HealthTTS: SYMPTOM-AWARE TEXT-TO-SPEECH WITH CLINICAL ACOUSTIC CUES

Anonymous authors
Paper under double-blind review

ABSTRACT

Speech in healthcare interactions conveys not only verbal content but also symptom-related acoustic cues, such as coughs, throat clearing, and sighs. However, our evaluation of representative text-to-speech (TTS) models shows that existing models often fail to reproduce the requested cue type, count, and order, while even correct cue sequences may not integrate naturally into speech. To address these challenges, we propose **HealthTTS**, a symptom-aware TTS model that directly generates complete utterances from text, a speaker reference, and an optional cue configuration. HealthTTS follows a three-stage training procedure that combines cue-aware language-model adaptation, localized acoustic adaptation, and global preference refinement to improve both cue generation and natural cue–speech integration. We construct SymCue, a paired speech resource for training, and **SymCueBench**, an evaluation benchmark covering seven symptom-related acoustic cue types. Under the shared-cue setting, HealthTTS achieves 87.3% Cue Sequence Accuracy, outperforming Fish Speech S2 Pro by 25.1 percentage points while maintaining high Speech Naturalness and Cue Realism. Beyond benchmark evaluation, a virtual standardized patient study with 30 medical students across 14 scenarios shows that HealthTTS achieves higher Patient Realism and Perceived Usefulness than Cue Insertion, a post-hoc insertion reference. To support future research and the open-source community, we will make our code publicly available.

1 INTRODUCTION

In healthcare interaction, patient speech conveys information beyond verbal content. Symptoms may also be expressed through localized vocal manifestations, such as coughing, throat clearing, or pain-related vocalizations. For healthcare interaction systems, faithfully capturing such acoustic cues is therefore crucial, as they complement verbal content and provide a more complete characterization of patients' symptoms. However, collecting speech with symptom-related acoustic cues in real clinical settings is challenging, as patient recordings are difficult to obtain at scale and rarely provide controlled coverage of diverse symptom types, cue combinations.

Accordingly, we focus on these non-lexical vocalizations, including coughs, throat clearing, sighs, hiccups, retching, sneezing, and pain-related groans Liao et al. (2025); Xue et al. (2026), and refer to them as *symptom-related acoustic cues*. Unlike persistent pathological voice or speech characteristics such as hoarseness or dysarthria Halpern et al. (2026); Wang et al. (2026), these cues occur locally within an utterance and may appear multiple times or in different combinations. Symptom-aware TTS therefore requires structured control over cue types, counts, order, and positions while preserving linguistic content, speaker identity, and natural speech quality.

Existing TTS models are primarily designed for general-purpose speech synthesis rather than healthcare-oriented symptom presentation. Although modern models achieve strong intelligibility, naturalness, and speaker consistency Liao et al. (2026); Hu et al. (2026), these capabilities do not necessarily provide reliable structured cue control. Cue Insertion, a post-hoc cue-insertion reference that combines CosyVoice3 Du et al. (2025) speech with speaker-matched cue recordings, provides explicit cue placement but may introduce unnatural cue–speech transitions. To examine this gap, we audit seven representative TTS models on 252 generated speech using their native control interfaces (see Figure 2). The results reveal two distinct failure modes: cue matching does not imply exact sequence compliance, while sequence compliance does not imply natural speech. Together,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

these findings expose a mismatch between general-purpose TTS capabilities and the requirements of symptom-aware patient speech.

Addressing this mismatch requires both structured supervision and a model capable of realizing symptom-related acoustic cues under explicit cue configurations. We therefore construct SymCue, a paired speech resource with temporal cue annotations, and **SymCueBench**, an evaluation benchmark covering seven cue types and single, repeated, and mixed cue configurations. Building on this supervision, we introduce **HealthTTS**, a symptom-aware TTS model based on CosyVoice3 that combines cue-weighted language-model and acoustic adaptation with preference-based refinement. At inference, HealthTTS directly generates complete utterances from text, a speaker reference, and an optional cue configuration, without external cue recordings or post-hoc cue insertion.

Experiments on SymCueBench show that HealthTTS achieves 87.33% Cue Sequence Accuracy under the shared-cue setting while maintaining high Speech Naturalness and Cue Realism. In a virtual standardized patient (VSP) study with 30 medical students across 14 scenarios, HealthTTS maintains high Symptom Cue Recognition while receiving higher Patient Realism and Perceived Usefulness ratings than Cue Insertion. These results suggest that symptom-related acoustic cues are better generated as an integral part of patient speech rather than added post hoc.

Our contributions are threefold:

- We formulate symptom-aware TTS as structured generation of symptom-related acoustic cues and identify limitations of existing TTS models in jointly achieving cue-sequence compliance and natural speech through a controlled audit.
- We construct **SymCue**, a paired speech resource with temporal cue annotations, and **SymCueBench**, an evaluation benchmark covering seven cue types and single, repeated, and mixed cue configurations.
- We introduce **HealthTTS**, a symptom-aware TTS model for directly generating complete utterances under structured cue control, and demonstrate improvements in cue-sequence compliance and natural cue–speech integration, with VSP simulation as a representative healthcare application.

