

Article

Dual-Constrained Co-Planning of Virtual Health Avatars and User Interfaces for Older Adults with Chronic Diseases

Xi Zhang ¹, Hongyu Zhang ¹ and Yongtao Zheng ^{2,*}

¹ College of Architecture and Art Design, Hebei University of Technology, Tianjin 300401, China; zhangxi@hebut.edu.cn (X.Z.); 202322304013@stu.hebut.edu.cn (H.Z.)

² School of Advanced Manufacturing, Guangdong University of Technology, Guangzhou 510000, China

* Correspondence: jimzheng@gdut.edu.cn

Abstract

Digital health systems for older adults with chronic diseases must support task completion, risk communication, and emotional engagement. However, virtual avatars and user interfaces are usually generated separately, which can weaken the consistency between avatar behavior, interface hierarchy, and task risk. This study proposes a dual-constrained co-planning method that integrates continuous valence–arousal–dominance affective targets with task engineering constraints, including functional priorities and age-friendly boundaries. User needs are translated into three representative health-management tasks and encoded in a multi-source constraint graph. A two-stream conditional Transformer then jointly plans avatar behaviors and interface layouts, followed by constraint checking and automatic rollback. Across the three evaluated health-management scenarios, the framework produced task-sensitive changes in avatar expression, gaze, gesture, and interface priority and supported rule-based checking and rollback of predefined age-friendly constraints. Exploratory external comparisons and ablation analyses provided preliminary evidence regarding the roles of affective constraints, graph-based task representation, and cross-stream coordination in prototype-level affective matching, usability, constraint compliance, and avatar–UI coordination. The framework is intended as a structured approach for verifiable early-stage prototype generation and candidate screening rather than for direct clinical deployment.

Keywords: affective computing; age-friendly interaction; generative design; graph neural networks; user interface co-planning; virtual health assistants



Academic Editor: Hyokyung Bahn

Received: 6 August 2026

Revised: 23 September 2026

Accepted: 25 September 2026

Published: 30 September 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Digital health services for older adults with chronic diseases are evolving from single-purpose recording tools into systems that combine health monitoring, medication reminders, risk alerts, and ongoing guidance [1–3]. Virtual health avatars can improve social presence and make instructions easier to follow through facial expression, gaze, posture, and gesture [4,5]. However, these benefits depend on whether the avatar’s behavior is consistent with the information and actions presented by the interface. Older users may also experience visual, motor, or cognitive difficulties that make small text, low contrast, dense information, and long interaction paths particularly problematic [6–8]. Designing the avatar and interface as separate outputs may therefore weaken both usability and task clarity. Their generation should instead be coordinated according to task risk, expected emotional tone, functional priority, and age-friendly requirements.

Generative AI supports health interaction prototyping, but its direct use in chronic-disease care for older adults remains risky. Prompts can generate emotional avatars or complete UI images, yet they cannot stably control font size, contrast, hotspot size, information density, or critical paths. Structured GUI methods improve layouts, but rarely link avatar identity, expression, gaze, and gestures with key UI functions. Although related generative techniques have advanced [9–14], design still needs to move from result generation to planning-layer constraint organization. Existing studies present three research gaps. First, avatar and UI generation are often modeled separately. Second, age-friendly requirements are checked after generation, not used as pre-generation constraints. Third, cross-modal generation under-models emotional goals, task structures, age-friendly bounds, and avatar cues.

To address these gaps, this paper formulates avatar-UI generation as a dual-constrained co-planning problem. VAD constraints describe emotional tone, reminder intensity, and control relation. Task engineering constraints define functional priorities and predefined age-friendly bounds. A multi-source constraint graph represents task modules, spatial relations, operation sequences, and avatar attention anchors. The system jointly plans avatar behavior and UI layout, then forms a validated prototype through identity-preserving rendering, UI rendering, constraint checking, and rollback.

The main contributions of this paper are as follows.

First, a dual-constraint representation method is proposed for chronic-disease management tasks for older adults. This method jointly transforms VAD emotional goals, functional module priorities, and age-friendly hard boundaries into computable conditions.

Second, a planning-layer avatar-UI co-generation method is proposed. It uses a multi-source constraint graph and a two-stream conditional Transformer to jointly generate avatar behavior parameters and UI layout parameters.

Third, a hierarchical evaluation protocol is constructed. It analyzes the co-generation performance from several aspects, including first-pass output, automatic rollback, final validation results, external baselines, and ablation studies.

2. Related Work

Generative interaction design covers continuous emotion modeling, structured GUI generation, and virtual-agent interaction, improving emotional control, interface constraint modeling, and avatar behavior generation. However, few studies encode emotional goals, health tasks, age-friendly boundaries, and avatar behavioral cues together for chronic-disease management in older adults.

2.1. Task-Oriented Generation with Emotional Control

Emotion representation is the basis of controllable generation. Early studies used discrete labels such as happy, sad, or fearful, but these labels cannot express intensity variation, mixed affective states, or cross-scene transfer. Recent work has shifted toward continuous emotion spaces [15]. EmoGen shows that diffusion models can encode abstract emotions [16]. EmotiCrafter injects continuous valence-arousal values into text-to-image generation for fine-grained control [17]. For interface design, Chen et al. used a valence-arousal model to generate static interfaces with different emotional tones and validated perceptibility through check-all-that-apply (CATA) and eye-tracking experiments [15].

Continuous dimensions describe both emotion type and intensity. Yet chronic-disease management for older adults also involves perceived control [18]. Daily records require user-led control, medication reminders require moderate interruption, and anomaly alerts require system-led guidance. The VA space cannot fully represent these control relations [19]. Therefore, this paper adopts VAD, using dominance to characterize user-led

versus system-led control. Existing work targets general images or static interfaces [15–17], without binding emotional conditions with health tasks, age-friendly boundaries, and avatar-UI co-generation.

2.2. Structurally Constrained and Age-Friendly Interface Generation

GUI generation increasingly treats interfaces as structured objects integrating semantics, spatial layout, and logical relations. GUILGET, Graph4GUI, and ShowUI use GUI arrangement graphs, graph neural networks, and connected UI graphs, respectively, to model layout constraints, element attributes, semantic-spatial relations, and vision–language–action localization [20–23]. These studies demonstrate the value of graph representations, but mainly emphasize layout rationality, visual consistency, and element relations, while task priorities and safety boundaries remain underexplored.

For older adults with chronic diseases, interface quality also depends on critical-path visibility, cognitive load, information density, and risk-alert effectiveness. Anomaly warnings need high visual priority and clear action paths; medication reminders should reduce irrelevant interference; routine recording should fit within one screen. Thus, this paper maps functional modules, logical priorities, spatial adjacency, and attention-directing relations into graph nodes and edges. Since age-friendly design shows that font size, contrast, hotspot size, and information density can be modeled during design, as demonstrated by AdaptForge, a model-driven approach evaluated with both software developers and senior end users [24], this paper parameterizes them as hard acceptance constraints and integrates them with functional modules and spatial relations in a multi-source constraint graph.

2.3. Digital Human-Interface Co-Generation in Health Interaction

Health reviews show that embodied conversational agents and conversational agents are widely used in chronic-disease support, self-management, and personalized intervention, with effectiveness affected by appearance, animation, communication style, and personalization. Avatars provide task guidance, social presence, and emotional support. Matching avatar appearance to message emotion enhances eHealth persuasion [25], while reinforcement-learning adaptation can improve emotional engagement [26].

Existing studies have advanced identity-preserving generation, behavior adaptation, and attention alignment [26–28], but these functions are often modular. In chronic-disease health management, identity, expression, gaze, gesture, functional modules, and critical paths should follow shared task semantics and safety bounds. This paper integrates avatar identity, expression, posture, and gaze with key UI functional areas in a unified condition space. A conditional Transformer, shared condition vectors, and a multi-source constraint graph support joint planning of avatar emotion, attention guidance, UI function, and age-friendly constraints. Overall, prior work supports continuous emotion modeling [15–17], structured GUI generation [21–24], identity preservation and attention alignment [26–28], and health-agent design [2,3,25]. However, planning-layer avatar-UI co-generation remains limited. This paper addresses this gap by modeling co-generation as a dual-constrained structured planning process.

3. Method

This section presents the proposed dual-constrained avatar-UI co-generation method. The method jointly models task emotional goals, functional module priorities, and age-friendly boundaries at the planning layer. It then generates avatar behavior sequences and UI layout sequences.

3.1. Problem Formulation and User Task Modeling

Health management needs of older adults with chronic diseases often involve multiple layers of information. These include task goals, emotional states, and usage constraints. Such information needs to be transformed into computable dual-constraint inputs before generation.

The user study in this paper is used to define the task space and constraint boundaries for the subsequent generation model. Interviews and contextual probes are adopted to identify typical tasks, trigger mechanisms, cognitive load, and risk differences in chronic-disease management for older adults. The SAM scale is used to convert users' expectations of system expression styles into continuous emotional constraints. Functional priorities and age-friendly boundaries are then used to form executable task engineering constraints. Through this process, the originally scattered user narratives are translated into structured conditions that can be fed into the co-planning module.

Figure 1 shows the user requirement modeling workflow in this study. First, raw user data are collected through participant recruitment, storyboard-based contextual probes, semi-structured interviews, and the SAM scale. Then, the workflow splits into two analytical paths. The first path is the task engineering constraint extraction path. It focuses on task attributes, user pain points, contextual features, functional module requirements, and rigid age-friendly boundaries. This path is used to form task types, functional module priorities, and verifiable engineering boundaries. The second path is the emotional constraint modeling path. It focuses on users' subjective expectations regarding the virtual agent's expression style, reminder intensity, and control relationship. Through SAM-VAD mapping and clustering, this path obtains task-level emotional targets. Finally, the two paths converge at the structured output layer. Together, they form a condition set that can be input into the subsequent collaborative generation module. This condition set includes task types, VAD emotional constraints, and task engineering constraints. Here, task engineering constraints consist of functional module priorities and rigid age-friendly boundaries.

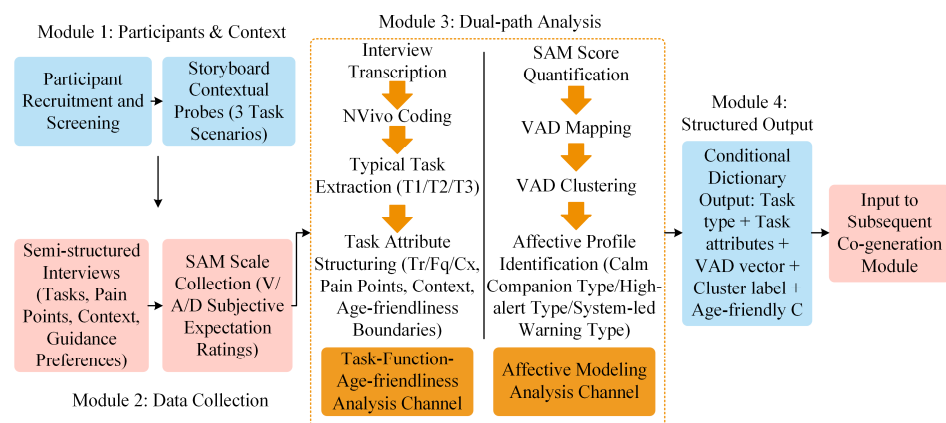


Figure 1. User requirement modeling pipeline. Colors distinguish workflow components: light blue indicates participant/context and structured output elements, pink indicates data collection/input elements, and orange indicates dual-path analysis components.

3.1.1. Participant Recruitment, Data Collection, and Typical Task Extraction

User requirement modeling involved 50 older adults with chronic diseases. Participants were recruited through community-based convenience sampling from the researchers' local social networks and residential communities surrounding the university. Inclusion criteria were age ≥ 60 years, hypertension or diabetes for at least one year, and prior smartphone experience. Exclusion criteria were severe visual or motor difficulty that prevented participation in assisted interviews, or inability to understand the task scenarios and ques-

tionnaires. All participants provided written informed consent, participation was voluntary, and participants could withdraw at any time without adverse consequences. Visual and fine-motor limitations were assessed by participant self-report using the four-level items in Questionnaire S1 rather than by clinical examination. The visual item assessed difficulty reading smartphone text while wearing usual glasses, and the motor item assessed hand tremor or difficulty with fine motor control. Sixteen participants reported moderate or severe difficulty reading smartphone text, and 14 reported moderate or severe hand tremor or fine-motor difficulty. No data were missing for the characteristics summarized in Supplementary Table S1. No formal a priori power analysis was used; the cohort size was selected for task and requirement modeling across the three predefined health-management scenarios rather than for population estimation. Cohort characteristics are summarized in Supplementary Table S1. These reported difficulties informed the later age-friendly constraints on font size, contrast, touch-target size, and information density.

