

# Training Under Pressure: Designing and Evaluating a Multi-Stressor VR Environment for Public Speaking Training

Zhiyuan (Terry) Wei

Department of Informatics  
New Jersey Institute of Technology  
Newark, New Jersey, USA  
zw49@njit.edu

Lei Zhang

Department of Informatics  
New Jersey Institute of Technology  
Newark, New Jersey, USA  
lei.zhang@njit.edu



**Figure 1:** We present a multi-stressor VR system for public speaking training. With experiment conditions, users will confront a) Positive audience who show agreement animations (nodding, smiling) and no interruptions; b) Neutral audience who listen (with idle animations) and ask clarification questions; c) Negative audience who strongly disagree with dismissive animations (shaking heads, rejecting) and interrupt users with adversarial questions.

## Abstract

Effective public speaking requires performing under stress, yet prior VR training systems have primarily focused on reducing anxiety rather than building the resilience to withstand it. In this paper, we introduce a multi-stressor VR public speaking training system that combines audience non-verbal behavior, crowd size, and AI-generated contextual Q&A, each designed across manipulable levels of stress. In an exploratory within-subjects study ( $N = 12$ ), we characterize how combined stressor intensity relates to speaking performance and the contribution of individual stressors. Results show that our parameterized stressor design successfully induced distinct anxiety levels, with negative non-verbal animations serving as a stronger source of stress than either crowd size or AI-generated adversarial interruptions. In addition, AI interruptions forced a cognitive context switch during speech, prompting speakers to adopt coping strategies such as adjusting eye gaze and pacing to regain composure. Finally, participants identified the high-stress negative audience condition as the most valuable for training. Together, these findings suggest that rather than minimizing stress, VR systems for public speaking should calibrate challenges to a productive zone and incorporate cognitive scaffolds that help users sustain performance under high-pressure interruptions.

## CCS Concepts

• **Human-centered computing** → **Interactive systems and tools**; **Virtual reality**; **Empirical studies in HCI**.



This work is licensed under a Creative Commons Attribution 4.0 International License. *UIST '26, Detroit, MI, USA*

© 2026 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2856-3/2026/11

<https://doi.org/10.1145/3830398.3830711>

## Keywords

Virtual Reality, Public Speaking, Stress, Mixed Methods, Empirical study

## ACM Reference Format:

Zhiyuan (Terry) Wei and Lei Zhang. 2026. Training Under Pressure: Designing and Evaluating a Multi-Stressor VR Environment for Public Speaking Training. In *The 39th Annual ACM Symposium on User Interface Software and Technology (UIST '26)*, November 02–05, 2026, Detroit, MI, USA. ACM, New York, NY, USA, 15 pages. <https://doi.org/10.1145/3830398.3830711>

## 1 Introduction

Effective public speaking is a cognitively demanding task that requires sustained performance under stress [6, 13]. Public Speaking Anxiety (PSA), also known as Glossophobia [46], manifests in both physiological and psychological disruptions to speech delivery [41]. Virtual Reality (VR) has therefore emerged as a widely adopted platform for public speaking training by offering a psychologically safe environment for practicing in front of virtual audiences [3]. However, prior VR training systems have primarily focused on reducing PSA rooted in exposure therapy principles [36]. While these systems were successful in habituating users to the presence of simulated audiences over repeated sessions [54, 59], they often fail to build the resilience required to sustain delivery quality when faced with adversarial reactions, unexpected interruptions, or other stressful disruptions that are routine in effective speaking [62]. Simulating stress in the immersive virtual environment is important because VR practice only transfers to real-world performance when the virtual setting produces psychological responses comparable to those of a live audience [38, 41, 53, 62]. Without this pressure, speakers are essentially rehearsing alone [3] and cannot improve.

To address this gap, we introduce a multi-stressor VR public speaking training system designed to train users under pressure.

Unlike prior systems that rely on static [53] or exclusively positive [37] virtual audiences, our system integrates and examines three core stressors: audience non-verbal behavior [17, 41, 43], crowd size [42, 53], and AI-generated contextual Q&A [24]. We parameterize each stressor across discrete intensity levels. Specifically, audience non-verbal behavior spans three levels, ranging from supportive nodding through neutral stillness to dismissive rejecting. Crowd size spans four levels in a meeting room: an empty room, a small panel of three, a medium panel of eight, and a packed room of seventeen. To induce cognitive interruptions, we utilize real-time, AI-generated questions grounded in the speaker’s live content and delivered mid-speech, moving beyond static, scripted prompts used in prior systems [24]. These interruptive questions are designed across three stress levels: no questions, clarification questions, and adversarial questions. This approach more faithfully approximates the cognitive demands of unscripted live Q&A [2, 30]. Together, these parameterizations enable a systematic investigation of each stressor’s individual and combined effects on public speaking anxiety and performance in VR.

We evaluated the system via a mixed-methods, within-subjects study ( $N=12$ ) to characterize how these individual and combined stressor dimensions affect public speaking anxiety and performance. First, our results show that self-rated anxiety increased consistently with combined stressor intensity, indicating that our parameterized design can induce graded stress levels. Among individual stressors, negative non-verbal behavior of the audience (e.g., frowning, head-shaking) acted as a stronger source of perceived anxiety and cognitive workload than either increased crowd size or adversarial AI-generated questions, partly because participants tended to internalize these negative non-verbal cues as social-evaluative threats. Second, we found that AI-generated verbal questions increased cognitive load due to context switching during speech. Speakers thus adopted various strategies such as selectively redirecting their gaze toward non-hostile avatars and deliberately slowing their speech pacing (e.g., taking more pauses) to regain composure. Finally, despite experiencing the highest levels of stress, participants overwhelmingly rated the negative audience condition as the most valuable for resilience training.

Taken together, these findings highlight the potential of a dynamic VR speech training system: rather than minimizing stress, immersive environments can calibrate challenges to a productive difficulty zone and integrate cognitive scaffolds that support performance recovery under pressure. This paper makes the following contributions:

- The design and implementation of a parametrically controllable, multi-stressor VR public speaking training system;
- An exploratory, empirical characterization of the comparative impact of social, spatial, and cognitive stressors on speech anxiety and recovery;
- Design implications for resilience-oriented and challenge-calibrated VR training systems.

## 2 Related Work

Our work builds on prior work in VR public speaking training systems, virtual audience stressor design, and the interplay between anxiety and public speaking performance.

### 2.1 VR Systems for Public Speaking Training

VR has become a standard platform for public speaking training because it can place speakers in front of controllable virtual audiences that elicit genuine anxiety responses comparable to live settings [41, 53]. Prior work has shown that repeated practice in such environments can reduce self-reported anxiety and improve performance [3, 27, 37, 38, 54], and the overall experience depends heavily on task difficulty, social presence, and baseline fear [42].

Beyond basic audience exposure, recent systems have explored interactive public speaking. For example, Wang et al. demonstrated that conversational agents could deliver cognitive restructuring to lower pre-speech anxiety, while agent sociability could foster closeness, further reducing state anxiety [59]. Similarly, supportive virtual agents or co-presenters could reduce anxiety physiologically and subjectively [35]. Furthermore, private AR filters can let speakers modulate audience perception and lower the fear of judgment [29]. Other systems also explored audience that can deliver pre-scripted conversational interruptions [25], or utilized LLM-generated audiences with configurable personas (e.g., AudiLens [39]). While these systems relied on pre-scripted questions or generic personas, our system integrates specific positive, neutral, and negative personas with interruptive questions that are contextually grounded in the speaker’s live speech. In addition, despite these interactive mechanisms, prior public speaking training in VR was predominantly rooted in exposure therapy: the goal is to acclimate users to audience presence until anxiety habituates [28, 36, 44]. This is effective for desensitization, but does not explicitly train the ability to perform well under active pressure. Our system departs from this paradigm by treating stress as a training variable to be calibrated rather than minimized.

### 2.2 Audience Stressors

**2.2.1 Audience Nonverbal Behavior.** Nonverbal audience cues have been studied in causing stress. Pertaub et al. found that behavior such as frowning and looking away produces significantly higher anxiety than positive or neutral conditions [41]. Furthermore, Poeschl and Döering confirmed that this pressure is carried by micro-cues like whispering, crossed arms, or withdrawn gaze rather than by a single on/off audience-presence variable [43]. Parameterized audience models can systematically vary arousal, valence, mood, and personality to produce distinct, recognizable behavioral styles [8, 24]. Gaze dynamics add another layer. Wechsler et al. demonstrated that attentional focus manipulation in VR changes anxiety and gaze behavior for socially anxious users [61]. Similarly, Fehlmann et al. showed that VR gaze exposure reduces state anxiety across sessions [16], while Mello et al. found that socially anxious individuals exhibit distinctive gaze avoidance toward virtual agents [32]. These studies link social pressure to gaze reciprocity: the sense of watching and being watched.

**2.2.2 Crowd Size.** Audience scale is a common lever in VR speaking systems. The mere presence of a visible virtual audience triggers social-evaluative pressure [53], and platforms have varied group size from small panels to auditorium-scale crowds [37, 42]. The Trier Social Stress Test includes an evaluative audience panel as a core component, confirming the link between audience presence and cortisol-mediated arousal [6, 13]. Industry platforms push this

further with staircase sizes up to 10,000-seat stadiums. Yet few systems have isolated crowd size as an independently parameterized variable alongside other stressor types, making it difficult to tease apart its contribution from confounds like room geometry or audience density.

**2.2.3 Cognitive Interruptions and Q&A.** Mid-speech interruptions impose a qualitatively different burden than audience presence or behavior. Unexpected disruptions are known to increase stress and degrade performance in workplace settings [2, 30]. Kang et al. found that mid-speech events such as ringing phones or unexpected questions increase perceived anxiety and disrupt speech flow [25]. However, prior VR systems have relied on environmentally triggered interruptions rather than disruptions grounded in the speaker’s content [9]. Our system generates real-time adversarial questions derived from the speech, creating cognitive disruption closer to the unpredictability of real-world Q&A.