2 Towards Symptom-Aware TTS for Healthcare Interaction

2.1 The Role of Symptom-Aware TTS in Healthcare Interaction

Symptom-aware Acoustic Cues. We refer to transient, non-lexical vocal manifestations of symptoms as symptom-aware acoustic cues. These cues complement verbal content by making symptoms directly perceptible in patient speech rather than only described in words. Unlike persistent pathological voice or speech characteristics (Halpern et al., 2026; Wang et al., 2026), symptom-related acoustic cues occur locally within an utterance and may recur or appear in combination, making them particularly relevant across speech-based healthcare settings, including clinical consultation (Ambady et al., 2002), telehealth interaction (Liu et al., 2020), and simulated patient encounters (Mukadam et al., 2025; Tyrrell et al., 2025). In this work, we consider seven representative cue types, including coughs, throat clearing, sighs, hiccups, retching, sneezing, and pain-related groans (Liao et al., 2025; Xue et al., 2026). Figure 1 illustrates representative healthcare interaction settings and the cue types considered in this work.

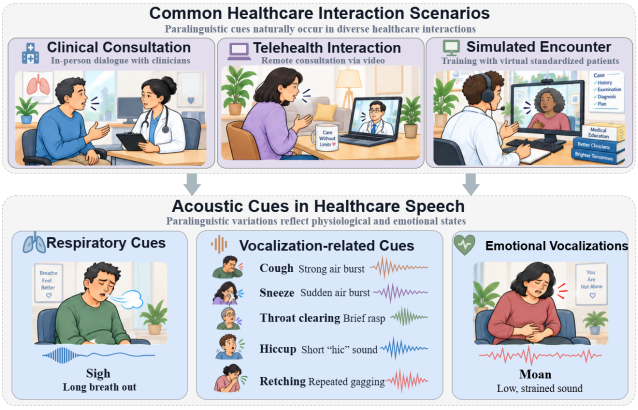


Figure 1: Seven symptom-related acoustic cue types considered in this work, adapted from Xue et al. (2026) and grouped into Respiratory, Throat/Physiological, and Emotional Vocalizations.

From Symptom Cues to Speech Generation. Many healthcare applications require patient speech to convey symptoms acoustically rather than through verbal descriptions alone. For example, in virtual standardized patient training, such cues can make patient presentations more realistic and expose learners to both verbal and acoustic symptom information. Symptom-aware TTS can also generate cue-conditioned speech for developing and evaluating speech-based healthcare systems. However, collecting real patient recordings is costly and difficult to scale, and existing recordings rarely cover different texts, speakers, symptoms, and cue combinations. Symptom-aware TTS can instead generate patient speech under controlled cue configurations. In our preliminary survey, 37 of 45 medical students (82.2%) considered voice necessary for VSP interaction (Appendix A).

2.2 Can Existing TTS Models Support Symptom-Aware Speech Generation?

In this setting, supporting symptom-aware speech generation requires both reliable compliance with structured cue requests and natural cue–speech integration.

Probing Existing TTS Models. We audit seven TTS models—GPT-4o-mini-TTS OpenAI (2025), Qwen3-TTS Hu et al. (2026), CosyVoice3 Du et al. (2025), Seed-TTS Anastassiou et al. (2024), Fish Speech S2 Pro Liao et al. (2026), Chatterbox-Turbo Resemble AI (2025), and Dia Nari Labs (2025)—on 36 requests per model involving cough, throat clearing, and sigh, balanced across single-cue, repeated-cue, and ordered mixed-cue configurations (252 utterances). Each request is generated once through the model's supported interface, without sample-specific prompt tuning or candidate selection; Further details are provided in Appendix B.

Qwen3-Omni Xu et al. (2025) serves as the primary evaluator, with human validation on a stratified subset. We evaluate Cue Match (whether the requested symptom cues are realized), Cue Count Accuracy (whether the number of cues is correct), Cue Sequence Accuracy (whether cue type, count, and order are correct), and Speech Naturalness (overall fluency and naturalness on a 1–5 scale).

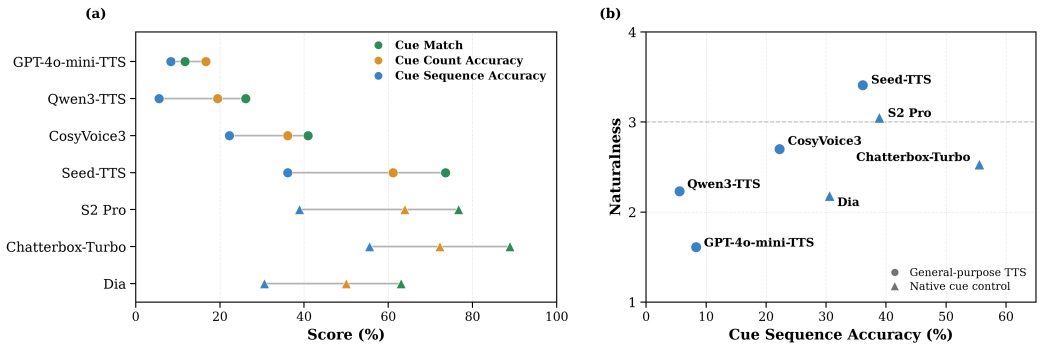


Figure 2: Pilot audit of seven TTS models on 252 utterances. Left: Cue Match, Cue Count Accuracy, and Cue Sequence Accuracy. Right: Speech Naturalness versus Cue Sequence Accuracy. Circles denote general-purpose TTS models, while triangles denote models with paralinguistic cue control.

Key Observation. Figure 2 reveals a consistent limitation across existing TTS models.

Observation: Existing TTS does not reliably combine cue control with natural speech. Models may fail to reproduce the requested cue sequence, while correct cue sequences can still result in unnatural speech.