Before formal interviews, three storyboards served as visual probes to standardize task recall and scenario description. The health values, medication-reminder content, and abnormal-alert situations used in the study were simulated research scenarios rather than individualized clinical instructions. No treatment decisions or medication changes were made on the basis of the study tasks or prototype outputs. They covered morning measurement and recording, medication reminders during distractions, and abnormal results when alone at home. These probes reduced differences from verbal descriptions and improved contextual relevance. Some older adults refined the storyboards, but formal participants did not join research question design or indicator selection. Semi-structured interviews collected requirements in four areas: smartphone-based health tasks, physical and attentional contexts, common difficulties or mistouches, and expectations for system and avatar expression under different risk levels. Transcripts were analyzed using thematic analysis. Two HCI researchers independently performed open and axial coding for task types, pain points, and contextual features. Agreement was high for task-type coding (Cohen's $\kappa = 0.85$) and pain-point coding (Cohen's $\kappa = 0.81$). Disagreements were reviewed against transcripts and, if unresolved, arbitrated by a third HCI researcher.

SAM scores were linearly transformed into VAD values for task-level affective modeling. Based on coding results, tasks were described by trigger mechanism, occurrence frequency, and operational load. Three tasks were extracted: T1 routine health data recording, T2 medication reminder, and T3 abnormal alert. They represent active recording, passive reminder, and urgent intervention, supporting later age-friendly bounds, VAD modeling, and task-level generation conditions.

3.1.2. Task Engineering Representation of Functional Priorities and Age-Friendly Constraints

After the typical tasks were identified, task semantics, functional priorities, and age-friendly requirements were translated into task engineering constraints. Functional-module priorities, derived from participant rankings and expert evaluation, were used as soft layout constraints, while font size, contrast, information density, target size, and critical-path length were specified as prototype-stage acceptance constraints. Contrast thresholds followed WCAG 2.2, with 4.5:1 corresponding to SC 1.4.3 and 7:1 to SC 1.4.6 [29]. The remaining acceptance criteria, including 32 px for ordinary targets, 44 px for critical CTAs, and the limits on font size, information blocks, and path length, were determined from the user-needs study and pilot evaluation, with the font-size criterion further informed by prior age-friendly design research [30].

The age-friendly acceptance vector C_{age} contains the minimum font size C_{font} , minimum contrast $C_{contrast}$, maximum information-block count $C_{density}$, minimum target size C_{target} , and maximum critical-path steps C_{steps} . Together with the functional weights, it

forms the task engineering constraints passed to the collaborative generation stage with the VAD conditions.

All font-size and target-size constraints were evaluated on the normalized 1080×1920 reference canvas. These pixel-based thresholds were defined on the prototype canvas and may require adjustment when implemented on devices with different display characteristics.

Text contrast was calculated using the WCAG 2.2 relative-luminance method [29]. For the foreground and background colors, the contrast ratio was computed as follows:

$$CR = \frac{L_{\text{lighter}} + 0.05}{L_{\text{darker}} + 0.05'} \quad (1)$$

where L_{lighter} and L_{darker} denote relative luminance of the lighter and darker colors, respectively. Relative luminance was derived from linearized sRGB values according to the WCAG definition, with $L = 0.2126R + 0.7152G + 0.0722B$. The resulting contrast ratio is dimensionless and was compared with the task-specific threshold in Table 1.

Table 1. Task-level generation conditions derived from user requirement modeling.

Task	Task Context/User Need	Representative Emotional Target	Functional Weight Vector [w_1, w_2, w_3, w_4, w_5]	Age-Friendly Hard Acceptance Constraints	Basis/Source
T1 Routine recording	User-initiated routine recording; difficulty recalling values and reading dense/small-text interfaces.	Medium positive valence, low arousal, high dominance; supportive and user-led interaction	[0.34, 0.28, 0.08, 0.20, 0.10]	Main text ≥ 24 px; contrast $\geq 4.5:1$; ordinary target ≥ 32 px; information blocks ≤ 3 ; interaction steps ≤ 2	Font/target/blocks/steps: user-needs study + pilot; contrast: WCAG 2.2 SC 1.4.3.
T2 Medication reminder	Scheduled reminder under divided attention; risk of missed reminders or medication confusion.	Medium positive valence, medium arousal, medium dominance; clear but non-alarming guidance	[0.10, 0.08, 0.42, 0.24, 0.16]	Persistent reminder until acknowledgement; main text ≥ 24 px; contrast $\geq 4.5:1$; ordinary target ≥ 32 px; information blocks ≤ 3 ; interaction steps ≤ 2	Reminder /font/target/blocks/steps: user-needs study + pilot; contrast: WCAG 2.2 SC 1.4.3.
T3 Abnormal data alert	High-risk abnormal alert; panic, uncertainty, and accidental touch under urgency.	Slightly positive valence, high arousal, low dominance; system-led single-path guidance	[0.08, 0.06, 0.08, 0.15, 0.63]	Main text ≥ 24 px; contrast $\geq 7.0:1$; information blocks ≤ 2 ; critical CTA size ≥ 44 px; interaction steps ≤ 1	Font/CTA/blocks/steps: user-needs study + pilot; contrast: WCAG 2.2 SC 1.4.6.

Note. w_1 = weight of the data input area; w_2 = weight of the trend/history area; w_3 = weight of the medication information area; w_4 = weight of the guidance/help area; w_5 = weight of the emergency action area.

Thus, functional priorities and age-friendly acceptance constraints are jointly encoded in the task engineering representation. Table 1 summarizes the task-level conditions and the basis of the numerical thresholds.

For the high-risk alert task, limiting the interface to no more than two primary information blocks and one critical interaction step was intended to reduce competing visual information and minimize the action path during time-sensitive decisions; these limits were operational design choices informed by the needs study and pilot evaluation rather than externally standardized thresholds.

The 44-px criterion defines the minimum visual target extent on the normalized reference canvas used for the static prototypes. Device-level implementation should account for the display and interaction characteristics of the target hardware.

3.1.3. Task-Level Quantification of Affective Goals and VAD Clustering

Because the three health-management tasks differed in emotional tone, reminder intensity, and perceived control, users' affective expectations were represented using a continuous valence-arousal-dominance (VAD) vector. The Self-Assessment Manikin (SAM)

scale [31] was used to obtain valence, arousal, and dominance ratings under the T1 routine-recording, T2 medication-reminder, and T3 abnormal-alert scenarios.

Let the valence, arousal, and dominance scores of participant p under task t be denoted as $S_{v,p,t}$, $S_{a,p,t}$, and $S_{d,p,t}$, respectively. The corresponding continuous VAD vector $E_{pt} = [V_{p,t}, A_{p,t}, D_{p,t}]$ is then defined as follows:

$$V_{p,t} = \frac{S_{v,p,t} - 5}{4}, A_{p,t} = \frac{S_{a,p,t} - 5}{4}, D_{p,t} = \frac{S_{d,p,t} - 5}{4}. \quad (2)$$

where the transformed VAD values represent participant-level affective coordinates used for task-level clustering. Higher valence indicates a more positive tone, higher arousal indicates stronger reminder or warning intensity, and lower dominance indicates a greater need for system guidance.

K-means clustering was performed separately for each task using participants' three-dimensional VAD coordinates. Candidate solutions with $K \in \{1, 2, 3\}$ were compared using SSE, silhouette coefficients, and task interpretability. $K = 2$ was retained for all three tasks because it provided the best balance between compactness and interpretability; the corresponding silhouette coefficients were 0.49, 0.43, and 0.48 for T1, T2, and T3, respectively. The resulting cluster centers and design interpretations are reported in Table 2.

Table 2. Task-specific VAD cluster centers and design interpretation.

Task	Cluster	n	V (Mean $\pm$ SD)	A (Mean $\pm$ SD)	D (Mean $\pm$ SD)	Design Interpretation
T1 Routine recording	C1	35	0.31 ± 0.16	0.19 ± 0.11	0.79 ± 0.10	Calm, supportive, user-led daily management
	C2	15	0.17 ± 0.12	0.30 ± 0.10	0.35 ± 0.13	Mild reassurance needed, but still non-urgent
T2 Medication reminder	C1	34	0.24 ± 0.17	0.38 ± 0.15	0.57 ± 0.12	Clear but non-alarming reminder with moderate guidance
	C2	16	0.16 ± 0.13	0.45 ± 0.14	0.09 ± 0.13	Stronger interruption needed for distracted users
T3 Abnormal alert	C1	28	0.10 ± 0.17	0.84 ± 0.12	-0.45 ± 0.16	High alert, system-led guidance, reduced user control
	C2	22	0.16 ± 0.12	0.67 ± 0.12	0.11 ± 0.17	Alert but still seeking partial control and information transparency

Table 2 shows task-specific affective patterns. T1 is characterized by positive valence, low arousal, and high dominance; T2 by moderate arousal and dominance; and T3 by high arousal with greater variation in dominance. The T3 clusters indicate different preferences for system-led guidance versus retained user control, supporting the inclusion of dominance in the affective representation. These task-level VAD patterns were subsequently used as conditions for collaborative generation.

3.2. Dual-Constraint Unified Representation

Task types, VAD goals, and task engineering bounds are organized into a unified representation for collaborative generation.

3.2.1. Interpretable Mapping from VAD to Generation Parameters

The VAD vector provides planning-level control rather than direct pixel-level control. Valence modulates expression and interface tone, arousal modulates gaze, gesture, posture, and alert salience, and dominance regulates the balance between user-led and system-led in-

teraction. Representative mappings are provided in Supplementary Figure S1. Continuous VAD targets are then encoded as affective latent conditions for collaborative planning.

3.2.2. Joint Conditional Encoding and Constraint Prioritization

Affective constraints, task engineering constraints, and identity anchors are encoded into a unified conditional representation shared by the multi-source constraint graph and dual-stream Transformer. Affective constraints represent task-level emotional goals, task engineering constraints specify module priorities and age-friendly bounds, and identity anchors preserve avatar consistency across tasks. Given the target vector $E_t = [V_t, A_t, D_t]$, we project it into an affective latent vector using a two-layer multilayer perceptron (MLP):

$$e_t^{emo} = f_{MLP}(E_t) \quad (3)$$

where e_t^{emo} denotes the affective latent vector. A two-layer MLP maps the 3D VAD input to a 128-dimensional latent representation through a 64-dimensional hidden layer.

Task engineering constraints combine functional weights and age-friendly bounds. Five interface modules are considered: data input m_1 , historical trend m_2 , medication information m_3 , help section m_4 , and emergency operation m_5 . Here, w_i denotes the functional weight assigned to module m_i . The functional weight vector for task t is defined as follows:

$$w_{func}^{(t)} = [w_1, w_2, w_3, w_4, w_5], \sum_{i=1}^5 w_i = 1 \quad (4)$$

Participant rankings were converted into task-specific module-priority scores, with ranks 1, 2, and 3 coded as 3, 2, and 1, respectively, and unranked modules coded as 0. For module i in task t , the participant-derived raw score was calculated as $U_{t,i}^{raw} = 3n_{1,t,i} + 2n_{2,t,i} + n_{3,t,i}$ and normalized by the maximum possible score across 50 participants. Three evaluators with industrial-design backgrounds independently assigned each module an overall 0–5 priority score based on task necessity, interaction sequence, and risk relevance; disagreements were discussed before the final expert scores were averaged. Participant- and expert-derived priorities were then combined using a fixed 0.60/0.40 weighting, after which softmax normalization was applied to obtain the final functional-weight vector. The task-specific rank counts, intermediate participant and expert scores, combined scores, and final normalized weights are reported in Supplementary Table S2.

For each task t , the age-friendly constraint vector c_t^{age} contains the task-specific bounds for font size, contrast, information density, target size, and interaction steps, as listed in Table 1. The functional priorities and age-friendly bounds are then combined into the task-engineering condition, which is further integrated with the affective latent representation and identity anchor:

$$c_t^{eng} = [w_t^{func}; c_t^{age}], z_t^{cond} = [e_t^{emo}; c_t^{eng}; z^{id}]. \quad (5)$$

where c_t^{eng} represents the task-engineering condition, and z_t^{cond} is the joint conditional representation used by the subsequent graph-encoding and dual-stream planning modules. The identity anchor z^{id} preserves avatar identity across task scenarios, whereas e_t^{emo} and c_t^{eng} provide the affective and task-engineering conditions, respectively.