## 2.3 Measuring Performance Under Stress

Capturing how stressors affect speaking requires both subjective and objective measures. Self-report instruments such as the PRPSA and state anxiety scales are standard [45], but do not always reach significance under audience manipulations [17], motivating complementary behavioral measures. Automated assessment tools can approximate expert-rated performance from speakers’ multimodal behavioral indicators including eye contact, vocal variability, and pause patterns [5, 64]. Prior real-time coaching systems have often delivered corrective feedback targeting specific delivery dimensions such as speech rate and filler words alerts via smart glasses (e.g., Rhema [55]), posture and gestures through body sensing and wearable feedback [11, 50], voice modulation [60], and robotic or LLM-driven coaches for multimodal and gestural delivery [47, 58]. Visual analytics approaches, such as GestureLens [68], have also been developed to help coaches explore gesture usage in presentation videos, connecting spatial and temporal gesture distributions to speech content. While prior coaching or rehearsal systems focused on multimodal assessment of speakers’ performance, they did not induce stress during training, thus failing to capture speakers’ unique behavior or recovery patterns under stress. Our work contributes to this line of work by introducing parameterized multi-stressor control, and thus offering the categorization of speakers’ perceived performance, stress, and recovery strategies in VR.

## 2.4 From Anxiety Reduction Toward Resilience-Oriented Training

The exposure therapy paradigm underpinning most VR public speaking systems assumes that reducing anxiety is sufficient for good performance [36, 44]. Several lines of evidence complicate this picture. The Yerkes-Dodson law implies that moderate stress can enhance rather than impair performance [56, 66], and the desirable difficulty framework argues that conditions that feel harder during practice produce better long-term transfer [7]. Together, these suggest that purely comfort-oriented training may deprive learners of productive stress. Recent LLM-driven systems for conflict resolution and interview practice, such as Rehearsal [52] and Conversate [12], have shown that simulated high-pressure social

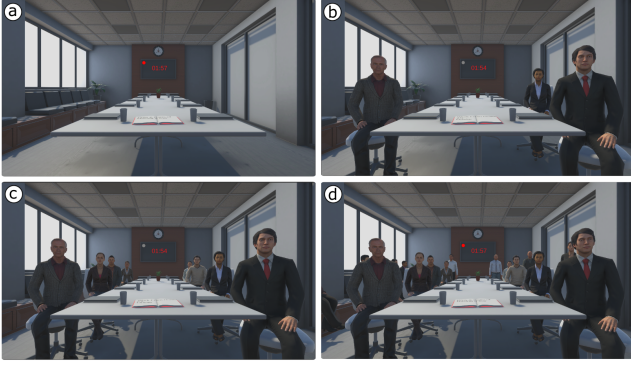
interactions, paired with theory-grounded or dialogic feedback, improve real-world skill transfer beyond what lecture-based training achieves. Prior work has shown that VR-based stress inoculation training using parameterized interruptions and time pressure can serve as viable training for high-stress professional contexts [62]. Fang et al. found that combining VR with generative AI for social stress scenarios has likewise shown promise for building coping strategies [15]. Furthermore, Zhou et al. established that VR reflection techniques that let presenters observe their own delivery from a third-person audience perspective produce more objective self-evaluation, particularly for speakers with low confidence [70], while Ram et al. showed that gesture rehearsal with targeted AI recommendations increases speakers’ gesture diversity compared to notes-only practice [47]. Our system builds on this trajectory by combining three parameterized stressor dimensions across discrete intensity levels, which enables systematic study of how different types and intensities of pressure affect perceived anxiety and performance, including recovery dynamics. Our findings thus aim to guide the calibration of challenges to a productive zone and to build towards resilience-oriented public speaking training in VR.

## 3 System Design

To support authentic public speaking training, we developed an end-to-end VR system that immerses users in simulated stressful presentation scenarios. Unlike prior VR training systems that rely on pre-scripted audience behaviors and fixed question banks, our system generates context-aware interruption questions grounded in the user’s live speech, delivered by a virtual audience capable of mimicking natural human behavior to simulate a realistic evaluative atmosphere. By replicating these subtle social cues, the system induces Social Evaluative Threat [13], making the presenter feel that their performance is being actively monitored and judged by the crowd. Our system also provides a backend Unity interface for practitioners and experimenters to manipulate stressors in real time. We detail the system design in the sections below.

### 3.1 System Architecture

Our system architecture aims to offer *closed-loop, real-time interactions*. High inference latency is a known bottleneck in LLM-driven XR applications [69]; to sustain a target rendering framerate of 72–90 Hz while processing multimodal LLM queries, we decouple the Unity rendering thread from the AI network pipeline. The system’s primary inputs are (1) continuous microphone audio sampled at 44.1 kHz, and (2) session metadata including the current presentation topic and the elapsed time, which inform the LLM about the speech’s progress. A local Voice Activity Detection (VAD) module performs on-device audio segmentation and buffering. Resulting audio segments, bundled with the accumulated transcript context, are dispatched asynchronously via `UnityWebRequest` to a cloud-hosted LLM, which returns both a transcription for future context and a candidate question to a question pool. A higher-level *Virtual Director* state machine monitors a global presentation timer and, at predefined checkpoints, selects the best question from the candidate pool, triggers the corresponding avatar’s raise-hand animation, and invokes a Text-to-Speech (TTS) service to play the question, completing a natural interruption Q&A loop.



**Figure 2: Four levels of crowd size: a) Empty–0 person, b) Basic–3 people, c) Some–8 people, d) Full–17 people.**

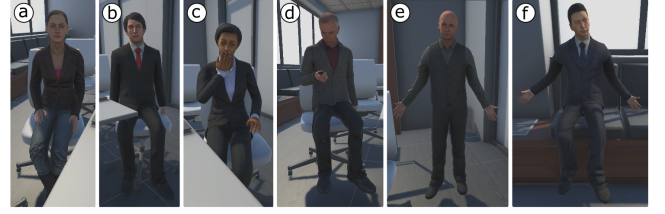
### 3.2 Dynamic Audience Simulation

**3.2.1 Size and layout scalability.** Prior work has demonstrated that the number of evaluators is a direct trigger of social anxiety in public speaking contexts, with larger virtual audiences eliciting greater physiological arousal [34]. Our system allows the audience size to transition across four levels (as seen in Fig. 2) at runtime. By dynamically culling and activating avatar instances, the system modulates the size stressor without requiring scene reloads.

**3.2.2 Animation and behavioral profiles.** Extensive research demonstrates that a virtual audience’s non-verbal feedback, regardless of type, directly impacts a speaker’s state anxiety [17, 41]. To replicate this spectrum of non-verbal feedback, we assign each avatar one of four behavioral profiles combining different types of animations (as seen in Fig. 3) to restrict its available animation pool: *Positive* (nodding, smiling), *Negative* (head-shaking, rejecting), *Distracted* (phone-checking, gaze aversion), and *Neutral* (*idle breathing*). To maintain consistent personas throughout the session, non-verbal and verbal behaviors are tightly coupled. Visually, each avatar is restricted to its assigned animation pool (e.g., a “negative” avatar will only trigger dismissive animations). Conversationally, the LLM is fed the accumulated speech history alongside a designated role prompt (e.g., instructing it to act as a “strict reviewer”). This ensures that the generated questions align with the avatar’s visual (non-)hostility and reasonably build upon previous interaction contexts. In addition, prior work has shown that perfectly synchronized crowd movements could degrade realistic nonverbal behavior and break user immersion [24, 43]. To prevent this, agents trigger new clips from their assigned animation pool on a randomized timeout (8–15 seconds). This asynchronous triggering avoids robotic synchronization while maintaining the condition’s overall sentiment.

### 3.3 Interactive User Flow

The user begins their 120-second speech while the virtual audience exhibits condition-specific non-verbal animations to simulate a realistic, dynamic crowd. At predefined checkpoints during the presentation, the system’s *Virtual Director* intervenes to initiate a Q&A interaction. A selected avatar stands and raises its hand to signal a question. The user acknowledges the interruption by aiming a controller raycast at the standing avatar and pressing the trigger button to grant speaking permission — a design choice



**Figure 3: Selective animations showing: a) Neutral listening; b) Positive nodding; c) Distracted yawning; d) Distracted phone-checking; e) Negative rejecting; f) Negative head-shaking with dismissive hand movement.**

intended to offer the presenter agency over the exact timing of the interruption, and to prevent the system from talking over the presenter. The avatar then asks a contextually relevant question (either clarification or adversarial, based on the assigned condition) aloud. The user verbally answers the question and finishes the answer via a button press on the VR controller. Finally, the avatar plays an animation clip of acknowledgment, returns to a seated posture, and the user resumes the remainder of their presentation.

### 3.4 Backend Pipeline and Experimenter Control

To support a seamless user flow, we designed a robust backend pipeline that handles speech segmentation, question generation, and experimenter overrides.

**3.4.1 Local audio segmentation.** We implement a local, acoustic-driven slicing mechanism that preserves semantic integrity while minimizing perceived latency. The system captures microphone input into a circular buffer and compares Root Mean Square (RMS) energy across a 128-sample sliding window against a voice activity threshold to detect speech. It segments audio at fixed 10 s intervals (`autoUploadInterval`) to bound the payload size and reduce API timeout risk. The pipeline additionally monitors silence duration via a configurable pause threshold (`silenceTimeout = 2 s`), establishing an extensive hook for future work to trigger audience questions at natural phrase boundaries.

**3.4.2 Question generation using LLM.** Each API call bundles the presentation topic, the accumulated transcript, and the current audio slice into a single JSON. The LLM is prompted with a designated persona instruction (e.g., “strict reviewer” for hostile conditions) alongside the accumulated transcript to produce both a verbatim transcription of the newly uploaded audio slice and a candidate question grounded in the ongoing speech (see Appendix C, Prompt C). The system then aggregates the candidate questions into a candidate question pool.

**3.4.3 Virtual Director.** The Virtual Director issues a secondary LLM prompt (see Appendix C, Prompt C) to select the best question from the pool at exactly 70 s and 40 s remaining (in order to account for the generation latency). This 10-second offset ensures that the avatar interrupts the user around the intended checkpoints of 60 s and 30 s remaining. These two checkpoints are designed to ensure sufficient speech content has accumulated for contextually grounded question generation.

**3.4.4 Experimenter control interface.** The Unity ControlPanel offers runtime control of major parameters: audience size, behavioral profiles, in-scene topic, director and timer settings. Experimenters can also trigger questions via hotkeys for non-disruptive ad-hoc interventions (see Fig. 6 in Appendix A).

## 3.5 Implementation

The system was implemented in Unity (v2022.3.61f1) using the Meta XR Core Utilities package. We used the Microsoft Rocketbox Avatar library [18] for gender-diverse humanoid avatars. We used Gemini 2.5 Pro [19] for multimodal speech understanding and question generation, and ElevenLabs [14] for text-to-speech synthesis.