Together, these results show that existing TTS models do not meet the needs of symptom-aware healthcare interaction. We therefore formulate symptom-aware TTS as speech generation conditioned on text, a speaker reference, and an optional cue configuration specifying the requested cue sequence and lexical boundaries.

Task Requirement. Reliably reproduce the requested cue sequence while generating natural cue-conditioned speech and preserving linguistic content.

3 HEALTHTTS

Building on the task formulation above, HealthTTS directly generates complete utterances from text, a speaker reference, and an optional cue configuration. We train HealthTTS in three stages: Stage I performs cue-aware language-model adaptation to learn how requested cues are represented and generated in the speech-token sequence (Section 3.1); Stage II applies localized acoustic adaptation to refine the acoustic realization and integration of cue regions (Section 3.2); and Stage III performs global preference refinement over complete utterances to improve cue-sequence correctness and overall speech naturalness (Section 3.3).

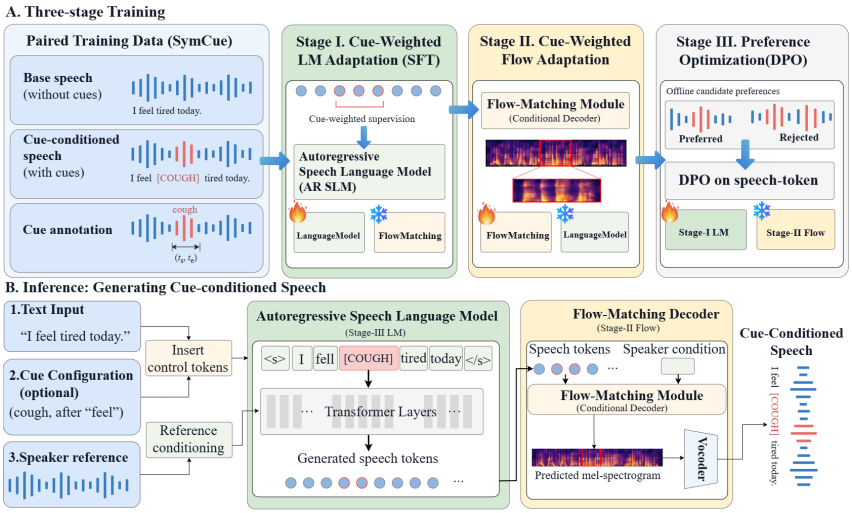


Figure 3: HealthTTS overview. (a) Speech generation from text, a speaker reference, and an optional cue configuration. (b) Sequential training with cue-weighted language-model adaptation, cue-weighted flow adaptation, and direct preference optimization (DPO).

3.1 STAGE-I: CUE-AWARE LM ADAPTATION

Stage I enables the language model to learn how requested symptom-related cues are represented and generated in the speech-token sequence. Given input text x and a speaker reference s, we represent the requested cues using a cue configuration

$$C = \big((c_1, l_1), \ldots, (c_K, l_K)\big), \tag{1}$$

where c_k denotes the cue type and l_k its requested lexical boundary. We insert one non-lexical control token for each cue to obtain the cue-conditioned input $\widetilde{x}$. The token identity encodes cue type, while its number, order, and placement encode the requested cue sequence and lexical boundaries.

Training uses paired base and cue-conditioned speech under the same text and speaker reference. Base speech corresponds to an empty cue configuration, while cue-conditioned speech follows the assigned configuration. Temporal annotations identify the cue-aligned speech tokens used for localized supervision.

Cue-weighted LM adaptation. To emphasize learning of cue-related speech tokens, we update only the autoregressive speech language model while keeping the flow decoder fixed. Let $z = (z_1, \ldots, z_T)$ denote the target speech-token sequence, and let m_j indicate whether token z_j overlaps an annotated cue region:

$$\mathcal{L}_{\mathrm{LM}} = -\frac{1}{\sum_{j=1}^{T} w_j} \sum_{j=1}^{T} w_j \log p_{\theta_{\mathrm{LM}}}\left(z_j \mid z_{<j}, \widetilde{x}, s\right), \qquad w_j = 1 + \alpha m_j, \tag{2}$$

where $\alpha \geq 0$ assigns greater weight to cue-aligned speech tokens.

3.2 Stage-II: Localized Acoustic Adaptation

Stage II refines the acoustic realization and integration of the learned symptom-related cues by emphasizing cue-aligned acoustic regions.

Cue-weighted flow adaptation. We freeze the adapted language model and update only the flow decoder. Let M_n indicate whether acoustic frame n overlaps an annotated cue region. To emphasize learning within these regions, we weight the backbone's frame-level flow-matching loss (Lipman et al., 2022) as

$$\mathcal{L}_{\text{flow}} = \mathbb{E}\left[\frac{1}{\sum_{n=1}^{N} v_n}\sum_{n=1}^{N} v_n\,\ell_{\text{flow}}(n)\right], \qquad v_n = 1 + \beta M_n, \tag{3}$$

where $\beta \geq 0$ assigns greater weight to cue-aligned acoustic frames.

Stages I and II are optimized sequentially: Stage I learns cue-aware speech-token generation, while Stage II refines the local acoustic realization of the requested cues. For an empty cue configuration, both weighted objectives reduce to their standard forms.

3.3 Stage-III: Global Preference Refinement

Stage III refines complete utterances using preferences that jointly consider cue-sequence correctness and natural cue–speech integration. While the localized objectives in Stages I and II emphasize cue-related regions, they do not directly compare the overall quality of different generations under the same cue configuration.