3.3. Avatar-UI Collaborative Generation Architecture

Figure 2 summarizes the proposed avatar-UI co-generation framework. Task type, VAD target, functional weights, age-friendly bounds, and identity anchors are encoded into a joint condition and multi-source constraint graph. A GAT encoder and dual-stream conditional Transformer jointly predict avatar behavior and structured UI layouts, which

are then rendered and checked against predefined accessibility and alignment constraints. Noncompliant outputs are revised through bounded rollback. The VAD encoder, GAT encoder, and dual-stream Transformer are trainable, whereas rendering, constraint checking, and rollback are execution modules.

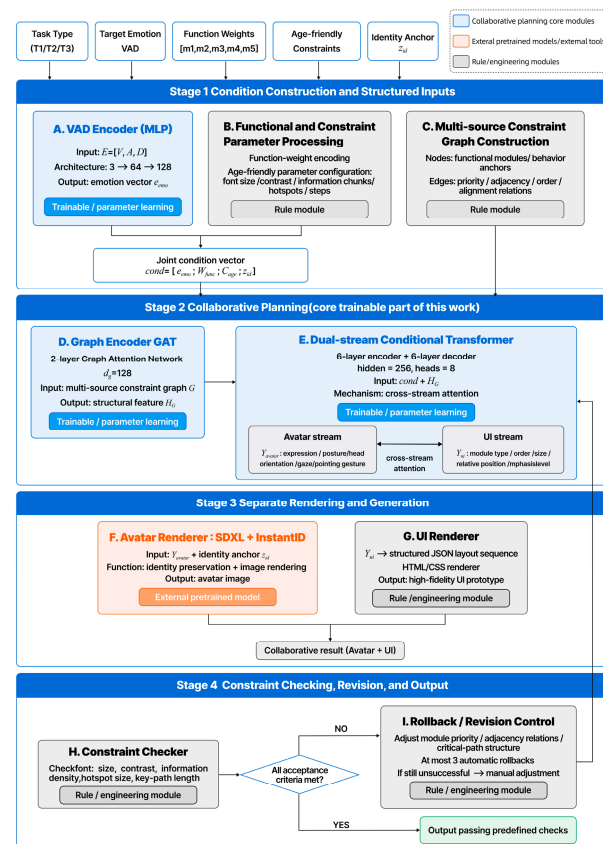


Figure 2. System boundary and module division of the proposed framework.

3.3.1. Formal Definition of the Multi-Source Constraint Graph

The multi-source constraint graph represents UI modules, avatar behavioral anchors, spatial relations, task sequences, and cross-modal alignment. For sample n of task $t \in \{1,2,3\}$, the multi-source constraint graph is defined as follows:

$$G_t^n = (V_t^n, E_t^n, X_t^n, R_t^n, C_t^n) \quad (6)$$

where V_t^n and E_t^n denote the node and edge sets, X_t^n and R_t^n denote node and edge attributes, and C_t^n denotes graph-level conditions. The node set contains UI functional modules and avatar behavioral anchors. The edge set contains priority, adjacency, sequence, and attention-alignment relations. Table 3 summarizes the corresponding roles and fields.

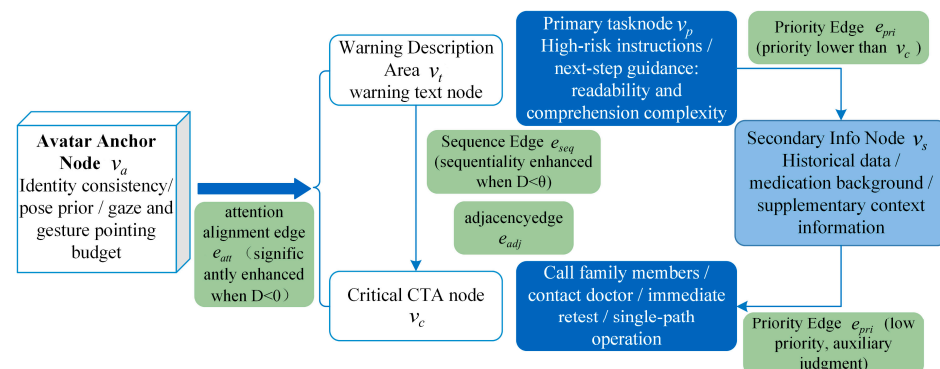
The node feature x_i consists of node type, module category, semantic attributes, normalized bounding boxes, age-friendly attributes, task-level VAD conditions, and gaze/gesture anchor features. Edge feature r_{ij} consists of edge type, direction or order, spatial distance, and hard-constraint labels. Graph-level conditions C_t^n consist of affective latent vectors, functional weights, age-friendly boundaries, and identity anchors.

GAT encoding produces graph context H_G , which provides structural information to the dual-stream Transformer. The attention weights characterize the relative importance of graph relations during encoding.

Table 3. Definitions of nodes, edges, and key fields in the multi-source constraint graph.

Component	Role	Main Fields
UI/task node	Represents interface modules for data input, trend/history information, medication information, guidance/help, and emergency actions.	node type; module id; priority; bbox; font; contrast; target size; key-path step
Avatar node	Represents the identity anchor, facial expression, head orientation, gaze landing point, and gesture target of the virtual health assistant.	identity anchor; expression; head pose; gaze target; gesture target
Priority edge	Represents the display priority of task-related modules.	edge type; source; target; weight; hard flag
Adjacency edge	Represents spatial adjacency or grouping relationships between interface modules.	gap; distance; overlap; group id
Sequence edge	Represents the step order in the main task path.	step index; max steps; next action
Alignment edge	Represents the alignment between avatar gaze/gesture and key UI regions or CTA components.	gaze/gesture target; key region; CTA id

Figure 3 illustrates graph instantiation for the T3 abnormal-alert task. Under its high-arousal, low-dominance condition, emergency-action priority, task sequence, and attention-alignment relations are strengthened, while secondary information is assigned lower priority.

**Figure 3.** Multi-source constraint graph instance for the abnormal-alert task.

3.3.2. Graph Encoding, GAT Attention Interpretation, and Dual-Stream Collaborative Planning

A two-layer graph attention network (GAT) encodes the multi-source constraint graph G into the structural representation $H_G \in \mathbb{R}^{N \times d_g}$, with $d_g = 128$. The graph representation H_G and joint conditional representation z_t^{cond} are used as shared inputs to the two planning streams:

$$\begin{aligned} Y_{avatar} &= \text{Transformer}(H_G, z_t^{cond}; \text{avatar stream}), \\ Y_{ui} &= \text{Transformer}(H_G, z_t^{cond}; \text{UI stream}). \end{aligned} \quad (7)$$

Y_{avatar} predicts expression, posture, gaze, and gesture parameters, whereas Y_{ui} predicts module type, order, geometry, and salience. The six-layer encoder–decoder Transformer uses hidden dimension 256 and 8 attention heads. Cross-stream attention at each decoding layer allows avatar tokens to access key UI structure and UI tokens to access avatar directive, gaze, and gesture states.

At decoding layer l , let T_a^l and T_u^l denote the hidden states of the digital human stream and UI stream, respectively. The cross-stream attention is formulated as follows:

$$Z_{a \leftarrow u}^l = \text{Attn}(T_a^l, T_u^l, M_{a \leftarrow u}), Z_{u \leftarrow a}^l = \text{Attn}(T_u^l, T_a^l, M_{u \leftarrow a}) \quad (8)$$

The masks $M_{a \leftarrow u}$ and $M_{u \leftarrow a}$ are derived from attention-alignment and key-path graph edges, enabling mutual constraint between avatar behavioral cues and key UI regions.

In the graph encoding process, GAT attention weights $a_{ij}^{(l,k)}$ have more explicit interpretations. High weights on priority edges indicate that the model relies more on these modules to determine interface hierarchy and visual salience. High weights on adjacency edges indicate that the model relies more on spatial grouping relations to control information density. High weights on sequence edges indicate greater reliance on these edges to compress key interaction paths. High weights on attention-alignment edges indicate that gaze or gesture tokens should remain aligned with corresponding UI key regions or CTA elements. Accordingly, gaze and gesture alignment was assessed using planning-level target relations, while coordination consistency was evaluated separately as an expert-rated measure of avatar-UI correspondence. These measures can also be traced to planning-level evidence from edge types, edge attributes, and cross-stream token interactions.

Figure 4 illustrates token-level cross-stream interaction: avatar gaze and gesture tokens access key UI regions, while UI tokens access avatar directive and attention states to adjust CTA salience and hierarchy.

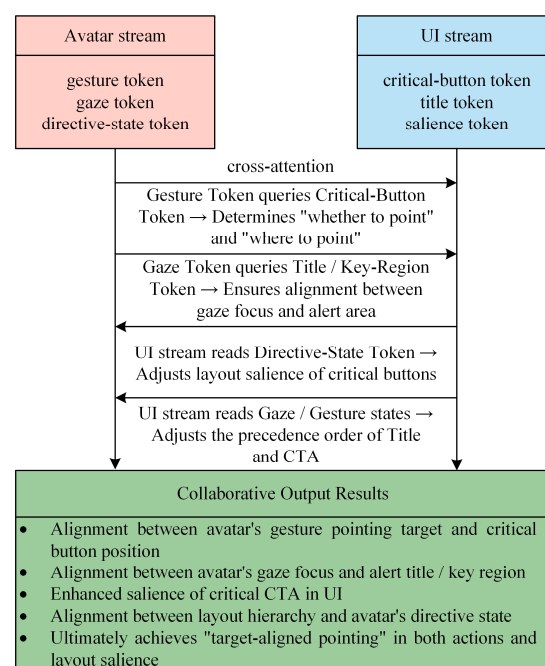


Figure 4. Cross-stream attention mechanism between the avatar stream and the UI stream.

The conditional Transformer outputs mid- and high-level planning results, whereas pixel-level images and high-fidelity interfaces are generated by backend rendering modules. For the digital humans, Y_{avatar} is converted into structured prompts or parameter sequences and combined with identity anchor z^{id} as input to SDXL+InstantID, thereby preserving identity consistency across task scenarios. For interfaces, Y_{ui} is represented as a structured JSON component tree and rendered into high-fidelity prototypes using an HTML/CSS renderer. Figma auto-layout is used only for presentation, whereas quantitative evaluation is based on the structured component data. In the present examples, all layouts used a 1080×1920 portrait canvas as the reference coordinate system, and all stimuli were

presented at a standardized display size and resolution. The component tree contains screen metadata, component identifiers, module type and role, bounding boxes, z-order, text and font attributes, foreground/background colors, contrast ratio, visibility and interactivity flags, target dimensions, path-step/action fields, and graph edges defined by source, target, and relation. Font size, contrast ratio, target dimensions, information-block identifiers, and key-path steps are read directly by the rule-based constraint checker rather than estimated from the rendered bitmap. These properties are checked against the predefined task-specific constraints before and after rendering.

3.3.3. Training Data, Loss Function, and Rendering Implementation

Each sample corresponds to a participant-task base case and its controlled augmented variants. Inputs include task type, VAD vector, functional weights, age-friendly bounds, identity anchors, and a multi-source constraint graph. Outputs include expert-reviewed digital-human behavior, a UI component tree, avatar-UI alignment, and hard-constraint pass status. Behavior parameters cover expression, posture/head orientation, gaze, and gestures. UI parameters cover component type, order, bbox, hierarchy, font size, contrast, hotspot, visibility, and key-path steps. The total loss combines the following:

$$L = 1.00L_{ui} + 0.75L_{aff} + 0.55L_{align} + 0.65L_{constraint} \quad (9)$$

where L_{ui} penalizes errors in UI component type, position, size, and visual attributes; L_{aff} penalizes inconsistency between digital-human expression and the VAD condition; L_{align} penalizes misalignment between gaze/gesture targets and key UI regions; $L_{constraint}$ penalizes violations of the font-size, contrast, information-density, hotspot-size, and key-path-length criteria. Coordinate errors are normalized by the canvas diagonal, whereas font-size, contrast, hotspot-size, and path-length penalties are normalized by their task-specific thresholds. Final hard-constraint pass/fail status is determined separately by the rule checker.

The trainable planning modules were optimized with AdamW using an initial learning rate of 8×10^{-5} , a batch size of 32, and a maximum of 180 epochs, with an early-stopping patience of 18 epochs. The loss weights for L_{ui} , L_{aff} , L_{align} , and $L_{constraint}$ were 1.00, 0.75, 0.55, and 0.65, respectively. Full software, architecture, optimization, seed, and rendering settings are provided in Supplementary Table S3, while controlled data construction and participant-level splitting are described below.

The 50 participants contributed 150 participant-task base cases across T1, T2, and T3. Each base case was paired with one expert-reviewed reference target and nine controlled variants, yielding 1500 condition-planning pairs.