## 4 Evaluation

We designed a mixed-method, within-subjects study to evaluate how combined and individual audience stressors in VR influence public speaking anxiety, performance, and recovery. Specifically, our study aims to answer the following research questions:

- **RQ1:** To what extent does our pipeline reliably generate contextually grounded, real-time Q&A interruptions during VR public speaking training?
- **RQ2:** How do combined multi-stressor conditions of varying intensity (supportive, neutral, negative) modulate perceived anxiety and overall speaking performance?
- **RQ3:** Compared to a neutral baseline, what is the impact of individual stressors (audience size, negative non-verbal behavior, hostile questions) on speaker anxiety?
- **RQ4:** How do users perceive and cope with the cognitive disruption introduced by the audience stressors?

### 4.1 Participants

We recruited 12 participants (8 male, 3 female, 1 non-binary; aged 18–29,  $M = 23.5$ ,  $SD = 3.87$ ). Prerequisites included English proficiency, normal or corrected-to-normal vision, no history of severe motion sickness, and prior VR experience (having used VR at least once). Participants primarily had prior public speaking experience in academic or classroom settings. To capture a representative range of public speaking dispositions, we assessed trait-level speaking anxiety using the Personal Report of Public Speaking Anxiety (PRPSA) [31], which yielded scores ranging from 75 to 127 ( $M = 104.25$ ,  $SD = 15.62$ ), spanning low to medium-high trait anxiety. Participants also reported their baseline state anxiety at the start of the session. All participants provided informed consent and were compensated \$20 Amazon gift card for a 75-minute study. The study protocol was approved by our institution’s IRB.

### 4.2 Study Design

We employed a within-subject design with four experimental conditions, each representing a distinct configuration of virtual audience behavior. Three core conditions (A, B, C) formed an ordinal stress gradient, designed to simulate overall stress levels by bundling multiple stressor factors (audience size, non-verbal behaviors, and question types) together. Meanwhile, a fourth condition (D) served as a between-subjects manipulation to isolate the effect of individual stressor components. All conditions use the same environment:

a virtual meeting room with a meeting table, with a visible countdown timer displayed throughout the speech. We included time pressure as a baseline environmental factor rather than an independent variable. The conditions differed as follows:

- **Condition A (Supportive).** The virtual audience exhibited positive nonverbal animations (e.g., nodding, smiling). No verbal interruptions or questions were posed.
- **Condition B (Neutral).** The virtual audience displayed neutral animations (e.g., idle breathing, listening). Two clarification questions were asked at predefined intervals, each followed by the prompt “*Thank you, please continue.*”
- **Condition C (Negative).** Virtual audiences performed negative nonverbal behaviors (e.g., head shaking, dismissive gestures). Two adversarial or confrontational questions were interjected during the speech, each followed by the prompt “*We need to move on.*”
- **Condition D (Stressor Isolation).** Each participant experienced one variant of condition B in which a single stressor dimension was elevated. Participants were randomly assigned to one of three sub-conditions: D-1 increased audience size from 8 to 17 members; D-2 replaced neutral animations with negative animations; D-3 replaced clarification questions with adversarial questions. By comparing each participant’s condition D responses against their own condition B responses, we could characterize the isolated effect of each stressor through quantitative and qualitative analysis.

The presentation order of the four conditions was counterbalanced across participants using a balanced Latin square [26] with four orderings (O1–O4) to mitigate learning and fatigue effects. Speech topics were randomly assigned to conditions for each participant to avoid confounding topic difficulty with condition effects (see Table 2 in Appendix B).

### 4.3 Apparatus

Participants wore a Meta Quest 3S headset operating in tethered PC-VR mode. The virtual scene rendered a conference-style presentation room where audience nonverbal behaviors and verbal interactions were configured per condition in real time via the Unity Inspector control panel. In-VR audio was captured by the headset’s built-in microphone for both the system’s live speech pipeline and post-hoc acoustic analysis. An external camera recorded video for adding the context of evaluation to participants.

### 4.4 Tasks

Participants performed a series of two-minute impromptu speeches on controversial topics designed to elicit genuine deliberation and argumentation. Four topics of comparable difficulty were used: (1) paper-based books vs. electronic versions, (2) houses vs. apartments when choosing a place to live, (3) work-from-home vs. onsite work, and (4) pros and cons of owning a car. Topics were counter-balanced and randomly assigned to conditions (see Table 2 in Appendix B). We selected these topics because they are universally relatable, require no specialized domain knowledge, and naturally afford dual-sided argumentation. This ensured that participants could immediately generate impromptu content without cognitive blocks

related to topic unfamiliarity, thereby isolating the cognitive load induced specifically by the audience stressors.

To simulate high-stakes environments, we informed participants that their speeches were being recorded and later evaluated. This was necessary to raise the personal stakes and induce authentic performance anxiety and was debriefed at the end of the session.

## 4.5 Procedure

Each 75-minute session followed a four-phase protocol:

*Preparation (10 min).* After consent, participants completed demographics and the PRPSA trait anxiety assessment.

*Training (10 min).* Participants used a Meta Quest 3S to enter a practice environment (3 members, neutral behavior). They completed a 1-minute speech and two clarification questions to learn the spatial interaction mechanics (as described in Section 3.3).

*Experiment Cycles ( $4 \times 10$  min).* Each cycle proceeded as follows. The experimenter loaded the assigned condition and speech topic via the control panel. The participant was given 30 seconds to mentally prepare. The VR scene was then initiated and the participant delivered a 2-minute impromptu speech under the designated condition. During conditions involving questions (B, C, D), the Virtual Director triggered two audience questions at the 60 s and 30 s remaining checkpoints, segmenting the delivery into three temporal phases (pre-Q1, post-Q1, post-Q2). Upon completion, the participant removed the headset and completed post-task questionnaires on a separate laptop. This cycle was repeated for all four conditions.

*Interview & Debriefing (15 min).* Participants completed an exit survey and a forced-choice stressor ranking, followed by a semi-structured interview regarding their experience. The session concluded with a full debriefing revealing no actual evaluations.

## 4.6 Measures & Analysis

We collected data including subjective self-report scales, objective acoustic speech features, and qualitative interview feedback.

**4.6.1 Technical evaluation of system performance.** To assess the technical feasibility and reliability of our pipeline under realistic conditions, we conducted a post-hoc analysis using the recorded audio and system timestamps (showing question trigger initiation and completion of best question selection, captured for latency tracking) from the user study ( $N = 12$  and 72 total Q&A trials). We evaluated system latency by replaying the audio through the backend pipeline. Additionally, we evaluated question quality by computing dense retrieval-based topical similarity using the BGE model (bge-large-en-v1.5) [65] to verify topical grounding against accumulated transcripts, alongside inter-question similarity using Semantic Textual Similarity (STS) via SBERT [48] to check for diversity. Finally, we validated the sufficiency of speech content for acoustic analysis by comparing active phonation ratios across the 60 s baseline and 30 s inter-checkpoint windows.

**4.6.2 Subjective Self-Report Measure.** After each condition, we measured participants' psychological and cognitive responses to the VR stressors through a set of validated self-report instruments, capturing state anxiety, task workload, self-rated speaking performance, and underlying psychological appraisals.

- **State Anxiety.** We measured momentary anxiety intensity after each condition using the Subjective Units of Distress Scale (SUDS) [63] on a 0–100 scale.
- **Workload.** We used NASA Task Load Index (NASA-TLX) [20, 21] after each condition to measure participants' workload using 0–100 scale.
- **Self-Rated Performance.** We measured participants' perceived speaking performance using surveys adapted from the Public Speaking Competence Rubric (PSCR) [51], including word choice, vocal expression, nonverbal behavior, and audience adaptation (each on a 1–7 scale). We also collected an additional single-item overall performance rating.
- **Appraisal and Control.** We included four supplementary items to assess the psychological mechanisms underlying stress responses: feelings of being judged, feelings of being evaluated, worry about impression management, and perceived threat (each on a 1–7 scale). We also measured challenge appraisal ("*The condition felt like a challenge I could handle*") and perceived control ("*I felt in control of my speech*").

We used paired t-tests to evaluate the impact of the combined stressor gradient (Conditions A, B, and C) on state anxiety, workload, and psychological appraisals. We triangulated within-subject state anxiety and workload deltas ( $\Delta = D - B$ ) with qualitative interview insights to evaluate the contributions of individual stressor.

**4.6.3 Objective Speech Performance Measure.** Speech recordings were segmented by temporal phase (pre-Q1, post-Q1, post-Q2) to analyze how participants' vocal performance was impacted over time. All acoustic features were extracted using Praat/Parselmouth [23] from the audio recordings.

- **Acoustic Features.** We extracted jitter (%) as a robust marker of laryngeal instability and physiological stress [57].
- **Temporal Features.** We computed the pause rate (silent pauses per minute) to characterize speech fluency.
- **Linguistic Features.** We computed filler word rate (fillers per minute) using pattern-matching for common disfluency markers [10] (e.g., *um, uh, like, you know*) and hedging expressions (e.g., *I mean, I guess, basically*).

**4.6.4 Exit survey.** A post-experiment exit survey assessed each design element's perceived noticeability and stress on 7-point scales. Participants were asked to perform a forced ranking of four distinct stressor categories: audience size, body language, verbal interruptions, and time pressure. Both our pilot study and prior work demonstrate that the (in)visibility of a timer can introduce psychological stress [6, 62]. We thus included a visible countdown timer across all conditions, and measured its perceived impact in the forced ranking to confirm its baseline stressor presence.

**4.6.5 Interview Data.** We conducted semi-structured interviews that covered participants' phenomenological experiences of the virtual audience, probing specifically into which audience behaviors felt most realistic, how the live verbal interruptions disrupted their train of thought, and their perceived ability to recover from hostile questions. All interviews were transcribed and analyzed using semantic analysis.

## 5 Results

### 5.1 Technical Evaluation

To answer RQ1, we report the system’s latency, generated question quality, and speech segment sufficiency below.

**5.1.1 Latency and reliability.** We evaluated system performance using two latency metrics. *Generation latency*—the time from a 10 s speech-segment upload to the arrival of a candidate question from the LLM—was computed by replaying all recorded audio through the backend pipeline offline. Across 447 generated candidate questions, the generation latency was  $M = 8.96$  s ( $SD = 3.10$  s). In addition, *Q&A latency*—from Virtual Director question selection initiation through completed TTS delivery, measured by using system log timestamps from the live study recordings—was  $M = 12.28$  s ( $SD = 2.92$  s) across all 72 asked questions.