We generate multiple candidates for each cue-conditioned input while keeping the text, speaker reference, cue configuration, and decoding settings fixed. Preference pairs first favor candidates that correctly reproduce the requested cue sequence. When both candidates satisfy the requested sequence, preference is given to the one with more realistic cues and more natural cue–speech integration. Ambiguous comparisons are discarded.

DPO-based LM refinement. Let z^+ and z^- denote the speech-token sequences of the preferred and dispreferred candidates. We optimize the language model using Direct Preference Optimization (DPO) (Rafailov et al., 2023):

$$\mathcal{L}_{\text{DPO}} = -\log\sigma\left(\gamma\left[\log\frac{\pi_\theta(z^+ \mid \widetilde{x}, s)}{\pi_{\text{ref}}(z^+ \mid \widetilde{x}, s)} - \log\frac{\pi_\theta(z^- \mid \widetilde{x}, s)}{\pi_{\text{ref}}(z^- \mid \widetilde{x}, s)}\right]\right), \tag{4}$$

where π_{ref} is the frozen Stage-I language model and $\gamma > 0$ controls the preference strength. Only the language model is updated, while the Stage-II flow decoder remains fixed.

4 Data Engineering: Constructing Cue-Aware Supervision for Symptom-Aware TTS

Training HealthTTS requires paired supervision that shows how requested symptom-related cues alter speech under matched linguistic and speaker conditions, together with controlled cue configurations and temporal cue annotations. We therefore construct SymCue for training and SymCueBench for evaluation. The data pipeline consists of three steps: collecting speaker-matched speech and cue recordings (Section 4.1), constructing paired base and cue-conditioned speech under controlled cue configurations (Section 4.2), and annotating the realized cue regions while maintaining leakage control between training and evaluation (Section 4.3). Figure 4 summarizes the overall pipeline.

4.1 Data Source

Each unit contains a base text, speaker reference, cue configuration, base speech, cue-conditioned speech, and temporal cue annotations. The base speech provides the target for an empty cue configuration, while the cue-conditioned speech follows the assigned non-empty cue configuration under the same linguistic and speaker conditions.

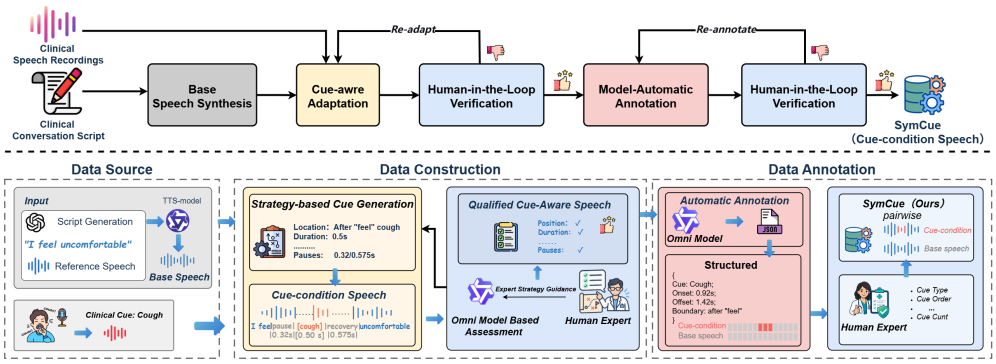


Figure 4: Construction of paired supervision for symptom-related acoustic cues. Base speech and cue-conditioned speech share the same base text and speaker reference, while the cue-conditioned speech follows an assigned cue configuration with temporal cue annotations for localized supervision.

Recording sources. We recruit six speakers, including three female and three male speakers, under consistent recording conditions. Each speaker records ten reference utterances used as speaker references for base-speech synthesis. The same speaker pool provides recordings of seven symptom-related acoustic cue types, with each type recorded by two speakers. Cue-conditioned speech is constructed using speaker-matched cue recordings; a multi-cue configuration is assigned to a speaker only when recordings are available for all requested cue types.

Cue configurations. Cue configurations contain one, two, or three cues, covering same-type repetitions and mixed-type sequences. For mixed-type configurations, matched groups include distinct cue orders while sharing the same base text, speaker reference, and requested lexical boundaries. This design varies cue order while holding the linguistic and positional context fixed. Cue positions cover sentence-initial, sentence-medial, and sentence-final regions and are represented by requested lexical boundaries, with multiple cues assigned to distinct boundaries. Figure 5 summarizes the corpus.

4.2 Data Construction

Speech construction. We synthesize base speech using Qwen3-TTS, conditioned on the base text and speaker reference. Requested lexical boundaries determine where speaker-matched cue recordings are integrated with the synthesized base speech using cue-specific composition strategies to construct the corresponding cue-conditioned speech. These strategies control cue duration, amplitude, local pauses, and cue–speech transitions. Appendix C.1 provides the detailed composition procedures.

Human-in-the-Loop Refinement. All constructed cue-conditioned speech is automatically verified by Qwen3-Omni for linguistic consistency, cue-sequence compliance, and natural cue–speech integration. Failed samples are reconstructed using revised cue composition strategies and verified again.

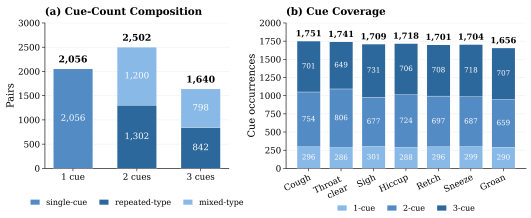


Figure 5: Overview of SymCue and SymCueBench. (a) Distribution by cue count and composition. (b) Per-cue occurrence counts across cue configurations.