Each base case corresponded to one participant under one task and contained the task condition, VAD target, functional weights, age-friendly bounds, identity anchor, and constraint graph. The corresponding reference target was prepared by the authors based on the predefined task conditions and the evaluators' design recommendations. The same three evaluators, including two graduate students and one faculty member, independently checked the avatar-behavior plan and UI structure. Disagreements were discussed before the reference target was finalized.

Nine predefined controlled variants were constructed for each base case by modifying prespecified variables within allowed ranges while keeping task identity and core constraints fixed. The complete variant definitions are provided in Supplementary Table S4.

Participant-level splitting was performed before variant construction, and all task cases and derived variants from the same participant remained within the same training, validation, or test split. Derived variants from the same participant were therefore not treated as additional independent participants. This yielded 1200 training pairs from

40 participants, 150 validation pairs from five participants, and 150 held-out test pairs from five participants. Constraint/rollback evaluation samples were generated after training and managed separately. Table 4 summarizes the dataset split and evaluation samples.

Table 4. Dataset split and evaluation samples.

Subset	Participants/Source	Samples	Purpose
Training set	40 participants	1200 pairs	Model training
Validation set	5 participants	150 pairs	Early stopping and model selection
Held-out test set	5 participants	150 pairs	First-pass internal evaluation
Constraint/rollback evaluation set	Generated after training	90 outputs, 30 per task	Constraint recovery and qualitative/boundary-case analysis

Note. Outputs used for the external baseline comparison and ablation analysis were generated separately after model training and were not included in the 90-output constraint/rollback evaluation set.

Table 5 reports first-pass performance on the participant-level held-out test set before rollback or manual correction. Position and size errors are computed on the 1080×1920 reference canvas, while gaze and gesture errors are derived from structured planning outputs. Here, landing denotes the predicted intersection of the gaze ray or gesture vector with the UI plane. Thus, both metrics measure planning-layer alignment.

Table 5. Held-out test performance of the co-planning model.

Metric	T1	T2	T3	Overall
Layout accuracy, module type	0.91 ± 0.04	0.89 ± 0.05	0.87 ± 0.06	0.89 ± 0.05
Position MAE (px)	32.40 ± 12.10	36.80 ± 14.30	41.20 ± 16.50	36.80 ± 14.76
Size MAE (px)	28.60 ± 10.40	31.20 ± 11.80	34.50 ± 13.20	31.43 ± 12.02
Gaze landing error (px)	14.20 ± 6.80	18.50 ± 8.20	22.10 ± 10.40	18.27 ± 9.13
Gesture-CTA center distance (px)	24.60 ± 11.20	28.30 ± 13.50	31.80 ± 15.70	28.23 ± 13.82
First-pass constraint pass	32/50, 64.0%	29/50, 58.0%	19/50, 38.0%	80/150, 53.3%

3.4. Constraint Checking and Rollback Mechanism

Age-friendly requirements are treated as task-specific hard acceptance constraints on font size, contrast, information-block count, target/hotspot size, and key-path length. Here, “hard constraint” denotes a rule-checker acceptance criterion and does not imply that every first-pass generative output is intrinsically violation-free. After rendering, any detected violation triggers updates to the graph structure or condition parameters, followed by replanning and rerendering.

Font-size violations triggered text scaling and local space expansion; contrast violations triggered foreground–background adjustment; information-density violations triggered compression of secondary content; undersized targets triggered CTA enlargement and local space redistribution; and excessive key-path length triggered branch compression. Up to three automatic rollback rounds were allowed; remaining failures underwent manual correction and re-checking. Constraint recovery results are reported in Section 4.2.2, with output-level rollback records provided in Supplementary Data S5.

Algorithm 1 summarizes the bounded automatic rollback procedure over the constraint graph, avatar parameters, and Y_{ui} component tree. Local revisions address predefined age-friendly and avatar–UI alignment criteria. Processing stops when all criteria are satisfied, the rollback limit is reached, or no available action further reduces the violations; unresolved outputs then require manual correction and re-checking.

Algorithm 1. Bounded dual-constraint co-planning and rollback search for avatar-UI design

Input: task type t , VAD vector E_t , functional weight vector w_t^{func} , age-friendly constraint vector c_t^{age} , identity anchor z^{id} ; maximum rollback $K_{max} = 3$.

Search space: $S = \{G, Y_{avatar}, Y_{ui}\}$, where G is the multi-source constraint graph, Y_{avatar} is the structured avatar behavior sequence, and Y_{ui} is the UI JSON/component tree.

Action set: $A = \{increase_{font}, adjust_{contrast}, reduce_{density}, enlarge_{target}, shorten_{path}, reorder_{priority}, update_{salience}, redirect_{gaze}, redirect_{gesture}, revise_{alignment_edge}\}$.

Termination: stop if all hard acceptance constraints pass, if $k = K_{max}$, or if no feasible action reduces the current violation vector.

Output: The current avatar image I_{avatar} and UI layout I_{ui} , a check status indicating either “passed” or “requires manual review”, the remaining constraint violations, and a log of first-pass checks, rollback actions, and the stopping reason.

1. Encode E_t into e_t^{emo} , and combine w_t^{func} and c_t^{age} into c_t^{eng} , and construct the joint conditional representation z_t^{cond} according to Equation (5);
2. Instantiate G from the task template, node/edge schema and condition dictionary;
3. Use GAT to obtain graph context H_G and initialize the planning state with G and z_t^{cond} ;
4. Generate first-pass Y_{avatar} and Y_{ui} using the dual-stream conditional Transformer;
5. Render Y_{ui} into a component-level prototype and compute violation vector v_0 with the constraint checker;
6. Set $k = 0$. While v_k is not empty and $k < K_{max}$, select the highest-priority violation according to task risk and hard-boundary severity;
7. Choose action a_k from A , update the affected graph nodes/edges or UI components, and regenerate only the affected planning tokens when possible;
8. Re-render the revised Y_{ui} , recompute v_{k+1} , record selected action, affected node IDs, rollback count and stop reason, then set $k = k + 1$;
9. If no constraint violations remain, set the check status to “passed”. Otherwise, set it to “requires manual review” and retain the unresolved violations;
10. Return I_{avatar} , I_{ui} , the check status, the remaining violations, and the recorded history.

4. Experiments

4.1. Experimental Design and Evaluation Protocol

The experimental protocol comprised user-needs modeling, collaborative generation, sample construction, subjective evaluation, external baseline comparison, ablation analysis, an additional user evaluation with a separate participant cohort, and statistical analysis. Constraint and rollback analysis used 30 initial samples per task, while identity consistency was evaluated across three tasks using 10 rendering seeds. Subjective evaluation involved the 50 participants from the needs-modeling cohort, and the additional user evaluation included a separate cohort of 30 older adults. The overall data-construction and evaluation workflow is summarized in Supplementary Figure S2.

For the additional user-evaluation cohort, formal educational level, digital-literacy scores, cognitive-screening results, and medication-regimen data were not collected. Smartphone experience and self-reported visual and fine-motor difficulty were therefore the available indicators of technology experience and functional limitation. The external baseline comparison used the T1 routine health-data recording condition as a common reference scenario and followed each method’s native input format. This was treated as a capability-specific descriptive comparison rather than a fully input-matched benchmark. Each baseline generated 10 first-pass outputs, with no automatic rollback or manual correction before evaluation. The statistical procedures are described below. Baseline configurations are summarized in Table 6, and implementation provenance, checkpoints, inference settings, and study-specific input adaptations are provided in Supplementary Table S10.

Table 6. Input configurations and applicable metrics for the T1 external comparison.

Method Category	Methods	T1 Input Configuration	Identity Input	Structured UI Input	Applicable Metrics
General text-to-image	SDXL	T1 task prompt, calm and supportive tone, large text, high contrast, simple layout, data input, trend viewing, and help information	None	None	Affective matching, functional usability, age-friendliness, and applicable first-pass constraint violations; identity consistency: N/A.
Affect-conditioned generation	EmoGen; EmotiCrafter	T1 task prompt plus discrete emotion label or corresponding VA condition	None	None	
Identity-preserving generation	PhotoMaker; ConsistentID	T1 task prompt plus the shared reference identity	Shared identity image	None	Affective matching, identity consistency, functional usability, age-friendliness, and applicable first-pass constraint violations.
Structured GUI generation	GUILGET; Graph4GUI; LACE	T1 module list and simplified spatial constraints	N/A	Module and spatial constraints	Functional usability and age-friendliness; affective matching, identity consistency, and first-pass hard-constraint violations: N/A.
Proposed framework	Ours	T1 task type, VAD target, functional weights, age-friendly hard acceptance constraints, identity anchor, and multi-source constraint graph	Shared identity image	Constraint graph and UI layout renderer	All applicable metrics.

Each baseline received the maximum task-relevant information supported by its native input format. Text-to-image methods used natural-language prompts, affect-conditioned methods their supported emotion inputs, identity-preserving methods the shared reference identity, and structured GUI methods task modules and spatial relations. Configurations were fixed before evaluation, and only first-pass outputs were compared. Full prompts and settings are provided in Supplementary Note S1.

Evaluation included four metric groups: subjective ratings (affective matching, usability, age-friendliness, and coordination), objective constraints, mechanism metrics (coordination and gaze/gesture alignment), and task-performance measures (completion time, mistouch rate, time to first correct CTA selection, and NASA-TLX). Constraint compliance was reported at first-pass, post-rollback, and final stages; held-out tests used first-pass compliance only, whereas the 90 rollback samples were evaluated at all three stages. Three HCI evaluators who were not involved in model development independently rated the outputs. Method identities were withheld, files were anonymized, sample order was randomized, and identical presentation settings were used. Evaluators reviewed a common rating rubric before scoring; no separate calibration or consensus session was conducted. The generated output was the primary analysis unit, and the median of the three ratings was used for each subjective dimension. Evaluator-level ratings were retained only for inter-rater reliability. Two-way absolute-agreement, average-measures ICC(A,3) values were 0.88 for affective matching, 0.84 for functional usability, 0.89 for age-friendliness, and 0.91 for overall coordination [32]. Label-masking and rating procedures are described in Supplementary Note S2 and Questionnaire S2.

Repeated-measures task comparisons used the Friedman test [33], followed by Wilcoxon signed-rank tests with Holm correction [34]; Kendall's W was reported as the omnibus effect size. Identity consistency was computed with $buffalo_l$ from the official InsightFace model package. All generated and reference images were first processed by the SCRFD face detector. Detected faces were then aligned and cropped to 112×112 RGB images. Next, 512-dimensional identity embeddings were extracted and L_2 -normalized. Let $\hat{z}^r$ be the normalized embedding of the reference identity image, and $\hat{z}_{i,k}^s$ be that of

the generated result under task t and random seed k . The identity similarity is defined as follows:

$$s_{t,k}^{(r)} = \cos(\hat{z}_r, \hat{z}_{t,k}) = \hat{z}_r^T \hat{z}_{t,k} \quad (10)$$

This metric characterizes identity stability of the same virtual health assistant across tasks and random seeds. Identity consistency was not evaluated for GUILGET, Graph4GUI, or LACE, as they output UI layouts rather than human images. This study reports three levels: external baselines, rollback, and final validation. The T1 first-pass comparison examined responses to affective targets, functional structure, age-friendly bounds, and identity conditions. This layered report separates first-pass capability, recovery cost, and final validated output. External baselines assess input control channels. The proposed method adds affective targets, task structure, age-friendly bounds, and identity anchors, and is evaluated by affective matching, usability, compliance, and avatar-UI behavioral consistency.

For the ablation analysis, the three seed-level outputs were first averaged within each held-out condition, yielding 30 condition-level observations per configuration and task. One mechanism-aligned full-ablation comparison was prespecified for each ablation. Wilcoxon signed-rank tests were conducted at the condition level, with $r = |Z|/\sqrt{30}$ reported as the effect size and paired-bootstrap 95% confidence intervals obtained from 10,000 resamples. Holm adjustment was applied to the five mechanism-aligned comparisons within each task.

For the additional user evaluation, the participant was the unit of analysis. Because the same participants completed T1, T2, and T3, overall task differences were assessed using the Friedman test with Kendall's W as the effect size. Significant omnibus effects were followed by pairwise Wilcoxon signed-rank tests with Holm adjustment across the three task comparisons within each metric.

4.2. Main Comparative Results

This section analyzes the generated results at the level of collaborative schemes. The proposed method is compared across the three health-management tasks. Each scheme includes both virtual-human behavior and the corresponding UI layout. This design supports analysis of how task conditions affect facial expression, posture, gaze direction, and priority of UI modules.