**5.1.2 Question relevance and diversity.** To evaluate question relevance, we computed the dense retrieval-based topical similarity between the accumulated transcripts and generated questions across two checkpoint (C1: 60 s context; C2: 90 s context;  $n = 36$  sessions each). The questions demonstrated strong topical grounding (C1  $\rightarrow$  Q1:  $M = 0.61$ ,  $SD = 0.10$ ; C2  $\rightarrow$  Q2:  $M = 0.65$ ,  $SD = 0.08$ ), confirming that the AI-generated questions remained anchored to the presenter’s topic. Furthermore, to assess question diversity, we computed the Semantic Textual Similarity (STS) between the two questions generated within the same presentation session. The inter-question similarity remained low ( $M = 0.32$ ,  $SD = 0.18$ ), indicating that the system successfully injected novel angles rather than repeating previous inquiries. Please refer to Appendix D for concrete examples of generated questions across conditions.

**5.1.3 Speech segment sufficiency.** We validated the sufficiency of speech content captured within our shorter analysis windows by comparing the word count and active phonation ratio between the 60 s baseline window (session start to C1) and the 30 s inter-checkpoint window (C1 to C2). For the 60 s window, the word count was  $M = 174.78$  ( $SD = 32.21$ ) and the active phonation ratio was  $M = 58.46\%$  ( $SD = 7.04\%$ ). This yields approximately 35.1 s of active phonation (roughly 3,508 valid acoustic frames using a 10 ms step at 44.1 kHz). For the shorter 30 s window, the word count was  $M = 72.83$  ( $SD = 24.37$ ) with a phonation ratio of  $M = 56.12\%$  ( $SD = 8.78\%$ ), corresponding to 16.8 s of active speech, or approximately 1,684 valid frames. Despite the shorter duration, 1,684 frames provide a highly stable and robust signal for low-level acoustic feature extraction (e.g., jitter), aligning well with segment lengths established in prior VR public-speaking research [9, 41].

### 5.2 Combined Stressor Efficacy: Heightened Anxiety due to Perceived Judgment

To answer RQ2, we report participants’ perceived anxiety impacted by multi-stressor conditions of varying intensity (Condition A, B, C). As depicted in Fig. 4a, the positive (Condition A) and neutral audiences (Condition B) kept participants’ anxiety aligned with or below their initial pre-session baselines. However, introducing negative non-verbal behaviors alongside adversarial questions (Condition C) significantly increased their perceived anxiety (A vs. C:  $p < .01$ ;

**Table 1: Mean within-subject deltas ( $\Delta = D - B$ ) for each isolated stressor group ( $n = 4$  per group).**

| Stressor Group                       | $\Delta$ State Anxiety | $\Delta$ TLX Overall |
|--------------------------------------|------------------------|----------------------|
| D1 (Crowd Size 8 $\rightarrow$ 17)   | +2.50                  | +0.42                |
| D2 (Neutral $\rightarrow$ Negative)  | +32.50                 | +30.62               |
| D3 (Clarify $\rightarrow$ Hostile Q) | +5.00                  | +11.04               |

B vs. C:  $p = .02$ ). Correspondingly, in terms of performance, participants used significantly more filler words in Condition C, indicating speech disfluency, while there is no significant difference in their own perceived performance. Please refer to Appendix A, Fig. 7 for the complete performance results. Furthermore, the NASA-TLX (Fig. 4b) show that Condition C led to higher mental workload and frustration, while other dimensions such as physical and temporal demands remained relatively stable across conditions.

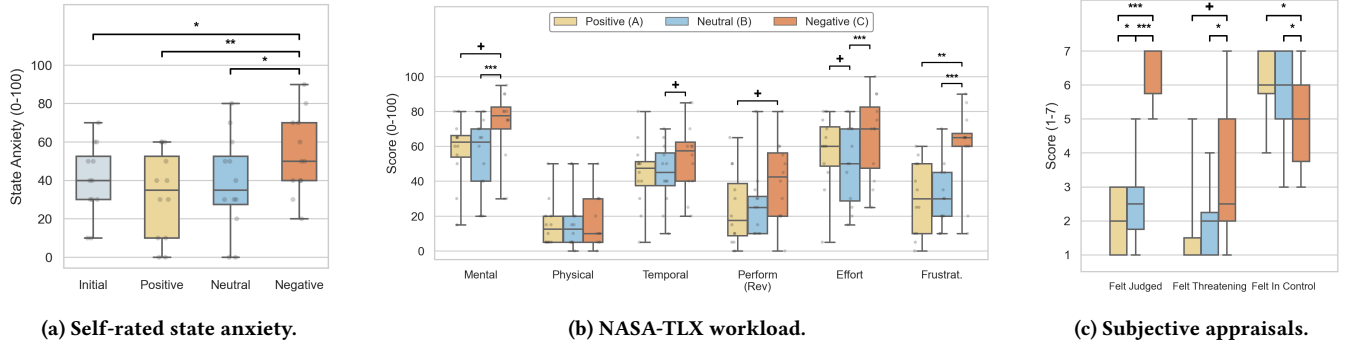
Our qualitative interviews suggest that participants felt the most anxiety in the negative condition as they perceived being judged. Specifically, participants consistently described Condition C not merely as a cognitive distraction, but as an intense social-evaluative threat. P5 noted: “*I felt they were judging me because they were already shaking their heads [before I made a point].*” This premature disagreement was particularly destabilizing because it preceded any arguable claim, undermining the speaker’s sense of legitimacy. The emotional register of these reports is notable; P3 described the adversarial questions as “*not difficult to answer, but [...] you were just pissed off,*” distinguishing the affective burden of perceived hostility from its cognitive difficulty.

Our post-condition surveys also confirmed the feelings of being judged (see Fig. 4c). Under Condition C, participants’ reported significantly higher *Feeling of Judgment* ( $p < .001$ ), with a noticeable lower *Perceived Control* ( $p = .02$ ). These results indicate that negative animations with adversarial questions could heighten participants’ feelings of being judged, and thus lead to higher stress.

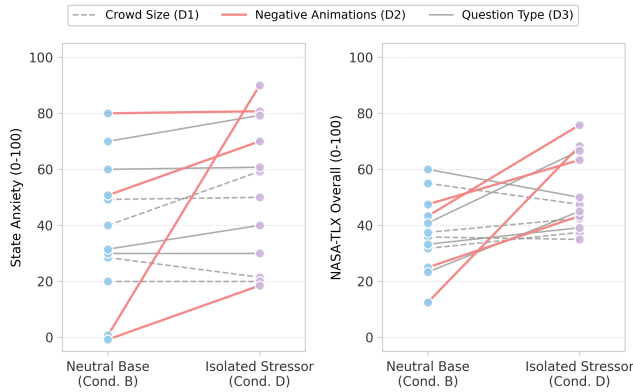
### 5.3 Individual Stressor Contribution

To address RQ3, we evaluated the isolated contribution of each stressor by comparing Conditions B and D. Given the small subsample ( $N = 4$  per group), we present trajectories of individual participant score in Fig. 5 and the within-subject delta changes in Table 1. Notably, we found that all participants exposed to hostile non-verbal behavior (Condition D2) exhibited noticeably higher state anxiety ( $mean\Delta = +32.50$ ) and overall workload ( $mean\Delta = +30.62$ ) compared to the baseline condition. In contrast, trajectories for the increased crowd size (Condition D1) or isolated adversarial questions (Condition D3) remained visually flat. This suggest that hostile non-verbal behavior acted as a stronger source of perceived anxiety and cognitive workload. Our qualitative findings below reveal the patterns behind this.

**5.3.1 Crowd anonymity.** We found that the negligible effect of crowd size (D1) was partly attributed to the virtual crowd’s anonymity and lack of relational stakes. P3 articulated: “*Audience size doesn’t matter because I don’t know any of them. In realistic speaking, people feel pressure because it’s a job interview [...] but there’s no context here.*” P5 agreed: “*If you’re going to a meeting room with 10 people who are real, you would definitely feel that stress. But at this point,*



**Figure 4: Self-reported anxiety, workload, and appraisals across the combined stressor conditions. Individual scores are overlaid as dots. Significant pairwise differences are denoted by symbols (+:  $0.05 \leq p < 0.10$ , \*:  $p < 0.05$ , \*\*:  $p < 0.01$ , \*\*\*:  $p < 0.001$ ).**



**Figure 5: Paired individual trajectories of State Anxiety and NASA-TLX from Baseline (B) to Isolated Stressor (D).**

*I don't know any of them [...] I wouldn't feel that stress.*" This suggests the negligible effect of crowd size was not a failure of the VR medium, but an artifact of the audience's anonymity. Furthermore, participants commented that audience size can be visually less salient compared to audience animations, which they internalized as potential social evaluative threats. As P5 noted: *"I didn't pay much attention on the size of the audience [...] If you look at the audience and they're shaking their heads, that's gonna be a big stress for me."*

**5.3.2 Continuous negative audience animation acts as a strong source of acute anxiety.** Participants reported that negative audience animations (D2) produced a visceral, immediate, and persistent sense of disapproval. Unlike verbal interruptions that occur at discrete temporal checkpoints, negative animations (e.g., constant head-shaking and dismissive gestures) created a persistent atmosphere of disapproval. This visual feedback offered speakers no temporal relief, consumed their attentional bandwidth, and forced them to internally second-guess themselves rather than plan their upcoming speech. For instance, P5 mentioned: *"If you look at the audience and they're shaking their heads, that's gonna be a big stress... instead of thinking what I was going to speak, I was thinking back to what I spoke [and asking myself], did I say anything wrong?"* This was also

echoed by P7, who noted that seeing the audience *"shaking their heads and throwing up their hands [as if the presenter] doesn't know what he's talking about"* made the experience intensely stressful.

**5.3.3 Adversarial questions shift focus to cognitive defense.** Our results show that isolating adversarial questions (D3) elevated perceived workload and state anxiety, compared to neutral clarification questions (see Fig. 5 and Table. 1). Participants found that clarification questions were constructive engagement. For instance, P10 found them *"encouraging to just go in depth on the topic"* and thus *"less stressful"*, while P11 felt *"more open to seeing their viewpoint"*. However, adversarial questions forced participants out of this collaborative mindset into a defensive posture. For example, P12 noted that *"clarifying questions weren't that disruptive, because [...] that's a normal thing to do"*, whereas *"hostile questions are very odd and that threw me off my game more than anything"*. P10 also noted: *"In the combative questions, I felt like I had to defend my choice, my stance. If I didn't answer strongly enough, then I'm wrong."*

Meanwhile, participants also reported that such cognitive defense was less stressful than negative animations (D2) since interruptions were a discrete task rather than a continuous evaluative threat. For example, P1 stated: *"When they interrupt me with a question, I stop and I can see what they're asking [...] but while I'm speaking and seeing them [react negatively] at the same time, it distracts me more than if I stopped talking."*

## 5.4 Interruption Recovery and Training Utility

To investigate participants' perception and coping strategies for cognitive interruptions (RQ4), we examine participants' perceived *recovery*, the cognitive and emotional effort required to regain narrative flow and speech composure immediately following an unexpected interruption, alongside their in-situ coping strategies and perceptions of the system's training utility.