As an additional quality check, two annotators independently inspect a stratified subset of 350 cue-conditioned utterances, with 50 samples per cue type and coverage across cue counts and repeated versus mixed-type configurations. The audit assesses cue recognizability, placement appropriateness, duration and pause patterns, and natural cue–speech integration. Issues identified during the audit are used to refine the corresponding cue composition strategies. A total of 1,128 samples require a second construction round, and 270 require a third. The 54 remaining failures are discarded and replaced within the same cue-configuration stratum and data partition.

4.3 Data Annotation

Temporal annotation and audit. Only verified cue-conditioned speech proceeds to temporal annotation. Qwen3-Omni automatically identifies the realized cue regions and associates each onset–offset interval with its cue type and requested lexical boundary. The intervals are mapped to speech-token and acoustic-frame resolutions for localized supervision, with zero masks assigned to base speech. To assess annotation quality, two annotators independently inspect a stratified subset of 350 utterances, with 50 samples per cue type and coverage across cue counts and repeated versus mixed-type configurations. The audit verifies cue type, count, and order together with the corresponding temporal intervals and lexical-boundary associations, and incorrect or incomplete annotations are corrected through re-annotation. Detailed criteria are provided in Appendix C.3.

Dataset Partitioning. Following the above construction, verification, and annotation pipeline, we obtain 4,958 training pairs in SymCue and 1,240 pairs in SymCueBench. To prevent leakage between training and evaluation, we define disjoint resource pools before speech construction, separating base texts, cue recordings, and speaker-reference recordings while allowing the same speaker identities across the two partitions. Each paired unit and all cue-order variants derived from the same base text are kept together.

5 Experiments

We evaluate HealthTTS in terms of cue controllability, speech quality, and healthcare applicability. We first assess whether the model can faithfully generate the requested symptom-related acoustic cues while maintaining natural and intelligible speech, then examine the contribution of each training component, and finally evaluate its effectiveness in VSP simulation.

5.1 Experimental Setup

Evaluation settings. HealthTTS is trained on SymCue and evaluated on SymCueBench in two settings. The *full-cue setting* evaluates CosyVoice3, Cue Insertion, and HealthTTS on 1,240 requests covering all seven cue types, while the *shared-cue setting* evaluates Fish Speech S2 Pro, Chatterbox-Turbo, Dia, and HealthTTS on 292 requests covering cough, throat clearing, and sigh. Methods within each setting use matched texts, speaker references, and cue configurations. The settings are reported separately because their cue coverage and cue-count distributions differ. Each request is generated with three independent seeds, with only technical failures retried; detailed request distributions and generation settings are provided in Appendix D.1.

Metrics and evaluation. We use the cue-control metrics defined in Section 2.2, together with WER and Speaker Similarity for speech preservation and Speech Naturalness and Cue Realism for perceptual quality. Cue Realism is scored only for successfully realized cues, and lexical-boundary accuracy is not evaluated separately. Qwen3-Omni serves as the primary evaluator; on 140 human-annotated utterances, it achieves 86.8% exact cue-sequence agreement and correlations of $r = 0.77$–0.81 across perceptual metrics (Appendix D.3).

5.2 Cue Controllability

HealthTTS obtains the highest Cue Match, Cue Count Accuracy, and Cue Sequence Accuracy in both settings (Table 1). In the full-cue setting, Cue Sequence Accuracy increases from 28.31% for CosyVoice3 to 79.41% for HealthTTS. In the shared-cue setting, HealthTTS achieves 87.33% Cue Sequence Accuracy, exceeding Fish Speech S2 Pro by 25.12 percentage points.

Cue realism alone does not imply structured cue control. HealthTTS and Fish Speech S2 Pro achieve similar Cue Realism (4.85 vs. 4.80), but substantially different Cue Sequence Accuracy (87.33% vs. 62.21%; Figure 6(a)). Thus, realistic individual cues do not guarantee faithful reproduction of the requested cue sequence.

Takeaway-I. HealthTTS substantially improves structured cue control while retaining high Cue Realism for successfully realized cues.

Table 1: Performance on SymCueBench under the full-cue and shared-cue settings. Cue Insertion is a post-hoc insertion reference. Bold indicates the best result in each setting.

Method	Cue Controllability			Speech Preservation		Perceptual Quality	
	Cue Match ↑	Cue Count Acc. ↑	Cue Sequence Acc. ↑	WER ↓	Speaker Sim. ↑	Speech Naturalness ↑	Cue Realism ↑
Full-cue setting							
CosyVoice3	33.64	47.18	28.31	7.93	**0.88**	2.95	–
Cue Insertion	–	–	–	10.61	0.86	1.55	4.62
HealthTTS	**83.87**	**80.11**	**79.41**	**7.33**	0.87	**4.13**	**4.84**
Shared-cue setting							
Fish Speech S2 Pro	76.58	81.39	62.21	**3.42**	0.85	3.10	4.80
Chatterbox-Turbo	85.73	66.32	59.59	10.97	0.75	3.13	2.77
Dia	57.41	49.77	25.11	14.36	0.69	2.43	2.68
HealthTTS	**91.03**	**88.58**	**87.33**	7.80	**0.88**	**4.25**	**4.85**

5.3 Speech Quality

In the full-cue setting, HealthTTS largely preserves linguistic content and speaker identity while improving perceptual quality. WER decreases from 7.93% to 7.33%, Speaker Similarity remains close to CosyVoice3 (0.87 vs. 0.88), and Speech Naturalness increases from 2.95 to 4.13. Cue Insertion achieves high Cue Realism (4.62) but low Speech Naturalness (1.55), showing that realistic cues alone do not ensure natural cue–speech integration.