4.2.1. Typical Collaborative Cases Across the Three Tasks

For T1 routine health data recording, the results showed preference for a low-pressure, user-led style. The digital human used a gentle expression and small guiding gestures. The UI focused on the (m_1) data-entry area and (m_2) historical-trend area, retained necessary (m_4) instruction/help information, and did not highlight the less relevant (m_3) medication-information area or (m_5) emergency-action area. For T2 medication reminders, the results emphasized reminder recognizability and confirmation accessibility. The digital human had a clear reminding posture and offered moderate guidance. The UI centered on (m_3) and (m_4), while retaining a confirmation entry. Medication name, dose, time, and confirmation actions were more prominent than trend information. In T3 abnormal data alert, the results focused on risk warning and next-step guidance. The UI highlighted (m_5), compressed noncritical information, and shortened the interaction path. Overall, task risk and control relation shaped both digital-human cues and UI structure. Figure 5 illustrates representative avatar-UI co-generation results for T1, T2, and T3, showing that task risk and control relation jointly shape digital-human behavior, UI module priority, and age-friendly constraint satisfaction.

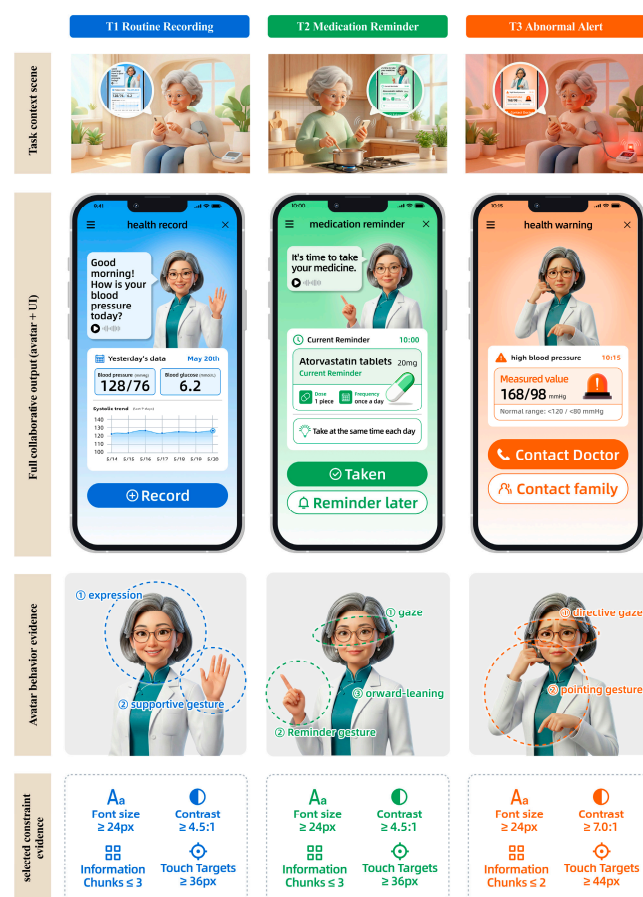


Figure 5. Core collaborative case display across three representative tasks.

Identity consistency was evaluated using 30 generated avatar images across three tasks and 10 random seeds per task. All outputs were compared with the same reference identity anchor using InsightFace cosine similarity. Mean similarity was 0.911 ± 0.005 for T1, 0.894 ± 0.005 for T2, and 0.878 ± 0.004 for T3, with individual values ranging from 0.8704 to 0.9181 (Figure 6).

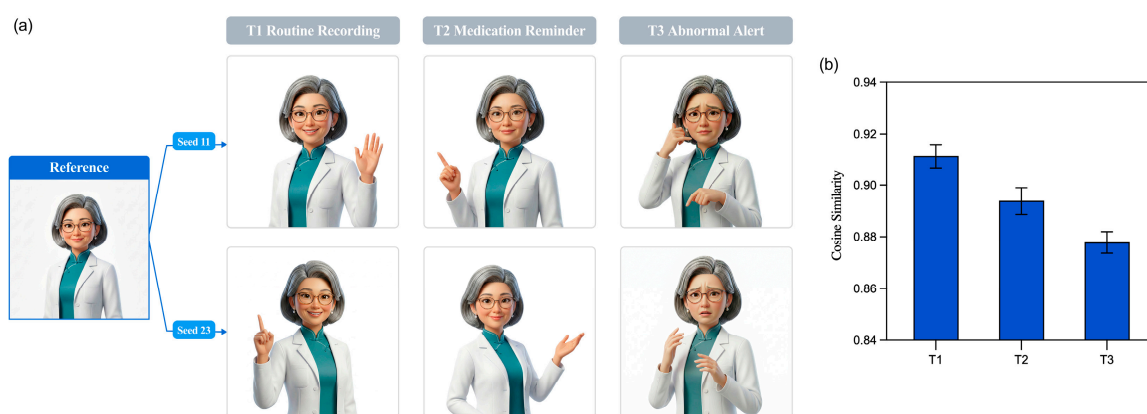


Figure 6. Identity consistency across task conditions and random seeds. (a) Representative outputs generated from the same reference identity anchor under T1 routine recording, T2 medication reminder, and T3 abnormal-alert conditions using two representative random seeds (11 and 23). (b) Reference-anchor cosine similarity across the three tasks. Bars represent mean $\pm$ SD across 10 random seeds per task.

4.2.2. Constraint Compliance and Rollback

Constraint compliance was evaluated on 30 samples per task. Sample-level first-pass compliance was 63.3%, 56.7%, and 36.7% for T1–T3, respectively, increasing to 93.3%, 86.7%, and 80.0% after up to three automatic rollback rounds. Constraint-level recovery and residual manual corrections are summarized in Table 7, with output-level rollback records provided in Supplementary Data S5.

Table 7. Constraint-level compliance and recovery across T1–T3 ($n = 30$ per task).

Task	Constraint	First-Pass Compliant	After Auto Rollback	Manual Correction
T1 Routine recording	Font	27/30 (90.0%)	30/30 (100%)	0/30
	Contrast	27/30 (90.0%)	30/30	0/30
	Density	26/30 (86.7%)	29/30 (96.7%)	1/30 (3.3%)
	Target	28/30 (93.3%)	29/30 (96.7%)	1/30
	Path	27/30 (90.0%)	30/30	0/30
	Overall	19/30 (63.3%)	28/30 (93.3%)	2/30 (6.7%)
T2 Medication reminder	Font	28/30 (93.3%)	30/30	0/30
	Contrast	20/30 (66.7%)	26/30 (86.7%)	4/30 (13.3%)
	Density	26/30 (86.7%)	30/30	0/30
	Target	24/30 (80.0%)	30/30	0/30
	Path	26/30 (86.7%)	30/30	0/30
	Overall	17/30 (56.7%)	26/30 (86.7%)	4/30 (13.3%)
T3 Abnormal alert	Font	24/30 (80.0%)	30/30	0/30
	Contrast	24/30 (80.0%)	30/30	0/30
	Density	26/30 (86.7%)	30/30	0/30
	Target ≥ 44 px	22/30 (73.3%)	27/30 (90.0%)	3/30 (10.0%)
	Path ≤ 1	15/30 (50.0%)	27/30 (90.0%)	3/30 (10.0%)
	Overall	11/30 (36.7%)	24/30 (80.0%)	6/30 (20.0%)

Note. After manual correction and re-validation by the rule checker, all samples satisfied the applicable constraints (30/30 per task). Constraint-specific counts indicate the number of samples requiring manual correction for each constraint and may overlap across constraint types, whereas the Overall row reports the number of unique samples requiring at least one manual correction.

4.2.3. User Evaluation

The 50 older adults rated T1–T3 in affective matching, functional usability, age-friendliness, and overall coordination. For each task, all participants evaluated the same fixed task-level prototype stimulus. Repeated-measures comparisons used the Friedman test, followed by Wilcoxon signed-rank tests with Holm correction; Kendall's W was reported as the effect size. Table 8 reports the results.

Table 8. User evaluation scores and statistical comparisons across three tasks.

Evaluation Dimension	T1	T2	T3	Friedman $\chi^2(2)$	p -Value	Kendall's W	Holm-Corrected Post Hoc Result
Affective matching	4.22 $\pm$ 0.58	4.12 $\pm$ 0.56	3.98 $\pm$ 0.62	4.85	0.088	0.049	All ns; T1 vs. T3, $p_{\text{adj}} = 0.141$
Functional usability	4.30 $\pm$ 0.51	4.24 $\pm$ 0.52	4.48 $\pm$ 0.50	8.67	0.013	0.087	T3 > T2, $p_{\text{adj}} = 0.008$; others ns
Age-friendliness	4.44 $\pm$ 0.50	4.38 $\pm$ 0.49	4.66 $\pm$ 0.48	12.54	0.0019	0.125	T3 > T1, $p_{\text{adj}} = 0.023$; T3 > T2, $p_{\text{adj}} = 0.001$
Overall coordination	4.44 $\pm$ 0.50	4.34 $\pm$ 0.48	4.50 $\pm$ 0.51	2.97	0.227	0.03	All pairwise comparisons ns

Significant task differences were observed in functional usability and age-friendliness. T3 was rated higher than T2 for functional usability and higher than both T1 and T2 for age-friendliness. Affective matching and overall coordination did not differ significantly across tasks.

4.2.4. Additional User Evaluation with a Separate Participant Cohort and Comparison with External Methods

A separate cohort of 30 older adults with chronic diseases was recruited using the same community-based convenience sampling procedure and eligibility criteria as the needs-modeling cohort; cohort characteristics are reported in Supplementary Table S1. All 30 participants completed the three-task evaluation, and no task-level observations were missing. Participants received no financial compensation. This cohort was separate from the needs-modeling cohort but was evaluated within the same research project.

Each participant completed T1–T3 in random order in a within-participant design, yielding 90 task observations, with the participant as the analysis unit. Before timed evaluation, participants received standardized task instructions but no formal task-specific training or practice trials. No formal a priori power analysis was performed for this exploratory additional user evaluation; the cohort was used to provide repeated-measures prototype-level evaluation across the three predefined tasks. For T3, “Contact Doctor” was predefined as the target CTA to standardize task completion and CTA-selection measurements; this evaluated target localization rather than medical decision appropriateness. Behavioral measures included completion time, participant-level mistouch rate, and time to first correct CTA selection. Subjective workload was assessed immediately after each task using raw NASA-TLX data [35], with the instrument provided in Questionnaire S3. Each of the six subscales was scored from 0 to 100 in five-point increments; Performance was anchored from 0 (Perfect) to 100 (Failure). No pairwise weighting was applied, and the raw NASA-TLX total score was calculated as the unweighted mean of the six subscales. Timing definitions, mistouch criteria, and operational task differences are summarized in Table 9.

Table 9. Operational task instructions and success criteria for the additional user evaluation.

Feature	T1: Routine Recording	T2: Medication Reminder	T3: Abnormal Alert
Required actions	Enter three values and confirm	Review medication information and confirm	Locate and select “Contact Doctor”
Screens	1, scrollable	1, scrollable	1, no scrolling
Decision points	Data entry and confirmation	Medication confirmation	Action selection
Key CTA initially visible	No	Yes	Yes
Scrolling required	Yes	Yes, for medication details	No
Competing controls	Trend, help, history	History, help	“Contact family” button
Success criterion	Successful record displayed	“Taken” confirmation displayed	Selection of “Contact Doctor”

Note. For T3, “Contact Doctor” was predefined as the target CTA for standardized timing, whereas Questionnaire S1 used a broader needs-elicitation scenario. Because CTA configuration differed across tasks, CTA-selection time was treated as a task-specific descriptive measure. Mistouch rate was calculated as unintended or incorrect touches divided by total touches $\times 100\%$.

The three tasks therefore differed in action count, scrolling requirement, decision structure, and CTA visibility. These operational differences were considered when interpreting between-task timing and error measures.

Table 10 and Figure 7 show significant overall task differences in completion time, mistouch rate, and raw NASA-TLX total score. CTA-selection time is reported descriptively only because CTA visibility, scrolling requirements, competing controls, and decision structure were not matched across tasks. Pairwise Wilcoxon signed-rank tests with Holm

correction for the inferentially compared metrics are provided in Supplementary Table S5. Raw NASA-TLX total scores were higher in T3 than in T1 and T2, with a smaller difference between T1 and T2. These results reflect workload differences across the three evaluated scenarios rather than the effect of any single task or interface factor. For descriptive reference, the T1/T2/T3 means were 49.8/54.3/68.7 for mental demand, 38.0/35.8/39.7 for physical demand, 45.2/49.8/68.7 for temporal demand, 50.2/50.3/50.3 for performance, 57.0/64.3/72.7 for effort, and 53.8/55.7/62.5 for frustration. The largest descriptive increases in T3 were observed for mental demand, temporal demand, effort, and frustration, whereas physical demand and performance were comparatively similar across tasks. Participant-level subscale scores are provided in Supplementary Data S2.