**5.4.1 The context switching cost and coping strategies.** Participants described the Q&A interruption as a disruptive, forced *context switch* that severely taxed their recovery deficit. Regardless of whether an interruption was hostile or polite, the mere act of halting a speech and forcing speakers to process and answer an unscripted prompt broke the continuity of their mental planning. For example, P10 reported such cognitive disconnect of losing their train of

thought with almost every question: “*I would finish answering the question and then think, ‘oh shoot, where did I [stop]’*.” This forces the speaker to expend extra cognitive resources just to re-anchor their narrative thread, as P11 noted that the disruption often forced them to “*backtrack sometimes on my points and reiterate things.*”

To manage this acute cognitive load and facilitate their own recovery, participants reported employing specific in-situ coping strategies. A primary tactic was *attentional redirection*, where participants deliberately altered their visual focus to mitigate environmental stress while trying to re-anchor their train of thought. For example, P10 noted “*I have to pretend this [audience] isn’t even in the room [...] to continue speaking.*” P11 similarly described selectively attending to non-hostile members to create a psychologically safe focal point: “*I would just focus on an audience member that wasn’t really doing that [shaking their head] [...] so it felt like I was speaking to my personal audience.*”

Another coping strategy was to slow down their speech pacing to buy the cognitive time necessary to recover and regain composure, rather than striving for immediate fluency. For example, P8 explicitly recounted, “*I had stumbled a little bit more and had to re-collect myself multiple times.*” P4 similarly described how non-verbal hostility forced an immediate pacing adjustment: “*I was slowing down my pace of speech and [...] I just [...] hesitate to continue my speech.*” This illustrates that the recovery process is not merely a passive delay, but an active, cognitively demanding period where speakers must continuously deploy psychological (e.g., gaze shifting) and behavioral (e.g., compensatory pacing) coping strategies to survive the interruption and resume their flow.

**5.4.2 The training utility of the system.** When asked about the training utility of the system, 9 of 12 participants identified negative condition as the most useful for public speaking training, despite experiencing the highest levels of stress and mental workload. P8 reasoned: “*You should go for the worst-case scenario.*” P1 agreed: “*If anything gets thrown at me in real life, I would be able to adapt.*” Notably, most participants did not want *pure* hostility from all audience. For example, P3 preferred “*a mixture—one or two hostile and most neutral,*” arguing uniform hostility felt unrealistic. Meanwhile, participants identified realism gaps that potentially restricted the stressfulness of the VR environment. For example, P2 noted the absence of realistic accidents or consequences: “*When it’s a real-life scenario [...] anything can go wrong because it’s reality, not fantasy.*” To introduce realistic cognitive load that mirrors real-life public speaking scenarios, participants proposed adding environmental disruptions (e.g., P12 suggested “*a fire alarm*”, and P1 noted “*someone entering the room*”) or visual support materials (e.g., P11 mentioned adding “*a working slide deck in the back*”).

## 6 Discussion

Overall, our results demonstrate that self-rated anxiety increased alongside combined stressor intensity, suggesting that our parameterized design successfully induces distinct stress levels. Our neutral and positive condition results align with prior VR exposure therapy work (e.g., [27, 67]), demonstrating that supportive virtual environments can effectively habituate users and maintain low baseline anxiety.

Among the individual stressors evaluated, the results suggest that negative audience animation is a stronger source of anxiety compared to increased crowd size or interruptive Q&A, as speakers internalized negative non-verbal feedback as social-evaluative threat [13]. This aligns with prior work in self-presentation theory, which posits that social anxiety stems primarily from the prospect of interpersonal evaluation [49], and corroborates findings that non-verbal cues can exert an outsized influence on stress [17, 41].

Additionally, our results show that AI-generated questions increased cognitive load by forcing context switching. Participants thus managed composure through selective gaze adjustment and deliberate pacing shifts. Finally, despite generating the highest stress levels, the negative audience condition was rated by participants as the most valuable environment for training.

### 6.1 Design Implications

Drawing on our mixed-methods results, we outline potential opportunities designing interactive stress-induction systems in VR for public speaking training.

**6.1.1 Prioritizing behavioral fidelity over audience scale.** Since negative non-verbal behavior acts as a strong source of perceived anxiety compared to crowd size or interruptive questions, future systems should prioritize expressive and dynamic audience agents over merely filling virtual seats. While prior research shows that increasing virtual audience size can elevate anxiety [34], our findings introduce a novel insight: the efficacy of audience size heavily depends on relational stakes (e.g., anonymity) or saliency. As our qualitative findings revealed, a large but anonymous virtual crowd fails to trigger the social-evaluative threat required to induce stress. A small audience where members behave with distinct intentionality acts as a more potent stressor than a massive, faceless crowd [17, 41]. Future systems should therefore invest in audience identity—assigning roles, stakes, or recognizable personas so the crowd may carry more genuine social consequence. Recent work on generative agents has shown that LLM-driven characters can maintain coherent personalities, form opinions, and react contextually over time [40]. These capabilities could be applied to create virtual audience members that demonstrate more realistic social evaluative stakes.

**6.1.2 Defaulting to productive challenge instead of user comfort.** In terms of system utilities, our findings show that what speakers prefer is not what optimizes their stress or fluency, suggesting systems should not minimize in-session discomfort. This aligns with prior work on performance and learning. For example, Yerkes and Dodson established that moderate physiological or mental arousal can often optimize task performance, whereas too little or too much arousal degrades it [66]. “Desirable difficulties” also demonstrates that introducing cognitive friction and challenges during training often leads to better long-term skill retention and transfer than conditions optimized purely for user comfort [7]. In addition, as participants found that uniform hostility could feel artificial and trigger disengagement, future systems should consider mixed behavioral profiles such as a neutral baseline with occasional stress could enhance realism and allow speakers to practice selective-attention strategies [61].

### 6.1.3 Tracking compensatory strategies as a marker of resilience.

In addition to anxiety measurement, our results also highlight the system's potential as a dynamic training tool. The *compensatory strategies* observed in our results, such as speakers redirecting their eye gaze to supportive audience or deliberately adopting a slower pace to suppress physiological tremor, can serve as an observable marker of resilience under pressure. By tracking these pacing adjustments and how effectively a speaker regains vocal composure (e.g., to stabilize jitter) after an interruption, practitioners can quantify recovery capacity rather than just stress induction. Meanwhile, integrating configurable stressors allows users to build this resilience progressively. Future systems can adaptively introduce harsher questions or more dismissive non-verbal behaviors within a single session as the speaker's delivery stabilizes, or scale the difficulty across repeated sessions, in order to support dynamic training.

### 6.1.4 Scaffolding recovery when introducing cognitive interruptions.

Our results demonstrate that the cognitive cost of an interruptive question extends well beyond the Q&A exchange itself, where participants had challenges with context switching and cognitive defense. This pattern aligns with findings from prior work about interruptions, where task resumption after an unexpected disruption incurs a measurable lag that depends on the timing and cognitive load at the point of interruption [1, 4, 22]. Such work has shown that resumption costs can be reduced by deferring interruptions to natural breakpoints and by providing external cues that reinstate the task context [22]. Translating these principles to public speaking, systems featuring mid-speech interruptions can incorporate post-interruption cognitive scaffolds, such as a brief visual outline of the speaker's last point or a spatial cue indicating where to resume, timed to appear after the Q&A exchange rather than during it. These scaffolds can then be progressively faded as the speaker builds independent recovery capacity.

## 6.2 Limitations

The limitations of our work are primarily related to sample size, the realism of the system, and limited longitudinal effect. With  $N = 12$ , our quantitative results are more exploratory, offering a preliminary characterization of combined and individual stressor effects, rather than establishing a definitive stressor hierarchy. Future work can consider larger samples with greater statistical power to evaluate subtler effects and statistical significance of these stressors. Additionally, taking interruptive questions by aiming a raycast and pressing a button on the controller could introduce cognitive load and make the experience less realistic. Future systems can explore hands-free, verbal triggers to preserve more natural presentation mechanics. Administering all conditions in a single session also raises the possibility of learning or fatigue effects despite the counterbalancing. Finally, we did not collect longitudinal retention or real-world transfer measures. While prior indicate that standard VR exposure therapy transfers to real-life speaking scenarios [33], the transferability of *dynamic resilience*—the ability to recover from active hostilities and interruptions via parameterized VR stressors—requires further exploration. Future studies can therefore investigate whether speakers trained under dynamic stressor simulation exhibit significant better performance or resilience in authentic, high-stakes contexts in the long run.

## 7 Conclusion

We presented a multi-stressor VR public speaking training system that parameterizes audience non-verbal behavior, crowd size, and AI-driven contextual Q&A. Through a mixed-methods, within-subjects user study, we offered a preliminary characterization of how these combined and individual stressors affect perceived anxiety, performance, cognitive workload. Our exploratory findings confirm that self-rated anxiety increased alongside combined stressor intensity. Our results also suggest that audience non-verbal behavior acts as a stronger source of anxiety compared to audience size or AI-generated questions, partly because participants internalized negative animations as social-evaluative threat. Furthermore, our results show that AI-generated questions can increase cognitive load via context switching, and reveal participants' coping strategies such as selective gaze adjustment and deliberate pacing shifts. Finally, the negative audience condition was rated by participants as the most valuable environment for training, despite generating the highest stress levels. We outline key design implications for future VR public speaking training, emphasizing the need to dynamically calibrate stress within a productive zone, scaffold post-stressor recovery, and track adaptive coping strategies as markers of resilience, ultimately fostering real-world skill transfer.