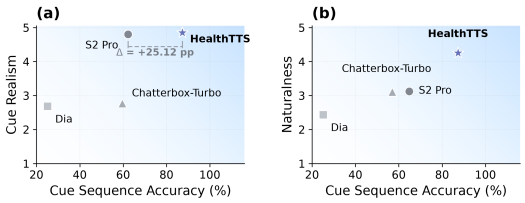


Figure 6: Cue controllability and perceptual quality on the shared-cue set. (a) Cue Realism versus Cue Sequence Accuracy; (b) Speech Naturalness versus Cue Sequence Accuracy. Cue Realism is evaluated only for successfully realized cues; the dashed annotation in (a) indicates the difference in Cue Sequence Accuracy.

In the shared-cue setting, HealthTTS achieves the highest Speech Naturalness (4.25) and Speaker Similarity (0.88). Fish Speech S2 Pro obtains lower WER (3.42% vs. 7.80%), indicating that linguistic content preservation and perceptual naturalness capture complementary aspects of speech quality.

Takeaway-II. HealthTTS combines strong structured cue control with high Speech Naturalness while largely preserving linguistic content and speaker identity.

5.4 Ablation Study

We ablate LM cue weighting ($\alpha = 0$), flow cue weighting ($\beta = 0$), and preference-based refinement (DPO) while keeping training and evaluation settings unchanged.

The three components play complementary roles. Removing LM cue weighting causes the largest loss in structured cue control, reducing Cue Sequence Accuracy from 79.41% to 63.28%. Removing flow cue weighting affects sequence control less but causes the largest degradation in Speech Naturalness (4.13 to 3.82) and Cue Realism (4.84 to 4.12). Removing preference-based refinement reduces Cue Sequence Accuracy to 73.55% and Cue Realism to 4.55, while WER and Speaker Similarity remain largely stable across variants.

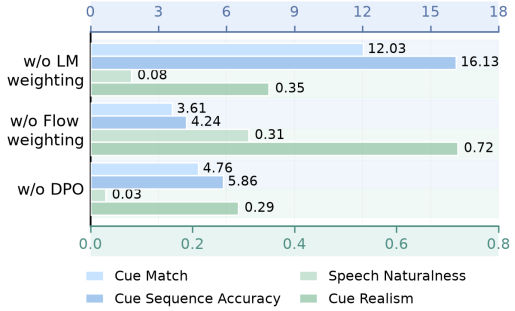


Figure 7: Ablation drops on the full-cue benchmark. Blue and green bars correspond to cue controllability and perceptual quality, using the top and bottom axes, respectively.

Takeaway-III. LM cue weighting mainly strengthens structured cue control, flow cue weighting improves perceptual quality, and preference-based refinement benefits both.

5.5 Virtual Standardized Patient Evaluation

Benchmark metrics capture cue controllability and speech quality, but not healthcare relevance. We therefore evaluate HealthTTS in VSP simulation on cue recognizability, patient realism, and perceived usefulness. Thirty Year 3–5 medical students evaluate 14 VSP scenarios across seven cue categories and three conditions: CosyVoice3 base speech, Cue Insertion, and HealthTTS. Linguistic content and speaker identity match across conditions. With counterbalancing, participants hear one condition per scenario, yielding 140 judgments per condition. Participants identify cues and rate Patient Realism and Perceived Usefulness on 1–5 scales. See Appendix D.5.

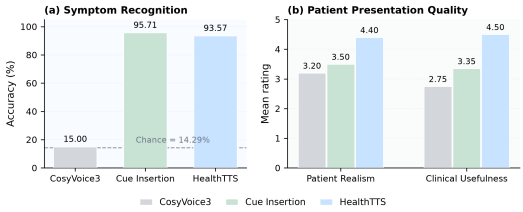


Figure 8: Virtual standardized patient evaluation with 30 medical students across 14 scenarios (140 judgments per condition). (a) Symptom Cue Recognition, with the dashed line indicating the 14.29% chance level. (b) Mean Patient Realism and Perceived Usefulness.

HealthTTS and Cue Insertion achieve 93.57% and 95.71% Symptom Cue Recognition, respectively, compared with 15.00% for CosyVoice3 and a 14.29% chance level. Despite similar cue recognizability, HealthTTS achieves higher Patient Realism (4.40 vs. 3.50) and Perceived Usefulness (4.50 vs. 3.35), suggesting that natural cue–speech integration also matters for convincing patient presentation.

Takeaway-IV. HealthTTS maintains high cue recognizability while achieving higher Patient Realism and Perceived Usefulness than post-hoc Cue Insertion.

6 Related Work

Controllable Speech Generation. CosyVoice3 and Qwen3-TTS improve prosody and controllable delivery (Du et al., 2025; Hu et al., 2026). Fish Speech S2 Pro enables localized vocal control via inline instructions (Liao et al., 2026), while Chatterbox-Turbo and Dia support non-lexical vocalizations (Resemble AI, 2025; Nari Labs, 2025). NVSpeech provides word-level paralinguistic supervision for recognition and synthesis (Liao et al., 2025), and NVBench evaluates cue controllability, placement, salience, and speech quality (Xue et al., 2026). HealthTTS targets structured control over cue type, count, and order with natural cue–speech integration.