Table 10. Additional user evaluation across T1–T3 ($n = 30$).

Metric	T1	T2	T3	Friedman $\chi^2(2)$	Exact p	Kendall's W
Completion time (s)	188.7 $\pm$ 68.2; 196.2 [153.0, 251.8]	136.3 $\pm$ 52.6; 148.1 [110.7, 171.3]	94.6 $\pm$ 38.1; 90.2 [71.0, 112.9]	24.87	3.98×10^{-6}	0.414
Mistouch rate (%)	9.7 $\pm$ 6.1; 9.2 [5.5, 13.8]	6.8 $\pm$ 3.9; 7.7 [3.3, 9.4]	3.9 $\pm$ 2.8; 3.9 [1.2, 6.2]	17.27	1.78×10^{-4}	0.288
CTA-selection time (s)	3.6 $\pm$ 1.8; 3.5 [2.3, 4.4]	2.8 $\pm$ 1.4; 2.7 [1.8, 3.5]	1.9 $\pm$ 0.9; 1.9 [1.2, 2.5]	-	-	-
Raw NASA-TLX total	49.0 $\pm$ 13.6; 50.0 [38.5, 59.0]	51.7 $\pm$ 12.8; 52.1 [42.5, 60.6]	60.4 $\pm$ 12.0; 60.0 [52.7, 68.3]	30.59	2.28×10^{-7}	0.510

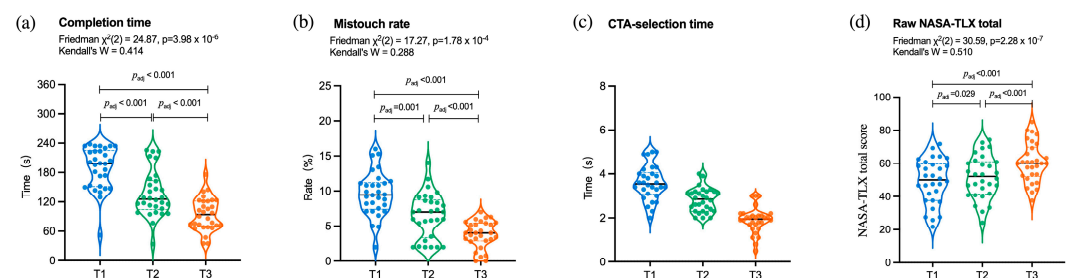


Figure 7. Participant-level performance and workload across T1–T3 in the additional user evaluation ($n = 30$): (a) completion time; (b) mistouch rate; (c) CTA-selection time; and (d) Raw NASA-TLX total score. Violin plots show participant distributions; Friedman statistics, Kendall's W , and Holm-adjusted pairwise comparisons are shown for (a,b,d), whereas (c) is descriptive only.

The T1 external comparison included 10 first-pass outputs per method. Table 11 summarizes the applicable evaluation metrics as mean $\pm$ SD. Implementation provenance, checkpoints, inference settings, and study-specific input adaptations for the external baselines are provided in Supplementary Table S10.

Among avatar-generating methods, the proposed framework had numerically higher mean T1 scores for affective matching and functional usability, while identity-consistency scores were within the range observed for identity-preserving methods. GUI-only methods were evaluated only on the interface-related dimensions applicable to their outputs. The comparison therefore reflects differences in supported capabilities rather than a unified performance ranking. Representative outputs are provided in Supplementary Figure S3.

The compared methods support different forms of conditioning. Text-to-image and affect-conditioned methods primarily use textual or affective inputs, identity-preserving methods incorporate reference identity information, and structured GUI methods focus on interface structure without avatar generation. The proposed framework combines affective, identity, functional, and accessibility conditions within a joint planning process.

First-pass constraint violations for applicable avatar-generating methods are reported in Supplementary Table S6.

Table 11. External comparison based on first-pass outputs under T1 condition ($n = 10$ per method).

Method	Affective Matching	Identity Consistency (Cosine Similarity)	Functional Usability	Age-Friendliness	Any Applicable First-Pass Hard-Constraint Violation
SDXL [36]	3.10 ± 0.74	N/A	2.70 ± 0.48	2.50 ± 0.53	10/10
EmoGen [16]	3.60 ± 0.52	N/A	2.60 ± 0.52	2.30 ± 0.82	7/10
EmotiCrafter [17]	3.80 ± 0.42	N/A	2.80 ± 0.63	2.50 ± 0.85	7/10
PhotoMaker [37]	3.40 ± 0.52	0.894 ± 0.016	2.50 ± 0.53	2.60 ± 0.52	10/10
ConsistentID [38]	3.50 ± 0.53	0.911 ± 0.025	2.90 ± 0.57	2.80 ± 0.42	9/10
GUILGET [21]	N/A	N/A	3.50 ± 0.53	3.40 ± 0.52	N/A
Graph4GUI [22]	N/A	N/A	3.60 ± 0.84	3.50 ± 0.53	N/A
LACE [39]	N/A	N/A	3.60 ± 0.52	3.70 ± 0.82	N/A
Ours	4.00 ± 0.47	0.916 ± 0.011	4.00 ± 0.47	3.70 ± 0.67	7/10

Note. Metrics are reported only where applicable to each method. GUI-only methods were evaluated for functional usability and age-friendliness only because they do not generate avatar content. Identity consistency was evaluated only for methods using a reference identity. N/A indicates that the corresponding metric was not applicable to the method.

4.3. Ablation and Coordination Analysis

Ablation experiments were conducted across T1–T3 using the same participant-level split as the main evaluation. The same 30 held-out conditions were used across configurations, and each configuration was retrained with three random seeds, yielding 90 first-pass outputs per task. Seed values are reported in Supplementary Table S3.

Five ablations were evaluated, as follows: A1 removed the VAD constraint, A2 removed the functional and age-friendly constraints, A3 removed cross-stream attention, A4 replaced the multi-source constraint graph with a flat MLP, and A5 used independent avatar/UI planning followed by late concatenation. A1–A2 tested constraint inputs, whereas A3–A5 tested coordination structures. Evaluation included the four subjective ratings, gaze and gesture alignment, and first-pass constraint pass rate.

Gaze alignment was calculated from the relation between each predicted gaze landing point and its graph-linked UI target region. For sample i , $GA_i = K_{i,g}^{-1} \sum_{j=1}^{K_{i,g}} \mathbb{I}(p_{ij}^{gaze} \in R_{ij}^{target})$, where $K_{i,g}$ denotes the number of gaze anchors. Gesture alignment was calculated analogously from the relation between each predicted gesture target and its linked CTA or key-operation region. Rates were first calculated within each sample and then averaged across evaluation conditions.

Because the number of applicable gaze and gesture anchors varied across outputs, alignment was calculated using the output-specific anchor count as the denominator before averaging across outputs. Denominator distributions by task and configuration are reported in Supplementary Table S7.

The two constraint-input ablations showed distinct patterns. A1 produced the largest reductions in affective matching, whereas A2 produced larger reductions in functional usability, age-friendliness, and first-pass constraint compliance.

Table 12 summarizes the descriptive ablation results, while mechanism-aligned full-ablation comparisons are reported in Supplementary Table S8 and output-level results are provided in Supplementary Data S4. The coordination ablations mainly affected coordination consistency and gaze/gesture alignment. A3 showed lower values for both alignment measures, A4 showed broader changes across quality and coordination metrics, and A5 showed the largest reduction in coordination consistency. The full model achieved first-pass constraint pass rates of 65.6%, 61.1%, and 41.1% for T1, T2, and T3, respectively.

The lower pass rate for T3 reflects its stricter combined requirements for contrast, target size, information density, and critical-path length and is examined further in the rollback and boundary-case analysis.

Table 12. Descriptive ablation results across T1, T2, and T3.

Task	Config.	Affective	Functional	Age-Friendly	Coordination	Gaze (%)	Gesture (%)	First-Pass Pass
T1	A1	2.43 ± 0.37	4.18 ± 0.34	4.34 ± 0.28	4.12 ± 0.30	79.8 ± 12.6	81.3 ± 15.5	45/90 (50.0%)
	A2	4.16 ± 0.36	2.93 ± 0.18	2.48 ± 0.35	4.01 ± 0.24	80.7 ± 19.6	77.8 ± 18.7	27/90 (30.0%)
	A3	4.10 ± 0.35	4.16 ± 0.26	4.23 ± 0.25	3.62 ± 0.27	70.5 ± 18.0	64.4 ± 17.4	43/90 (47.8%)
	A4	3.63 ± 0.32	3.36 ± 0.26	3.54 ± 0.30	3.87 ± 0.24	74.9 ± 16.3	73.2 ± 18.4	33/90 (36.7%)
	A5	3.53 ± 0.31	3.61 ± 0.34	3.71 ± 0.32	2.88 ± 0.32	56.1 ± 17.4	35.4 ± 19.3	47/90 (52.2%)
	Full	4.32 ± 0.33	4.31 ± 0.34	4.44 ± 0.32	4.52 ± 0.34	92.8 ± 9.1	91.5 ± 10.3	59/90 (65.6%)
T2	A1	2.41 ± 0.35	4.12 ± 0.22	4.34 ± 0.32	4.21 ± 0.31	74.5 ± 15.5	70.8 ± 15.6	42/90 (46.7%)
	A2	4.10 ± 0.34	2.84 ± 0.30	2.46 ± 0.31	3.94 ± 0.32	78.4 ± 10.4	66.6 ± 19.8	23/90 (25.6%)
	A3	3.96 ± 0.38	4.06 ± 0.26	4.24 ± 0.30	3.51 ± 0.27	69.1 ± 13.1	59.3 ± 15.3	34/90 (37.8%)
	A4	3.74 ± 0.27	3.33 ± 0.29	3.54 ± 0.31	3.77 ± 0.38	75.6 ± 15.1	70.4 ± 18.1	31/90 (34.4%)
	A5	3.37 ± 0.31	3.57 ± 0.35	3.59 ± 0.29	2.76 ± 0.33	51.1 ± 22.1	40.7 ± 21.0	37/90 (41.1%)
	Full	4.29 ± 0.26	4.31 ± 0.31	4.46 ± 0.27	4.44 ± 0.35	89.7 ± 10.7	87.4 ± 9.8	55/90 (61.1%)
T3	A1	2.32 ± 0.28	4.38 ± 0.34	4.42 ± 0.29	3.97 ± 0.34	74.1 ± 12.3	65.5 ± 17.1	33/90 (36.7%)
	A2	4.11 ± 0.34	2.81 ± 0.27	2.41 ± 0.31	3.90 ± 0.29	74.1 ± 14.1	64.8 ± 21.8	18/90 (20.0%)
	A3	4.01 ± 0.30	4.11 ± 0.24	4.28 ± 0.33	3.51 ± 0.29	62.4 ± 15.4	60.8 ± 17.1	35/90 (38.9%)
	A4	3.84 ± 0.31	3.37 ± 0.28	3.62 ± 0.29	3.70 ± 0.37	70.5 ± 15.6	64.2 ± 13.7	18/90 (20.0%)
	A5	3.41 ± 0.34	3.52 ± 0.32	3.63 ± 0.29	2.58 ± 0.36	48.1 ± 16.0	39.9 ± 18.5	23/90 (25.6%)
	Full	4.24 ± 0.33	4.34 ± 0.31	4.60 ± 0.28	4.46 ± 0.33	90.5 ± 8.9	81.5 ± 12.1	37/90 (41.1%)

Note. A1, without the VAD constraint; A2, without functional-priority and age-friendly constraints; A3, without cross-stream attention; A4, without the multi-source constraint graph; A5, late-concatenation baseline. Subjective and alignment values are reported as mean ± SD across 30 held-out conditions after averaging the three seed-level outputs within each condition. First-pass constraint compliance is reported descriptively across the 90 generated outputs per configuration and task. Inferential full-ablation comparisons are reported in Supplementary Table S8.

Random-seed robustness was evaluated for the full model across six independent training seeds using the same 30 held-out conditions per task. Seed settings and complete results are reported in Supplementary Tables S3 and S9, respectively, and representative coordination outputs are provided in Supplementary Figure S4. Across the six seeds, subjective ratings, alignment measures, and first-pass compliance rates showed relatively limited variation across the tested initializations.