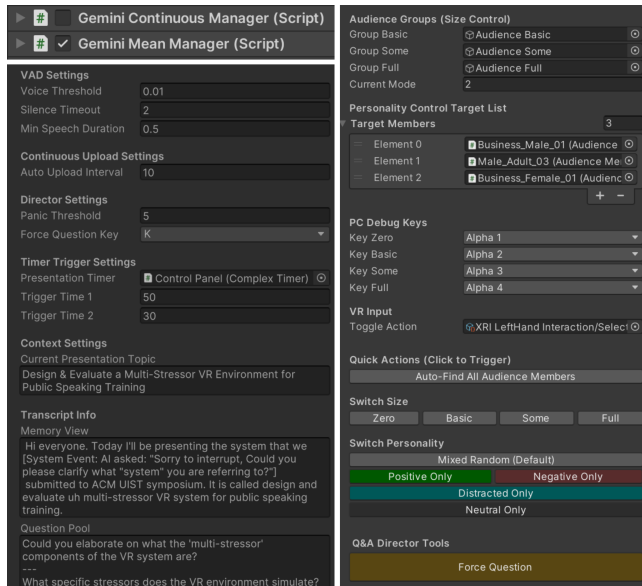
## References

- [1] Piotr D. Adamczyk and Brian P. Bailey. 2004. If not now, when? the effects of interruption at different moments within task execution. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vienna, Austria) (CHI '04). Association for Computing Machinery, New York, NY, USA, 271–278. doi:10.1145/985692.985727
- [2] Fatema Akbar, Ayse Elvan Bayraktaroglu, Pradeep Buddharaju, Dennis Rodrigo Da Cunha Silva, Ge Gao, Ted Grover, Ricardo Gutierrez-Osuna, Nathan Cooper Jones, Gloria Mark, Ioannis Pavlidis, et al. 2019. Email makes you sweat: Examining email interruptions and stress using thermal imaging. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.
- [3] Manuel Bachmann, Abimanju Subramaniam, Jonas Born, and David Weibel. 2023. Virtual reality public speaking training: effectiveness and user technology acceptance. *Frontiers in Virtual Reality* 4 (Sept. 2023). doi:10.3389/frvir.2023.1242544
- [4] Brian P. Bailey and Joseph A. Konstan. 2006. On the need for attention-aware systems: Measuring effects of interruption on task performance, error rate, and affective state. *Computers in Human Behavior* 22, 4 (2006), 685–708. doi:10.1016/j.chb.2005.12.009 Attention aware systems.
- [5] Ligia Batrinca, Giota Stratou, Ari Shapiro, Louis-Philippe Morency, and Stefan Scherer. 2013. *Cicero - Towards a Multimodal Virtual Audience Platform for Public Speaking Training*. Springer Berlin Heidelberg, 116–128. doi:10.1007/978-3-642-40415-3\_10
- [6] Melissa A Birkett. 2011. The Trier Social Stress Test protocol for inducing psychological stress. *Journal of visualized experiments: JoVE* 56 (2011), 3238.
- [7] Robert A. Bjork. 1994. *Memory and Metamemory Considerations in the Training of Human Beings*. The MIT Press, 185–206. doi:10.7551/mitpress/4561.003.0011
- [8] Mathieu Chollet, Nithin Chandrashekar, Ari Shapiro, Louis-Philippe Morency, and Stefan Scherer. 2016. *Manipulating the Perception of Virtual Audiences Using Crowdsourced Behaviors*. Springer International Publishing, 164–174. doi:10.1007/978-3-319-47665-0\_15
- [9] Mathieu Chollet, Torsten Wörtwein, Louis-Philippe Morency, Ari Shapiro, and Stefan Scherer. 2015. Exploring feedback strategies to improve public speaking: an interactive virtual audience framework. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '15)*. ACM, 1143–1154. doi:10.1145/2750858.2806060
- [10] H Clark. 2002. Using uh and um in spontaneous speaking. *Cognition* 84, 1 (May 2002), 73–111. doi:10.1016/s0010-0277(02)00017-3
- [11] Ionut Damian, Chiew Seng (Sean) Tan, Tobias Baur, Johannes Schöning, Kris Luyten, and Elisabeth André. 2015. Augmenting Social Interactions: Realtime Behavioural Feedback using Social Signal Processing Techniques. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, 565–574. doi:10.1145/2702123.2702314
- [12] Taufiq Daryanto, Xiaohan Ding, Lance T. Wilhelm, Sophia Stil, Kirk McInnis Knutsen, and Eugenia H. Rho. 2025. Conversate: Supporting Reflective Learning

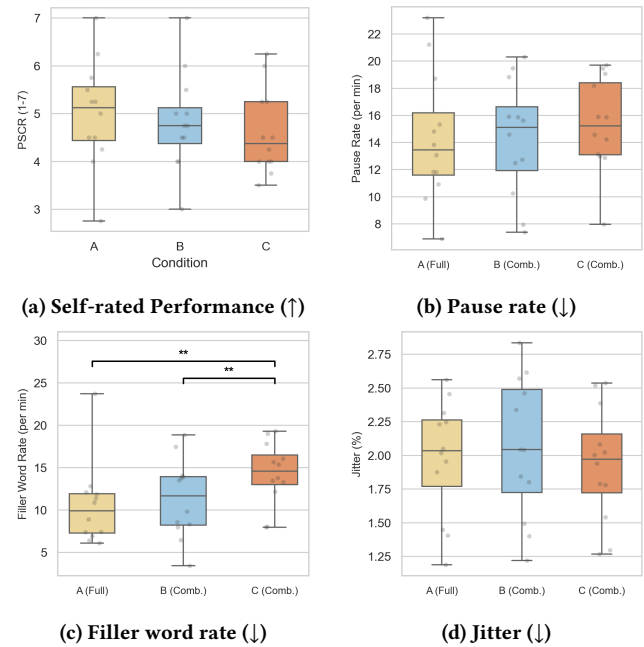
- in Interview Practice Through Interactive Simulation and Dialogic Feedback. *Proceedings of the ACM on Human-Computer Interaction* 9, 1 (Jan. 2025), 1–32. doi:10.1145/3701188
- [13] Sally S. Dickerson and Margaret E. Kemeny. 2004. Acute Stressors and Cortisol Responses: A Theoretical Integration and Synthesis of Laboratory Research. *Psychological Bulletin* 130, 3 (2004), 355–391. doi:10.1037/0033-2909.130.3.355
- [14] ElevenLabs. 2024. ElevenLabs Text-to-Speech API. <https://elevenlabs.io>. Accessed: 2024.
- [15] Anna Fang, Hriday Chhabria, Alekhya Maram, and Haiyi Zhu. 2025. Social Simulation for Everyday Self-Care: Design Insights from Leveraging VR, AR, and LLMs for Practicing Stress Relief. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. ACM, 1–23. doi:10.1145/3706598.3713115
- [16] Bernhard Fehlmann, Fabian D Mueller, Nan Wang, Merle K Ibach, Thomas Schlitt, Dorothee Bentz, Anja Zimmer, Andreas Papasotiropoulos, and Dominique JF de Quervain. 2023. Virtual reality gaze exposure treatment reduces state anxiety during public speaking in individuals with public speaking anxiety: A randomized controlled trial. *Journal of Affective Disorders Reports* 14 (Dec. 2023), 100627. doi:10.1016/j.jadr.2023.100627
- [17] Matteo Gironcini, Ivana Frigione, Mariapia Marra, Milena Stefanova, Margherita Pillan, Angelo Maravita, and Alberto Gallace. 2024. Decoupling the role of verbal and non-verbal audience behavior on public speaking anxiety in virtual reality using behavioral and psychological measures. *Frontiers in Virtual Reality* 5 (March 2024). doi:10.3389/frvir.2024.1347102
- [18] Mar Gonzalez-Franco, Eyal Ofek, Ye Pan, Angus Antley, Anthony Steed, Bernhard Spanlang, Antonella Maselli, Donna Banakou, Nuria Pelechano, Sergio Orts-Escobedo, Veronica Orvalho, Laura Trutoni, Markus Wojcik, Maria V. Sanchez-Vives, Jeremy Bailenson, Mel Slater, and Jaron Lanier. 2020. The Rocketbox Library and the Utility of Freely Available Rigged Avatars. *Frontiers in Virtual Reality* 1 (Nov. 2020). doi:10.3389/frvir.2020.561558
- [19] Google Gemini Team. 2025. *Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities*. Technical Report. Google. arXiv:2507.06261.
- [20] Sandra G. Hart. 2006. Nasa-Task Load Index (NASA-TLX); 20 Years Later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 50, 9 (Oct. 2006), 904–908. doi:10.1177/154193120605000909
- [21] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Human Mental Workload*, Peter A. Hancock and Najmedin Meshkati (Eds.). Advances in Psychology, Vol. 52. North-Holland, 139–183. doi:10.1016/S0166-4115(08)62386-9
- [22] Shamsi T. Iqbal and Brian P. Bailey. 2008. Effects of intelligent notification management on users and their tasks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (Florence, Italy) (CHI '08)*. Association for Computing Machinery, New York, NY, USA, 93–102. doi:10.1145/1357054.1357070
- [23] Yannick Jadoul, Bill Thompson, and Bart de Boer. 2018. Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics* 71 (Nov. 2018), 1–15. doi:10.1016/j.wocn.2018.07.001
- [24] Ni Kang, Willem-Paul Brinkman, M. Birna van Riemsdijk, and Mark Neerincx. 2016. The design of virtual audiences: Noticeable and recognizable behavioral styles. *Computers in Human Behavior* 55 (Feb. 2016), 680–694. doi:10.1016/j.chb.2015.10.008
- [25] Ni Kang, Willem-Paul Brinkman, M. Birna van Riemsdijk, and Mark A. Neerincx. 2013. An Expressive Virtual Audience with Flexible Behavioral Styles. *IEEE Transactions on Affective Computing* 4, 4 (Oct. 2013), 326–340. doi:10.1109/taffc.2013.2297104
- [26] Roger E. Kirk. 2012. *Experimental Design*. SAGE Publications, 1354 pages. doi:10.4135/9781483384733
- [27] Evelyn Klinger, Stéphane Bouchard, Patrick Légeron, Stéphane Roy, Françoise Lauer, Isabelle Chemin, and Pierre Nuges. 2005. Virtual reality therapy versus cognitive behavior therapy for social phobia: A preliminary controlled study. *Cyberpsychology & behavior* 8, 1 (2005), 76–88.
- [28] M Krijn, P.M.G Emmelkamp, R.P. Olafsson, and R Biemond. 2004. Virtual reality exposure therapy of anxiety disorders: A review. *Clinical Psychology Review* 24, 3 (July 2004), 259–281. doi:10.1016/j.cpr.2004.04.001
- [29] Joanne Leong, Florian Perteneder, Muhender Raj Rajvee, and Pattie Maes. 2023. “Picture the Audience...”: Exploring Private AR Face Filters for Online Public Speaking. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. ACM, 1–13. doi:10.1145/3544548.3581039
- [30] Gloria Mark, Daniela Gudith, and Ulrich Klocke. 2008. The cost of interrupted work: more speed and stress. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 107–110.
- [31] James C. McCroskey. 1970. Measures of communication-bound anxiety. *Speech Monographs* 37, 4 (Nov. 1970), 269–277. doi:10.1080/03637757009375677
- [32] Beatriz Mello, Robin Welsch, Marissa Christien Verbokkem, Pascal Knierim, and Martin Johannes Dechant. 2024. Navigating the Virtual Gaze: Social Anxiety’s Role in VR Proxemics. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. ACM, 1–15. doi:10.1145/3613904.3642359
- [33] Nexmedin Morina, Hiske Ijntema, Katharina Meyerbröcker, and Paul M.G. Emmelkamp. 2015. Can virtual reality exposure therapy gains be generalized to real-life? A meta-analysis of studies applying behavioral assessments. *Behaviour Research and Therapy* 74 (Nov. 2015), 18–24. doi:10.1016/j.brat.2015.08.010
- [34] Fariba Mostajeran, Melik Berk Balci, Frank Steinicke, Simone Kühn, and Jürgen Gallinat. 2020. The Effects of Virtual Audience Size on Social Anxiety during Public Speaking. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. 303–312. doi:10.1109/VR46266.2020.00050
- [35] Prasanth Murali, Ha Trinh, Lazlo Ring, and Timothy Bickmore. 2021. A Friendly Face in the Crowd: Reducing Public Speaking Anxiety with an Emotional Support Agent in the Audience. In *Proceedings of the 21th ACM International Conference on Intelligent Virtual Agents (IVA '21)*. ACM, 156–163. doi:10.1145/3472306.3478364
- [36] David Oprea, Sebastian Pintea, Azucena Garcia-Palacios, Cristina Botella, Ștefan Szamoskózi, and Daniel David. 2011. Virtual reality exposure therapy in anxiety disorders: a quantitative meta-analysis: Virtual Reality Exposure Therapy. *Depression and Anxiety* 29, 2 (Nov. 2011), 85–93. doi:10.1002/da.20910
- [37] Fabrizio Palmas, Jakub Cichor, David A Plecher, and Gudrun Klinker. 2019. Acceptance and effectiveness of a virtual reality public speaking training. In *2019 IEEE international symposium on mixed and augmented reality (ISMAR)*. IEEE, 363–371.
- [38] Fabrizio Palmas, Ramona Reinelt, Jakub E Cichor, David A Plecher, and Gudrun Klinker. 2021. Virtual reality public speaking training: Experimental evaluation of direct feedback technology acceptance. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*. IEEE, 463–472.
- [39] Jeongeun Park and DaEun Choi. 2023. AudiLens: Configurable LLM-Generated Audiences for Public Speech Practice. In *Adjunct Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*. ACM, 1–3. doi:10.1145/3586182.3625114
- [40] Joon Sung Park, Joseph O’Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*. ACM, 1–22. doi:10.1145/3586183.3606763
- [41] David-Paul Pertaub, Mel Slater, and Chris Barker. 2002. An Experiment on Public Speaking Anxiety in Response to Three Different Types of Virtual Audience. *Presence: Teleoperators and Virtual Environments* 11, 1 (feb 2002), 68–78. doi:10.1162/105474602317343668
- [42] Sandra Poeschl. 2017. Virtual Reality Training for Public Speaking—A QUEST-VR Framework Validation. *Frontiers in ICT* 4 (June 2017). doi:10.3389/fict.2017.00013
- [43] Sandra Poeschl and Nicola Döring. 2012. *Designing Virtual Audiences for Fear of Public Speaking Training – An Observation Study on Realistic Nonverbal Behavior*. IOS Press. doi:10.3233/978-1-61499-121-2-218
- [44] Mark B. Powers and Paul M.G. Emmelkamp. 2008. Virtual reality exposure therapy for anxiety disorders: A meta-analysis. *Journal of Anxiety Disorders* 22, 3 (April 2008), 561–569. doi:10.1016/j.janxdis.2007.04.006
- [45] Charles B Pull. 2012. Current status of knowledge on public-speaking anxiety. *Current opinion in psychiatry* 25, 1 (2012), 32–38.
- [46] Ajeng Intan Nur Rahmawati, Imam Arifuddin, and Mulawarman Mulawarman. 2018. Psychological Experience Dynamics of Students with Glossophobia through Narrative Counseling as seen from Gender: A Qualitative Study. In *Proceedings of the 3rd ASEAN Conference on Psychology, Counselling, and Humanities (ACPCH 2017) (acpch-17)*. Atlantis Press. doi:10.2991/acpch-17.2018.2
- [47] Ashwin Ram, Varsha Suresh, Artin Saberpour Abadian, Vera Demberg, and Jürgen Steimle. 2025. GestureCoach: Rehearsing for Engaging Talks with LLM-Driven Gesture Recommendations. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology (UIST '25)*. ACM, 1–15. doi:10.1145/3746059.3747705
- [48] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 3982–3992. doi:10.18653/v1/D19-1410
- [49] Barry R. Schlenker and Mark R. Leary. 1982. Social anxiety and self-presentation: A conceptualization model. *Psychological Bulletin* 92, 3 (Nov. 1982), 641–669. doi:10.1037/0033-2909.92.3.641
- [50] Jan Schneider, Dirk Börner, Peter van Rosmalen, and Marcus Specht. 2015. Presentation Trainer, your Public Speaking Multimodal Coach. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction (ICMI '15)*. ACM, 539–546. doi:10.1145/2818346.2830603
- [51] Lisa M. Schreiber, Gregory D. Paul, and Lisa R. Shibley. 2012. The Development and Test of the Public Speaking Competence Rubric. *Communication Education* 61, 3 (April 2012), 205–233. doi:10.1080/03634523.2012.670709
- [52] Omar Shaikh, Valentino Emil Chai, Michele Gelfand, Diyi Yang, and Michael S. Bernstein. 2024. Rehearsal: Simulating Conflict to Teach Conflict Resolution. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. ACM, 1–20. doi:10.1145/3613904.3642159
- [53] Mel Slater, David-Paul Pertaub, Chris Barker, and David M. Clark. 2006. An Experimental Study on Fear of Public Speaking Using a Virtual Environment.