Speech in Healthcare Interaction. Clinical speech conveys affective and interpersonal information (Ambady et al., 2002). Voice-based virtual patients support communication practice but face limitations in emotional authenticity (Mukadam et al., 2025), while EasyMED explores case-grounded patient dialogue (Zhang et al., 2026). Pathological TTS controls persistent characteristics such as severity and dysarthric articulation (Halpern et al., 2026; Wang et al., 2026). HealthTTS models transient symptom-related acoustic cues within continuous speech and evaluates their utility in virtual standardized patient simulation.

7 Conclusion

We presented **HealthTTS**, a symptom-aware TTS framework for structured generation of symptom-related acoustic cues, together with SymCue and **SymCueBench** for paired supervision and evaluation. HealthTTS combines cue-weighted language-model and acoustic adaptation with preference-based refinement, substantially improving cue-sequence compliance and natural cue–speech integration while largely preserving linguistic content, speaker identity, and speech quality. In virtual standardized patient simulation, HealthTTS maintains high symptom cue recognizability while achieving higher Patient Realism and Perceived Usefulness than post-hoc Cue Insertion, supporting the integration of symptom-related acoustic cues directly into generated patient speech.

LIMITATIONS

SymCue uses synthesized base speech and speaker-matched cue recordings to provide controlled supervision, but does not fully capture the variability and cue–speech transitions of naturally occurring patient speech. SymCueBench shares the same six speaker identities as SymCue, so generalization to unseen speakers remains untested. In addition, our controllability metrics evaluate cue type, count, and order but do not separately measure the accuracy of requested lexical boundaries.

Evaluation relies primarily on Qwen3-Omni, with independent human validation on stratified subsets. The VSP study includes 30 medical students and 14 scenarios and evaluates Patient Realism and Perceived Usefulness rather than learning outcomes or clinical effectiveness. Broader evaluation on natural patient speech, unseen speakers, and interactive settings is therefore needed.

ETHICAL STATEMENT

This work follows the ICLR Code of Ethics. HealthTTS is intended for research and educational use in controlled healthcare simulation, rather than for diagnosis or clinical decision-making. Human participants involved in speech collection and evaluation participated with informed consent. resources will follow applicable consent, privacy, and licensing requirements.

AI USE STATEMENT

Generative AI tools were used to assist with language editing and improving the readability of the manuscript. Generative AI models were also used as part of the research methodology: Qwen3-TTS was used to synthesize base speech for data construction, and Qwen3-Omni was used for automatic data verification, temporal cue annotation, preference-pair construction, and model evaluation. Human auditing and validation were conducted as described in the corresponding sections of the paper. All uses of generative AI were conducted under author supervision, with human auditing and validation as described in the corresponding sections of the paper. The authors designed the methodology and take full responsibility for the analysis and final manuscript.

REFERENCES

Nalini Ambady, Debi Laplante, Thai Nguyen, Robert Rosenthal, Nigel Chaumeton, and Wendy Levinson. Surgeons' tone of voice: A clue to malpractice history. *Surgery*, 132(1):5–9, 2002. doi: 10.1067/msy.2002.124733.

Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, et al. Seed-tts: A family of high-quality versatile speech generation models. *arXiv preprint arXiv:2406.02430*, 2024.

Zhihao Du, Changfeng Gao, Yuxuan Wang, Fan Yu, Tianyu Zhao, Hao Wang, Xiang Lv, Hui Wang, Chongjia Ni, Xian Shi, et al. Cosyvoice 3: Towards in-the-wild speech generation via scaling-up and post-training. *arXiv preprint arXiv:2505.17589*, 2025.

Bence Mark Halpern, Wen-Chin Huang, Lester Phillip Violeta, and Tomoki Toda. Severity-controllable pathological text-to-speech synthesis for clinical applications. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 34:573–582, 2026.

Hangrui Hu, Xinfa Zhu, Ting He, Dake Guo, Bin Zhang, Xiong Wang, Zhifang Guo, Ziyue Jiang, Hongkun Hao, Zishan Guo, et al. Qwen3-tts technical report. *arXiv preprint arXiv:2601.15621*, 2026.

Huan Liao, Qinke Ni, Yuancheng Wang, Yiheng Lu, Haoyue Zhan, Pengyuan Xie, Qiang Zhang, and Zhizheng Wu. Nvspeech: An integrated and scalable pipeline for human-like speech modeling with paralinguistic vocalizations. *arXiv preprint arXiv:2508.04195*, 2025.

Shijia Liao, Yuxuan Wang, Songting Liu, Yifan Cheng, Ruoyi Zhang, Tianyu Li, Shidong Li, Yisheng Zheng, Xingwei Liu, Qingzheng Wang, et al. Fish audio s2 technical report. *arXiv preprint arXiv:2603.08823*, 2026.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Shan Liu, Muyu Zhang, Baojun Gao, and Guoyin Jiang. Physician voice characteristics and patient satisfaction in online health consultation. *Information & Management*, 57(5):103233, 2020. doi: 10.1016/j.im.2019.103233.

Alifya Mukadam, Shruti Suresh, and Chris Jacobs. Beyond traditional simulation: An exploratory study on the effectiveness and acceptability of ChatGPT-4o advanced voice mode for communication skills practice among medical students. *Cureus*, 17(5):e84381, 2025. doi: 10.7759/cureus.84381.

Nari Labs. Dia: A TTS model capable of generating ultra-realistic dialogue in one pass. `https://github.com/nari-labs/dia`, 2025. GitHub repository.