4.4. Boundary Cases, Safety-Relevant Failure Analysis, and Rollback

Boundary-case analysis focused on T3 abnormal-alert samples after automatic rollback. Remaining issues mainly involved avatar-expression intensity, CTA hierarchy, and surrounding-information density. For the key CTA, 8/30 first-pass samples fell below the 44 px threshold. After automatic rollback, 3/30 remained below the threshold, and all were corrected after manual review and re-checking.

Figure 8 presents representative rollback cases from T1 and T3, illustrating how task sequencing, CTA salience, semantic consistency, and avatar behavior issues were identified and corrected. The quantitative results above describe the overall recovery pattern, while the figure provides examples of the corresponding failure and correction processes.

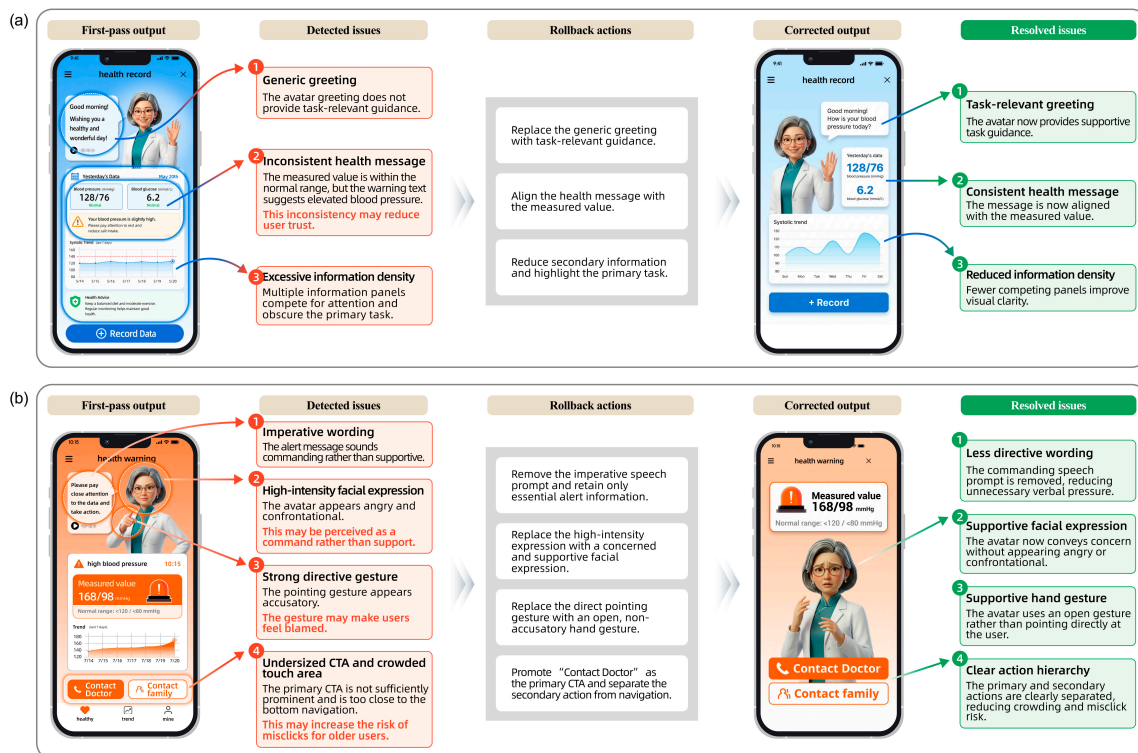


Figure 8. Representative rollback cases showing failure detection and correction in the generated avatar–UI schemes. **(a)** Recovery pattern represented by T1 output RB_T1_29, for which an information-density violation was resolved automatically in rollback round 2, with no manual constraint correction required. **(b)** Recovery pattern represented by T3 output RB_T3_13; the critical path was corrected automatically by round 3, whereas the CTA size remained below the 44-px threshold after automatic rollback and required manual correction before final re-checking. Semantic and behavioral issues shown in the callouts were identified during manual review rather than by the numerical rule checker. Both records passed final constraint validation.

Measurement-warning semantic mismatch and medication sequence disorder were classified as high-severity failure modes. Semantic and behavior-related risks were only partially captured by the numerical rule checker and therefore required additional human review. Table 13 summarizes the failure categories, their potential consequences, detectability, current mitigation, and residual risk.

The analysis was conducted as a prototype-stage hazard screening. Severity was defined according to the potential consequence of each failure, detectability according to whether it could be identified by the current rule checker, and residual risk according to the remaining concern after mitigation and review. High, Medium, and Low categories were used for the qualitative assessment.

Computational efficiency was also evaluated using the 90 rollback-analysis samples, with 30 samples per task and model-loading time excluded. Mean first-pass runtime was 6.505 ± 0.726 s, including condition encoding, graph construction, GAT and dual-stream planning, UI rendering, InstantID identity conditioning, SDXL diffusion rendering, and automatic constraint checking. Automatic rollback added an average of 0.532 ± 0.724 s, resulting in a mean total automatic runtime of 7.037 ± 0.978 s. Manual correction and human review were not included in the runtime estimates because their duration depended on operator involvement. Hardware, software, rendering, and optimization settings are provided in Supplementary Table S3, and the runtime breakdown is summarized in Table 14.

Table 13. Prototype-level failure and hazard screening.

Failure Mode	Potential Consequence	Severity	Detectability	Current Mitigation	Residual Risk
Measurement-warning semantic mismatch	Incorrect interpretation of an abnormal result or warning	High	Low	Task-conditioned planning; manual review and correction	Medium
Medication sequence disorder	Incorrect medication action or uncertainty about timing or dose	High	Medium	Operation-sequence edges; key-path checking; manual review and correction	Medium
Expression-content mismatch	Anxiety, confusion, or reduced trust	Medium	Medium	VAD-conditioned avatar planning; cross-stream coordination; manual review	Medium
Excessive urgency	Unnecessary stress or alarm fatigue	Medium	Low	Task-specific VAD conditioning; manual review and correction	Medium
Weak hierarchy	Delayed CTA discovery or missed next step	Medium	Medium	Functional priorities and priority edges; contrast, target-size, and key-path checks; rollback	Low

Table 14. Runtime of the complete prototype-generation pipeline.

Runtime Component	Mean (s)	SD (s)
Condition encoding	0.018	0.004
Graph construction	0.063	0.011
GAT + dual-stream planning	0.523	0.094
UI rendering	0.607	0.128
InstantID identity conditioning	1.43	0.267
SDXL diffusion rendering	3.805	0.636
Constraint checking/automatic validation	0.058	0.014
Total first-pass runtime	6.505	0.726
Rollback overhead	0.532	0.724
Total post-rollback automatic runtime	7.037	0.978

5. Discussion

5.1. Main Findings and Interpretation

This study examined whether affective goals and task engineering requirements could be jointly represented at the planning layer for avatar–UI prototype generation in older-adult chronic-disease management. Across the three scenarios, the framework produced task-sensitive avatar behavior and UI-priority changes while maintaining a shared structured representation of task conditions. Ablation results showed differentiated effects of affective constraints, task-engineering constraints, and coordination mechanisms on their corresponding evaluation dimensions. These findings are consistent with the intended roles of the model components but constitute within-framework evidence rather than proof of general superiority over other generation approaches.

The separate user evaluation provided complementary evidence on task performance, but the operational non-equivalence of T1–T3 means that timing, mistouch, CTA-selection, and workload differences should be interpreted as task-specific rather than framework-level efficiency effects. The external comparison is likewise capability-specific because the methods differ in conditioning and output scope. Rollback and boundary-case analyses further showed that first-pass violations and semantically unsafe outputs may remain

beyond numerical rule checking. Overall, the framework is best interpreted as a structured co-planning and prototype-screening approach that still requires human review and task-specific validation.

5.2. Methodological Limitations

Several limitations affect interpretation of the user-study results. The needs-modeling ($n = 50$) and additional user-evaluation ($n = 30$) cohorts were modest, non-probability samples. The needs-modeling cohort included only older adults with hypertension or diabetes and required prior smartphone experience, limiting generalizability to other chronic conditions and lower-digital-literacy populations. Visual and fine-motor difficulties were self-reported rather than clinically assessed. In addition, T1–T3 were operationally non-equivalent in actions, scrolling, decision structure, and CTA visibility; despite the randomized task order, residual learning, fatigue, and carry-over effects cannot be excluded. The study also covered only three predefined scenarios; broader cross-task, cross-language, and cross-cultural generalizability remains to be established.

The computational comparisons also have limited scope. External comparisons used only 10 first-pass outputs per method and involved heterogeneous baselines with different inputs, outputs, and applicable metrics; therefore, they should not be interpreted as a unified ranking. Ablation analysis covered five predefined removals with 30 held-out conditions per task and three training seeds, while robustness analysis used six seeds; these settings do not capture the full variability of training and generation. Subjective output quality was assessed by three HCI evaluators and therefore remains dependent on the pre-defined rubric and expert judgment despite masking and inter-rater reliability assessment.

Finally, the accessibility and safety findings should be interpreted at the prototype level and do not establish device-level accessibility compliance or clinical safety validation. Pixel-based thresholds were defined on the 1080×1920 reference canvas and require device-level revalidation before physical deployment. Rendering outcomes may also vary with backend models and versions. The reported boundary cases are illustrative rather than frequency estimates, and the observed first-pass violations show that constraint checking and rollback remain necessary parts of the workflow.

5.3. Safety and Applicability Boundaries

The framework should be interpreted as an early-stage prototype-generation and candidate-screening tool rather than a clinically validated or autonomous health-management system. Its rule checker addresses prototype-level properties such as font size, contrast, information density, target size, and interaction-path length, but does not replace clinical review, medication-safety assessment, semantic validation, or professional judgment. Downstream use will therefore require domain-expert review, explicit human oversight, deployment-specific safety evaluation [40], and longitudinal assessment of trust, alert fatigue [41], repeated-use behavior, and adaptation in home settings.

6. Conclusions

This study developed a dual-constrained co-planning framework for structured avatar–UI prototype generation under affective, functional, and age-friendly requirements. The framework integrates VAD-based affective targets, functional-module priorities, age-friendly engineering bounds, a multi-source constraint graph, and dual-stream planning within a shared task representation. Across the three evaluated health-management scenarios, the framework produced task-sensitive avatar behaviors and UI layouts and supported post-generation constraint checking and rollback.

External comparisons, ablation analyses, random-seed tests, and user evaluations provide preliminary prototype-level evidence for the roles of affective constraints, task representation, age-friendly bounds, and cross-stream coordination. These findings remain limited to the evaluated prototype conditions and do not establish clinical or deployment readiness. Future work should extend validation to larger and more diverse populations, additional tasks and devices, dynamic and multi-turn interaction, clinical review, device-specific accessibility testing, and longitudinal home-use evaluation.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/app16199695/s1>. The supporting information includes Supplementary Note S1: Prompts and input configurations for the T1 external comparison; Supplementary Note S2: Evaluator protocol, masking, rating aggregation, and inter-rater reliability; Figure S1: Exemplary virtual avatar expression patterns under different VAD conditions; Figure S2: Data-construction and evaluation protocol; Figure S3: Representative first-pass outputs from eight external baseline methods under the T1 common-reference condition; Figure S4: Illustrative avatar–UI coordination outputs for the Full model and coordination ablations under the same T3 abnormal-alert input and training seed; Questionnaire S1: User needs assessment; Questionnaire S2: Generated output evaluation; Questionnaire S3: Prototype test using NASA-TLX; Supplementary Data S1: De-identified participant-level needs-modeling data; Supplementary Data S2: Participant-level data from the additional user evaluation; Supplementary Data S3: Training-data manifest; Supplementary Data S4: Generated-output-level evaluation data; Supplementary Data S5: Constraint-checking and rollback logs; Table S1: Characteristics of the two participant cohorts; Table S2: Derivation of functional-module weights; Table S3: Training and rendering implementation settings; Table S4: Controlled construction of the nine variants; Table S5: Pairwise comparisons for the additional user evaluation; Table S6: First-pass hard-constraint checks in the T1 external comparison; Table S7: Output-level anchor counts used as denominators for gaze and gesture alignment; Table S8: Mechanism-aligned Full-model contrasts for the ablation analysis; Table S9: Random-seed robustness analysis across T1, T2, and T3; Table S10: Implementation provenance and inference settings for the external baselines.

Author Contributions: Conceptualization, X.Z.; methodology, X.Z. and Y.Z.; validation, X.Z.; investigation, H.Z.; data curation, H.Z.; visualization, H.Z.; writing—original draft preparation, H.Z.; writing—review and editing, X.Z. and Y.Z.; supervision, Y.Z.; project administration, Y.Z.; funding acquisition, X.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Youth Project of Humanities and Social Sciences Research of the Ministry of Education of China, grant number 24YJC760170; the 2026 Hebei University of Technology Undergraduate Education and Teaching Reform Research and Practice Project, grant number 202604058; and the 2027 Hebei University of Technology Postgraduate Education and Teaching Reform Research Incubation Project, grant number Yjg2027018.