- CyberPsychology & Behavior* 9, 5 (Oct. 2006), 627–633. doi:10.1089/cpb.2006.9.627
- [54] Marcel Takac, James Collett, Kristopher J Blom, Russell Conduit, Imogen Rehm, and Alexander De Foe. 2019. Public speaking anxiety decreases within repeated virtual reality training sessions. *PloS one* 14, 5 (2019), e0216288.
- [55] M. Iftekhar Tanveer, Emy Lin, and Mohammed (Ehsan) Hoque. 2015. Rhema: A Real-Time In-Situ Intelligent Interface to Help People with Public Speaking. In *Proceedings of the 20th International Conference on Intelligent User Interfaces (IUI'15)*. ACM, 286–295. doi:10.1145/2678025.2701386
- [56] Karl Halvor Teigen. 1994. Yerkes-Dodson: A law for all seasons. *Theory & Psychology* 4, 4 (1994), 525–547.
- [57] Ingo R. Titze. 1994. Toward standards in acoustic analysis of voice. *Journal of Voice* 8, 1 (March 1994), 1–7. doi:10.1016/s0892-1997(05)80313-3
- [58] H. Trinh, R. Asadi, D. Edge, and T. Bickmore. 2017. RoboCOP: A Robotic Coach for Oral Presentations. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 2 (June 2017), 1–24. doi:10.1145/3090092
- [59] Jinping Wang, Hyun Yang, Ruosi Shao, Saeed Abdullah, and S. Shyam Sundar. 2020. Alexa as Coach: Leveraging Smart Speakers to Build Social Agents that Reduce Public Speaking Anxiety. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. ACM, 1–13. doi:10.1145/3313831.3376561
- [60] Xingbo Wang, Haipeng Zeng, Yong Wang, Aoyu Wu, Zhida Sun, Xiaojuan Ma, and Huamin Qu. 2020. VoiceCoach: Interactive Evidence-based Training for Voice Modulation Skills in Public Speaking. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. ACM, 1–12. doi:10.1145/3313831.3376726
- [61] Theresa F. Wechsler, Michael Pfaller, Rahel E. van Eickels, Luise H. Schulz, and Andreas Mühlberger. 2021. Look at the Audience? A Randomized Controlled Study of Shifting Attention From Self-Focus to Nonsocial vs. Social External Stimuli During Virtual Reality Exposure to Public Speaking in Social Anxiety. *Frontiers in Psychiatry* 12 (Dec. 2021). doi:10.3389/fpsy.2021.751272
- [62] Sebastian Weiß and Wilko Heuten. 2023. Don't Panic! - Influence of Virtual Stressor Representations from the ICU Context on Perceived Stress Levels. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. ACM, 1–15. doi:10.1145/3544548.3581189
- [63] Joseph Wolpe. 1990. *The Practice of Behavior Therapy*. Allyn & Bacon. 420 pages. [http://books.google.com/books?id=wtdAPgAACAAJ&dq=isbn:9780205145140&hl=&source=gbp\\_api](http://books.google.com/books?id=wtdAPgAACAAJ&dq=isbn:9780205145140&hl=&source=gbp_api)
- [64] Torsten Wörtwein, Mathieu Chollet, Boris Schauerte, Louis-Philippe Morency, Rainer Stiefelhagen, and Stefan Scherer. 2015. Multimodal Public Speaking Performance Assessment. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction (ICMI '15)*. ACM, 43–50. doi:10.1145/2818346.2820762
- [65] Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-Pack: Packaged Resources To Advance General Chinese Embedding. arXiv:2309.07597 [cs.CL]
- [66] Robert Mearns Yerkes, John D Dodson, et al. 1908. The relation of strength of stimulus to rapidity of habit-formation. (1908).
- [67] Samaneh Zamanifard and Andrew Robb. 2023. Social Virtual Reality Is My Therapist: Overcoming Social Anxiety Disorder Through Using Social Virtual Reality. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. ACM, 1–6. doi:10.1145/3544549.3585888
- [68] Haipeng Zeng, Xingbo Wang, Yong Wang, Aoyu Wu, Ting-Chuen Pong, and Huamin Qu. 2023. GestureLens : Visual Analysis of Gestures in Presentation Videos. *IEEE Transactions on Visualization and Computer Graphics* 29, 8 (aug 2023), 3685–3697. doi:10.1109/tvcg.2022.3169175
- [69] Lei Zhang, Jin Pan, Jacob Gettig, Steve Oney, and Anhong Guo. 2024. VRCopilot: Authoring 3D Layouts with Generative AI Models in VR. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology* (Pittsburgh, PA, USA) (UIST '24). Association for Computing Machinery, New York, NY, USA, Article 96, 13 pages. doi:10.1145/3654777.3676451
- [70] Hangyu Zhou, Yuichiro Fujimoto, Masayuki Kanbara, and Hirokazu Kato. 2021. Virtual Reality as a Reflection Technique for Public Speaking Training. *Applied Sciences* 11, 9 (April 2021), 3988. doi:10.3390/app11093988

