with items marked † included in the shortened daily questionnaire.

Perceived Similarity:

- (Audio) The AI-generated voice is like me in many ways.†
- (Audio) The AI-generated voice resembles me.

- (Audio) I identify with the AI-generated voice.†
- (Audio) I feel that this AI voice is very close to my own personality and temperament.
- (Video) The AI-generated avatar in the video is like me in many ways.†
- (Video) The AI-generated avatar in the video resembles me.
- (Video) I identify with the AI-generated avatar.†
- (Video) I feel that this AI avatar is very close to my own personality and temperament.

Embodied Presence:

- (Video) When I see the video generated with the AI avatar, I feel like I am the person in the frame.
- (Video) Looking at the AI-generated me, I feel as if I have entered the character's body.
- (Video) While watching this video, I feel as if I have become one with the avatar.†
- (Video) I feel that the body in the AI video is my own body.

Wishful Identification:

- (Video) I wish I could be more like this AI-generated avatar.
- (Video) This AI avatar represents my ideal self.†
- (Video) The AI-generated me is a "better me."
- (Video) This AI avatar has characteristics that I wish I had.

C Post-study Interview

C.1 Experience Rating Questions

Participants were asked to rate the following items on a 7-point scale (Scale: 1 = lowest possible degree, 7 = highest possible degree).

- How physically demanding was it to complete the daily exercise tasks throughout the entire study period?
- Did you feel the overall pace of the study protocol (including daily exercise, logging, and submission) was too compact or rushed?
- To what extent did you feel the study protocol aligned with your personal exercise goals and circumstances?
- Overall, how satisfied were you with your experience throughout the study period?
- Please answer the version of the following question that applies to your group:
  - (Video) To what extent did the AI videos help you better understand and adhere to the exercises?
  - (Control) To what extent did the researchers' assistance help you complete the exercises?
- Did you find it difficult to adapt to the content and pace of the study?
- Was it difficult to complete the daily exercise and submission tasks during the study?
- What was your main motivation for participating in the study at the beginning? If you were to rate this motivation on a scale of 1-7, what score would you give? Did this motivation change over time?
- How significant a role did the monetary reward play for you? If there were no reward, would you still have persisted?
- Did you experience any periods of fatigue, numbness, or boredom? If so, how frequently did this occur (e.g., how many days a week)?
- If there were similar studies or long-term exercise programs in the future, how likely would you be to participate again?

C.2 Open Questions

- What factors do you think influenced your motivation curve? Can you recall the reasons for each rise and fall? [*Show the participant their objective and subjective charts*]
- Why do you think those rises and falls occurred?
- What impact, if any, has this experience had on forming a fitness habit or a new life rhythm for you?
- (For participants who dropped out) At what moment did you decide to stop? How were you feeling at that time?
- (Video) Do you usually watch fitness videos? If so, what types do you mainly watch?
- (Video) What differences did you notice between the AI-generated videos and regular fitness videos? What aspects did you like or dislike?
- (Video) During the study, were there any moments when you felt like quitting, similar to your past experiences with exercise? If so, what were the reasons, and do you feel the AI video had any motivational effect on you?

C.3 Qualitative Coding Procedure

To analyze the semi-structured exit interviews from both Study 1 and Study 2, we followed Braun and Clarke's six-phase reflexive thematic analysis approach [8]. Our goal was to understand participants' subjective experiences with AI self-modeling, clarify the mechanisms underlying performance changes, and find design implications grounded in participants' reflections.

All interviews were audio-recorded, fully transcribed, and anonymized. Transcripts were then segmented into meaning units and imported into a qualitative coding environment.

Initial Coding: two researchers independently conducted open coding, identifying segments segments that (a) had same or highly similar point raised by multiple participants, or (b) were insightful or informative for understanding AI self-modeling and its usage for long-term nudging. Codes remained close to participants' language.

Theme Development: Codes were iteratively clustered into broader themes. Because participants discussed both learning-related changes (e.g., posture awareness, technique refinement) and motivational dynamics (e.g., self-efficacy, identity cues, novelty), these mechanisms were intentionally separated. During this process, one recurring theme concerned how participants approached their daily goals. Some aimed for steady improvement, others for maintaining consistency, and others fluctuated with external constraints. These patterns were subsequently consolidated as the three "user archetypes" described in design implication (Section 6.2), emerging directly from repeated codes rather than being predefined categories.

Cross-Study Analysis: the analyses of Study 1 and Study 2 followed the same high-level structure but were allowed to diverge when long-term patterns appeared. Study 1 predominantly reflected short-term responses such as modality–task alignment or initial

fidelity impressions, whereas Study 2 introduced new themes related to longitudinal dynamics (e.g., early acceleration, stabilization, internalization).