Institutional Review Board Statement: The study was reviewed and approved by the Research Ethics and Technology Security Committee of the School of Architecture and Art Design, Hebei University of Technology (approval date 20 August 2025; approval number N/A, as no number was assigned by the Committee).

Informed Consent Statement: Written informed consent was obtained from all participants involved in the study.

Data Availability Statement: De-identified participant-level data supporting the needs-modeling and user-evaluation analyses, the training-data manifest, generated-output-level evaluation records, and constraint-checking and rollback logs are provided as Supplementary Data S1–S5. The accompanying data dictionary documents variable definitions, units, valid ranges, scoring rules, analysis units, and aggregation procedures. Direct personal identifiers are not included in the released files. Implementation settings and random seeds required to reproduce the reported computational experiments are provided in the Supplementary Information.

Acknowledgments: We thank all older adult participants for their time and valuable feedback. During the preparation of this manuscript, the authors used ChatGPT (OpenAI, GPT-5.4) for grammar checking, identifying typographical and language errors, and language refinement in selected parts of the manuscript. A representative prompt was: “Please check the following text for grammar, spelling, and wording without changing the scientific meaning or numerical values.” The authors reviewed and edited all AI-assisted output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Krutter, S.; Schuessler, N.; Kutschar, P.; Šabić, E.; Dellinger, J.; Klausner, T.; Nestler, N.; Beasley, M.; Henderson, B.; Pitzer, S.; et al. Piloting of the virtual telecare technology ‘Addison Care’ to promote self-management in persons with chronic diseases in a community setting: Protocol for a mixed-methods user experience, user engagement and usability pilot study. *BMJ Open* **2022**, *12*, e062159. [\[CrossRef\]](#)
2. Foukarakis, M.; Karuzaki, E.; Adami, I.; Ntoa, S.; Partarakis, N.; Zabulis, X.; Stephanidis, C. Quality assessment of virtual human assistants for elder users. *Electronics* **2022**, *11*, 3069. [\[CrossRef\]](#)
3. Zhou, C.; Guo, J.; Zhang, Y.; Huang, T.; Kaner, J. A home health management app designed in the post-epidemic era using empirical evidence based on the demand collection of elderly users. *Humanit. Soc. Sci. Commun.* **2024**, *11*, 619. [\[CrossRef\]](#)
4. Bin Sawad, A.; Narayan, B.; Alnefaie, A.; Maqbool, A.; Mckie, I.; Smith, J.; Yuksel, B.; Puthal, D.; Prasad, M.; Kocaballi, A.B. A systematic review on healthcare artificial intelligent conversational agents for chronic conditions. *Sensors* **2022**, *22*, 2625. [\[CrossRef\]](#)
5. Kokorelias, K.M.; Grigorovich, A.; Harris, M.T.; Rehman, U.; Ritchie, L.; Levy, A.M.; Denecke, K.; McMurray, J. Longitudinal coadaptation of older adults with wearables and voice-activated virtual assistants: Scoping review. *J. Med. Internet Res.* **2024**, *26*, e57258. [\[CrossRef\]](#)
6. Troncone, A.; Saturno, R.; Buonanno, M.; Pugliese, L.; Cordasco, G.; Vogel, C.; Esposito, A. Advanced assistive technologies for elderly people: A psychological perspective on older users’ needs and preferences (Part B). *Acta Polytech. Hung.* **2021**, *18*, 29–44. [\[CrossRef\]](#)
7. Peerbolte, T.F.; van Diggelen, R.J.; van den Haak, P.; Geurts, K.; Evers, L.J.; Bloem, B.R.; de Vries, N.M.; van den Berg, S.W. Conversational agents supporting self-management in people with a chronic disease: Systematic review. *J. Med. Internet Res.* **2025**, *27*, e72309. [\[CrossRef\]](#)
8. Loveys, K.; Sebaratnam, G.; Sagar, M.; Broadbent, E. The effect of design features on relationship quality with embodied conversational agents: A systematic review. *Int. J. Soc. Robot.* **2020**, *12*, 1293–1312. [\[CrossRef\]](#)
9. Kwon, Y.H.; Yoon, J.H.; Park, M.G. Text2Avatar: Articulated 3D avatar creation with text instructions. *IEEE Trans. Multimed.* **2025**, *27*, 3797–3806. [\[CrossRef\]](#)
10. Hong, F.; Zhang, M.; Pan, L.; Cai, Z.; Yang, L.; Liu, Z. AvatarCLIP: Zero-shot text-driven generation and animation of 3D avatars. *ACM Trans. Graph.* **2022**, *41*, 161. [\[CrossRef\]](#)
11. Lamberti, F.; Manuri, F.; Paravati, G.; Piumatti, G.; Sanna, A. Using semantics to automatically generate speech interfaces for wearable virtual and augmented reality applications. *IEEE Trans. Hum. Mach. Syst.* **2017**, *47*, 152–164. [\[CrossRef\]](#)
12. Gambino, O.; Rundo, L.; Cannella, V.; Vitabile, S.; Pirrone, R. A framework for data-driven adaptive GUI generation based on DICOM. *J. Biomed. Inform.* **2018**, *88*, 37–52. [\[CrossRef\]](#)
13. Wang, Z.; Li, J.S.; Pan, H.R.; Wu, J.Y.; Yan, W.A. Research on multimodal generative design of product appearance based on emotional and functional constraints. *Adv. Eng. Inform.* **2025**, *65*, 103106. [\[CrossRef\]](#)
14. Yoo, S.; Lee, S.; Kim, S.; Hwang, K.H.; Park, J.H.; Kang, N. Integrating deep learning into CAD/CAE system: Generative design and evaluation of 3D conceptual wheel. *Struct. Multidiscip. Optim.* **2021**, *64*, 2725–2747. [\[CrossRef\]](#)
15. Chen, J.; Tao, H.; Wu, J.; Yuan, Q.; Ma, L.; Chen, D.; Yu, M. Affective conveyance assessment of AI-generative static visual user interfaces based on valence-arousal emotion model. *Displays* **2026**, *91*, 103261. [\[CrossRef\]](#)
16. Yang, J.; Feng, J.; Huang, H. EmoGen: Emotional image content generation with text-to-image diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 17–21 June 2024; pp. 6358–6368. [\[CrossRef\]](#)
17. Dang, S.; He, Y.; Ling, L.; Qian, Z.; Zhao, N.; Cao, N. EmotiCrafter: Text-to-Emotional-Image Generation Based on Valence-Arousal Model. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Honolulu, HI, USA, 19–25 October 2025; pp. 15218–15228. [\[CrossRef\]](#)

18. Huo, Y.; Ge, Y. Multi-task emotion recognition based on dimensional model and category label. *IEEE Access* **2024**, *12*, 75169–75179. [CrossRef]
19. Moeller, S.K.; Lee, E.A.E.; Robinson, M.D. You never think about my feelings: Interpersonal dominance as a predictor of emotion decoding accuracy. *Emotion* **2011**, *11*, 816–824. [CrossRef]
20. Chen, J.; Chen, B.; Lei, J.; He, X.; Chen, L.; Kim, W.S. A behavior-driven adaptive user interface generation framework with iterative preference modeling and prompt fusion. *Signal Image Video Process.* **2026**, *20*, 128. [CrossRef]
21. Sobolevsky, A.; Bilodeau, G.-A.; Cheng, J.; Guo, J.L.C. GUILGET: GUI Layout Generation with Transformer. In Proceedings of the 36th Canadian Conference on Artificial Intelligence (Canadian AI 2023), Montréal, QC, Canada, 5–9 June 2023. [CrossRef]
22. Jiang, Y.; Zhou, C.; Garg, V.; Oulasvirta, A. Graph4GUI: Graph Neural Networks for Representing Graphical User Interfaces. In Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24), Honolulu, HI, USA, 11–16 May 2024; p. 988. [CrossRef]
23. Lin, K.Q.; Li, L.; Gao, D.; Yang, Z.; Wu, S.; Bai, Z.; Lei, S.W.; Wang, L.; Shou, M.Z. ShowUI: One vision-language-action model for GUI visual agent. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 11–15 June 2025; pp. 19498–19508. [CrossRef]
24. Wickramathilaka, S.; Grundy, J.; Madampe, K.; Haggag, O. Adaptive and accessible user interfaces for seniors through model-driven engineering. *Autom. Softw. Eng.* **2025**, *32*, 74. [CrossRef]
25. Tang, C.; Wang, L.; Pan, D.; Zhang, H.; Fang, Y. Matching embodied conversational agent appearance to message emotion enhances persuasion in eHealth. *Sci. Rep.* **2025**, *15*, 42723. [CrossRef]
26. Tamantini, C.; Umbrico, A.; Fracasso, F.; Beraldo, G.; Cortellessa, G.; Orlandini, A. Adapting virtual agent interaction style with reinforcement learning to enhance affective engagement. *Front. Digit. Health* **2026**, *7*, 1680605. [CrossRef]
27. Wang, Q.; Bai, X.; Wang, H.; Qin, Z.; Chen, A.; Li, H.; Tang, X.; Hu, Y. Instant ID: Zero-shot identity-preserving generation in seconds. *arXiv* **2024**, arXiv:2401.07519. [CrossRef]
28. Wang, Z.; Shi, Y.; Wang, Y.; Yao, Y.; Yan, K.; Wang, Y.; Ji, L.; Xu, X.; Yu, C. G-VOILA: Gaze-facilitated information querying in daily scenarios. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2024**, *8*, 1–33. [CrossRef]
29. World Wide Web Consortium (W3C). Web Content Accessibility Guidelines (WCAG) 2.2. W3C Recommendation. 12 December 2024. Available online: <https://www.w3.org/TR/WCAG22/> (accessed on 15 September 2026).
30. Hou, G.; Anicetus, U.; He, J. How to design font size for older adults: A systematic literature review with a mobile device. *Front. Psychol.* **2022**, *13*, 931646. [CrossRef]
31. Bradley, M.M.; Lang, P.J. Measuring emotion: The Self-Assessment Manikin and the Semantic Differential. *J. Behav. Ther. Exp. Psychiatry* **1994**, *25*, 49–59. [CrossRef]
32. Koo, T.K.; Li, M.Y. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *J. Chiropr. Med.* **2016**, *15*, 155–163. [CrossRef]
33. Friedman, M. The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance. *J. Am. Stat. Assoc.* **1937**, *32*, 675–701. [CrossRef]
34. Holm, S. A Simple Sequentially Rejective Multiple Test Procedure. *Scand. J. Stat.* **1979**, *6*, 65–70.
35. Hart, S.G.; Staveland, L.E. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Human Mental Workload*; Hancock, P.A., Meshkati, N., Eds.; North-Holland: Amsterdam, The Netherlands, 1988; Volume 52, pp. 139–183. [CrossRef]
36. Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; Rombach, R. SDXL: Improving latent diffusion models for high-resolution image synthesis. In Proceedings of the International Conference on Learning Representations (ICLR), Vienna, Austria, 7–11 May 2024.
37. Li, Z.; Cao, M.; Wang, X.; Qi, Z.; Cheng, M.-M.; Shan, Y. PhotoMaker: Customizing realistic human photos via stacked ID embedding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 17–21 June 2024; pp. 8640–8650. [CrossRef]
38. Huang, J.; Dong, X.; Song, W.; Chong, Z.; Tang, Z.; Zhou, J.; Cheng, Y.; Chen, L.; Li, H.; Yan, Y.; et al. ConsistentID: Portrait Generation with Multimodal Fine-Grained Identity Preserving. *IEEE Trans. Pattern Anal. Mach. Intell.* **2026**, *48*, 5639–5654. [CrossRef]
39. Chen, J.; Zhang, R.; Zhou, Y.; Jain, R.; Xu, Z.; Rossi, R.; Chen, C. Towards aligned layout generation via diffusion model with aesthetic constraints. In Proceedings of the International Conference on Learning Representations (ICLR), Vienna, Austria, 7–11 May 2024.

40. World Health Organization. *Ethics and Governance of Artificial Intelligence for Health: WHO Guidance*; World Health Organization: Geneva, Switzerland, 2021.
41. Hussain, M.I.; Reynolds, T.L.; Zheng, K. Medication safety alert fatigue may be reduced via interaction design and clinical role tailoring: A systematic review. *J. Am. Med. Inform. Assoc.* **2019**, *26*, 1141–1149. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.