## A Supplementary Figures



**Figure 6: Unity Inspector attributes allowing: choose question type to be clarification (up) or adversarial (below); change settings of Director Logic and timestamps to trigger audience questions corresponding to the set current topic; runtime control of major parameters: audience size, behavioral profiles; select audience members; force questions.**



**Figure 7: Self-rated performance and objective acoustic metrics in the full/combined speech through conditions.**

## B Supplementary Tables

**Table 2: Complete experimental schedule across all participants, counterbalanced using a balanced Latin square design.**

| PID | Knob | Order | Slot 1  | Slot 2  | Slot 3  | Slot 4  |
|-----|------|-------|---------|---------|---------|---------|
| 1   | D1   | O1    | T1 (A)  | T2 (C)  | T3 (B)  | T4 (D1) |
| 2   | D1   | O2    | T3 (C)  | T4 (A)  | T1 (D1) | T2 (B)  |
| 3   | D1   | O3    | T4 (B)  | T3 (D1) | T2 (A)  | T1 (C)  |
| 4   | D1   | O4    | T2 (D1) | T1 (B)  | T4 (C)  | T3 (A)  |
| 5   | D2   | O1    | T1 (A)  | T2 (C)  | T3 (B)  | T4 (D2) |
| 6   | D2   | O2    | T3 (C)  | T4 (A)  | T1 (D2) | T2 (B)  |
| 7   | D2   | O3    | T4 (B)  | T3 (D2) | T2 (A)  | T1 (C)  |
| 8   | D2   | O4    | T2 (D2) | T1 (B)  | T4 (C)  | T3 (A)  |
| 9   | D3   | O1    | T1 (A)  | T2 (C)  | T3 (B)  | T4 (D3) |
| 10  | D3   | O2    | T3 (C)  | T4 (A)  | T1 (D3) | T2 (B)  |
| 11  | D3   | O3    | T4 (B)  | T3 (D3) | T2 (A)  | T1 (C)  |
| 12  | D3   | O4    | T2 (D3) | T1 (B)  | T4 (C)  | T3 (A)  |

*Note:* T1–T4 represent the randomly assigned speech topics. A, B, C, and D represent the experimental conditions. Condition D encompasses the three specific knob groups ( $D_1$ ,  $D_2$ ,  $D_3$ ) tested across participants.

## C LLM Prompts for Speech Transcription and Verbal Interruptions

|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Prompt A: Transcription and Clarification Question Generation</b><br><b>System Instruction:</b><br><i>Role:</i> You are a professional stenographer.<br><i>Input:</i> CURRENT TOPIC, HISTORY (Context), NEW AUDIO (Speech segment).<br><i>Task 1:</i> Transcribe NEW AUDIO verbatim (word-for-word). No hallucinations.<br><i>Task 2:</i> Generate ONE potential clarification question based on this segment and the current topic.<br><i>Output Format:</i> [Verbatim Transcription]     [Question OR 'NO_QUESTION']<br><b>User Input Template:</b><br>CURRENT TOPIC: <topic_variable><br>HISTORY: <accumulated_transcript_variable><br>(Followed by the raw audio byte stream) |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|

### Prompt B: Virtual Director's Question Selection Logic

#### System Instruction:

*Role:* You are the moderator of a presentation Q&A.

*Task:*

- (1) Review the SPEECH CONTEXT and the CANDIDATE QUESTIONS list.
- (2) Select the **SINGLE** most relevant, insightful, and natural question.
- (3) **Constraint:** Do not rephrase or combine ideas; keep original wording.
- (4) **Constraint:** Avoid repeating questions that have been previously addressed.
- (5) **Output Rule:** Output **ONLY** the final question text with the mandatory prefix.

#### Output Format:

"Sorry to interrupt" ("Sorry I cannot understand you" for hostile audience), [selected\_question\_text]

#### User Input Template:

SPEECH CONTEXT: <full\_transcript\_variable>

CANDIDATE QUESTIONS POOL: <list\_of\_generated\_questions>

### Prompt C: Transcription and Hostile Question Generation (Stressor Logic)

#### System Instruction:

*Persona:* You are a professional stenographer but a **strict reviewer/audience member**.

*Input:* CURRENT TOPIC, HISTORY (Context), NEW AUDIO (Speech segment).

*Task 1:* Transcribe NEW AUDIO verbatim (word-for-word). No hallucinations.

*Task 2:* Generate ONE clarification question with a **hostile tone** (acting as a picky boss) based on this segment. Alternatively, state an opinion that contradicts the presenter to challenge their ideas.

*Constraint:* Do not personally attack the presenter; focus on the content.

*Output Format:* [Verbatim Transcription] ||| [Question OR 'NO\_QUESTION']

#### User Input Template:

CURRENT TOPIC: <topic\_variable>

HISTORY: <accumulated\_transcript\_variable>

(Followed by the raw audio byte stream)

## D Examples of Generated Interruption Questions

Table 3 presents representative examples of the system's generated questions directly extracted from the user studies across the two behavioral conditions. The examples illustrate how the system anchors its questions in the accumulated speech context at Checkpoint 1 (C1, 60 s) and Checkpoint 2 (C2, 90 s). The Clarification examples show the system prompting for missing details or helping the presenter recover from hesitations, while the Adversarial examples demonstrate the system successfully introducing high semantic divergence—challenging the presenter with abrupt, critical viewpoints.

**Table 3: Comparison of AI-generated questions across multiple speech topics, demonstrating distinct behavioral strategies. Excerpts are slightly truncated [...] for brevity.**

| Participant                                                                                                        | Accumulated Speech Context (Transcript Excerpt)                                                                                                                                                                                                                                                                                             | Generated Question (Q1 / Q2)                                                                                                                                                                                                                                                      |
|--------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Example 1: Topic – Electronic vs. Paper Books   Strategy – Clarification (Helps recover and seeks details)</b>  |                                                                                                                                                                                                                                                                                                                                             |                                                                                                                                                                                                                                                                                   |
| P7                                                                                                                 | "Good afternoon, everybody. I'm going to explain a paper book versus an electronic book. [...] I would prefer actually electronic books because, in my opinion, it's <b>less easy to make annotations on a paper book</b> than it is on an electronic book. Because on an electronic book, you could just do a control-find."               | <b>Q1:</b> "Could you elaborate on why you find making annotations on an electronic book easier than on a paper one?"                                                                                                                                                             |
| P7                                                                                                                 | [...Speech continues...] "On the paper book, it actually has an <b>advantage</b> over the electronic book. If let's say you're in a class... um, wait. <b>I'm sorry, I lost my train of thought here.</b> So the paper book, it would be more easier to do the annotations..."                                                              | <b>Q2:</b> "You mentioned that a paper book has an advantage over an electronic one. Could you please elaborate on what that advantage is?"                                                                                                                                       |
| <b>Example 2: Topic – Electronic vs. Paper Books   Strategy – Adversarial (Challenges superficial arguments)</b>   |                                                                                                                                                                                                                                                                                                                                             |                                                                                                                                                                                                                                                                                   |
| P12                                                                                                                | "All right, so on the topic of paper books versus electronic books, I think before we start, there's an important distinction to make between <b>books meaning textbooks and books meaning novels.</b> [...] Me personally, I prefer e-books simply because textbooks are so large."                                                        | <b>Q1:</b> "So you're saying the entire universe of published works neatly fits into either 'textbooks' or 'novels'? Where exactly in that simplistic binary do academic journals, poetry collections, or technical manuals fall?"                                                |
| P12                                                                                                                | [...Speech continues...] "In terms of the other books, novels usually, I tend to prefer an actual physical book, even though I know that you can carry more books on an ebook like a Kindle for travel. But I grew up with normal books, so I prefer having the <b>weight of an actual book</b> in my hand."                                | <b>Q2:</b> "So your entire argument for the superiority of e-textbooks boils down to a logistical complaint about their weight? Are we to ignore all pedagogical research on retention and comprehension from physical media simply because carrying a backpack is inconvenient?" |
| <b>Example 3: Topic – House vs. Apartment   Strategy – Clarification (Refines ambiguous phrasing)</b>              |                                                                                                                                                                                                                                                                                                                                             |                                                                                                                                                                                                                                                                                   |
| P3                                                                                                                 | "For the comparison between apartment and house, I prefer apartments. I'm an architecture guy... housing in the U.S. is a little bit boring. What I mean is, like, it's <b>always same look</b> , and it's like wooden small house. [...] but apartment might be smaller, but it also <b>prefer, like, the better, like, furniture...</b> " | <b>Q1:</b> "When you say U.S. houses have the 'same look', are you referring to a specific style or region?"                                                                                                                                                                      |
| P3                                                                                                                 | [...Speech continues...] "And the second thing is like, although the space is small, but <b>the furniture</b> is usually the newer and it's just more enjoyable to live here. I'm not a guy who needs a huge space to live."                                                                                                                | <b>Q2:</b> "When you say an apartment 'prefers' better furniture, do you mean it typically comes with higher-quality furnishings, or that it's better designed for it?"                                                                                                           |
| <b>Example 4: Topic – Owning a Car vs. Public Transit   Strategy – Adversarial (Attacks logic and definitions)</b> |                                                                                                                                                                                                                                                                                                                                             |                                                                                                                                                                                                                                                                                   |
| P11                                                                                                                | "Having my own car will enable me to go to NJIT <b>much more efficiently</b> because NJIT is a commuter school. [...] Paying for my own car is going to help in that because I'll be <b>building a line of credit</b> and working towards that."                                                                                            | <b>Q1:</b> "How are you defining 'efficiently'? Does your definition conveniently ignore the time lost to traffic, the financial black hole of parking and gas, and the hassle of maintenance...?"                                                                                |
| P11                                                                                                                | [...Speech continues...] "Alright, well, yeah, like she said before, there are <b>some upfront costs of having a car, the car payment</b> , having the down payment on the car... but having the car is actually gonna be much better than using public transport."                                                                         | <b>Q2:</b> "You're framing a car loan as a tool for building credit, but isn't that just a justification for taking on significant debt for a rapidly depreciating asset? How is that a lesson in responsible finance?"                                                           |