Thematic Consolidation for Design Implications. Themes identified across studies informed the interpretation of behavioral change stages, clarified mechanisms behind performance changes, and grounded the design implications, such as the role of representational fidelity and the emergence of user archetypes. Representative quotes were selected to illustrate each theme and ensure transparency in how findings were derived.

D Data Integrity and Attrition Details

D.1 Attrition Analysis: Tables and Survival Curves

This appendix details the statistical procedures used to validate the integrity of the dataset. We examined attrition from three perspectives: (1) Proportions: whether dropout rates differed by group or cohort; (2) Timing: whether dropout occurred earlier in specific groups; and (3) Characteristics: whether those who dropped out differed systematically from those who remained.

D.1.1 Dropout Contingency Tables and Fisher Exact Test. Table 4 summarizes the number of participants who completed or dropped out of the study across each *Cohort*×*Group* combination. "Dropped" denotes participants who did not finish all 28 study days.

To test whether attrition rates differed systematically, we applied Fisher exact tests (appropriate given the small cell counts). Results showed no evidence of differential dropout: for $Group \times Dropout$, $OR = 1.02$, $p > 0.050$; for $Cohort \times Dropout$, $OR = 2.10$, $p > 0.050$, indicating that neither cohort membership nor intervention condition was associated with higher or lower dropout probability.

Table 4: Dropout counts by cohort and group.

Cohort	Group	Completed	Dropped
1	Control	6	3
1	Video	5	4
2	Control	2	4
2	Video	4	4

D.1.2 Kaplan–Meier Survival Curve and Log-Rank Test. We next examined whether the *timing* of attrition differed across cohorts using Kaplan–Meier survival analysis, where "survival" indicates remaining active in the study (i.e., not yet dropped out). Figure 11 shows the survival functions for Cohort 1 and Cohort 2. The two curves track each other closely, suggesting no noticeable separation in dropout timing.

We further used a log-rank test to statistically compared the survival distributions. For both the VSM group ($\chi^2 = 0.05$, $p = 0.820$) and the Control group ($\chi^2 = 1.07$, $p = 0.301$), the log-rank tests indicated no significant differences between Cohort 1 and Cohort 2. These results further support that attrition timing was not cohort-dependent.

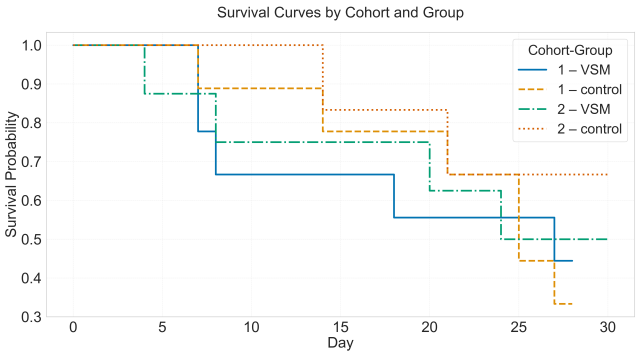


Figure 11: Kaplan–Meier survival curves for Cohort 1 and Cohort 2 over the 28-day study period.

D.2 Survivor Bias Check

To assess survivor bias, we performed independent t-tests comparing baseline metrics between Completers and Dropouts. The detailed statistical results are presented in Table 5.

Table 5: Comparison of baseline characteristics between Completers and Dropouts within VSM and Control groups. (Mean ± SD)

	VSM		
Metric	**Completed**	**Dropped**	***t*-value**
IMI-Interest/Enjoyment	4.87 (0.89)	4.71 (1.13)	0.39
IMI-Perceived Competence	4.90 (1.00)	4.48 (1.29)	0.92
ESES	2.99 (0.38)	3.18 (0.57)	-1.00
Crunch	44.10 (31.60)	46.00 (16.60)	-0.18
Wall-Sit	117.90 (53.70)	130.00 (68.40)	-0.49
	Control		
Metric	**Completed**	**Dropped**	***t*-value**
IMI-Interest/Enjoyment	4.62 (0.68)	3.40 (0.51)	4.76^{**}
IMI-Perceived Competence	4.78 (0.77)	4.09 (0.72)	2.18^{*}
ESES	3.08 (0.37)	3.02 (0.47)	0.37
Crunch	35.80 (14.40)	36.30 (14.80)	-0.09
Wall-Sit	108.10 (62.70)	108.10 (72.10)	0.00

Note: Positive t-values indicate Completers > Dropouts.