OpenAI. Introducing next-generation audio models in the API. `https://openai.com/index/introducing-our-next-generation-audio-models/`, March 2025. Accessed: 2026-09-14.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pp. 28492–28518. PMLR, 2023.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Resemble AI. Chatterbox-TTS. `https://github.com/resemble-ai/chatterbox`, 2025. GitHub repository.

Edward G. Tyrrell, Sardip K. Sandhu, Kathryn Berry, Suzan F. Ghannam, Sarah A. Lewis, Daniel Crowfoot, Gurvinder Singh Sahota, Julie Carson, Emma E. Wilson, and Jaspal Taggar. Web-based AI-driven virtual patient simulator versus actor-based simulation for teaching consultation skills: Multicenter randomized crossover study. *JMIR Formative Research*, 9:e71667, 2025. doi: 10.2196/71667.

Haoshen Wang, Xueli Zhong, Bingbing Lin, Jia Huang, Xingduo Pan, Shengxiang Liang, Nizhuan Wang, and Wai Ting Siok. Disentanglement matters: Controllable dysarthric speech synthesis via prototype-based representation learning. *arXiv preprint arXiv:2602.08696*, 2026.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025.

Liumeng Xue, Weizhen Bian, Jiahao Pan, Wenxuan Wu, Yilin Ren, Boyi Kang, Jingbin Hu, Ziyang Ma, Shuai Wang, Xinyuan Qian, Hung-yi Lee, and Yike Guo. NVV-SuperBench: Beyond words, beyond quality—benchmarking nonverbal vocalizations in speech generation. *arXiv preprint arXiv:2604.16211*, 2026. URL `https://arxiv.org/abs/2604.16211v3`. Version 3.

Bingquan Zhang, Xiaoxiao Liu, Yuchi Wang, Zhou Lei, Qianqian Xie, and Benyou Wang. Human or llm as standardized patients? a comparative study in medical education. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 26988–27012, 2026.

APPENDIX CONTENTS

A Preliminary Survey on Voice in VSP . . . 13

B Pilot Study Details . . . 13

B.1 Request Protocol . . . 13
B.2 Models and Control Interfaces . . . 13
B.3 Evaluation Protocol . . . 14

C SymCue Construction and Verification . . . 14

C.1 Cue Composition . . . 14
C.2 Construction Verification . . . 16
C.3 Temporal Annotation and Verification . . . 17
C.4 Leakage-Controlled Partitioning . . . 17

D Additional Experimental Details . . . 18

D.1 Training and Generation Details . . . 18
D.2 Evaluation Metrics . . . 19
D.3 Human Validation of Qwen3-Omni Evaluation . . . 21
D.4 Ablation Training Protocol . . . 21
D.5 Virtual Standardized Patient Evaluation . . . 22

Table 2: Responses to the preliminary survey on voice in virtual standardized patient interaction, by year of study (45 students; 15 per year). Values are respondent counts.

Year	Needed	Either	Not Needed
Year 3	10	5	0
Year 4	13	1	1
Year 5	14	1	0
Total	37	7	1

Table 3: Pilot cue configurations, with 36 requests per model across three request types. Cue types are cough, throat clearing, and sigh; c_1 and c_2 denote cue types.

Request type	Requests / Model	Configuration
Single cue	12	c_1
Repeated cue	12	$c_1 \rightarrow c_1$
Mixed cue	12	$c_1 \rightarrow c_2,\ c_1 \neq c_2$
Total	36	–

A Preliminary Survey on Voice in VSP

To examine whether voice is perceived as an important component of virtual standardized patient interaction, we conducted a preliminary survey with 45 medical students from Years 3–5, with 15 participants from each year. Participants were asked whether voice was needed when interacting with a VSP and selected one of three responses: *Needed*, *Either is acceptable*, or *Not needed*. The survey was used only to motivate the role of speech in VSP interaction and was conducted separately from the VSP evaluation in Section 5.5.

Overall, 37 of 45 participants (82.2%) considered voice necessary, seven (15.6%) considered either voice or text-only interaction acceptable, and one (2.2%) considered voice unnecessary.

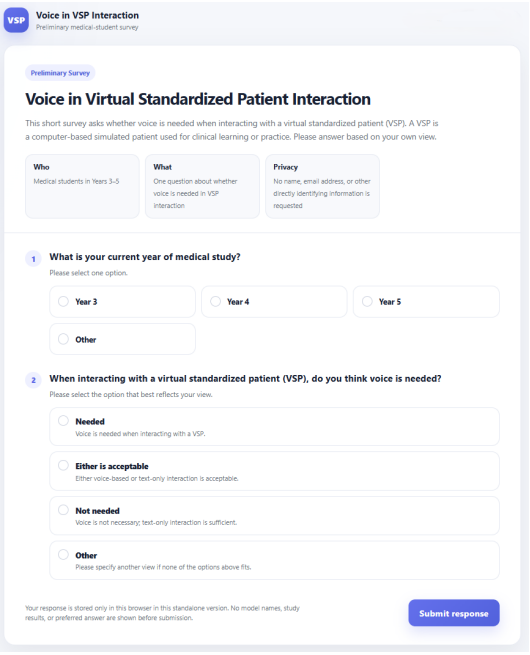


Figure 9: Questionnaire for the preliminary survey on the need for voice in virtual standardized patient interaction.

B Pilot Study Details

B.1 Request Protocol

The pilot study examines whether existing TTS models can reliably follow structured requests for symptom-related acoustic cues. We consider three representative cues—*cough*, *throat clearing*, and *sigh*—and construct 36 requests for each model. The requests are equally divided among single-cue, repeated-cue, and mixed-cue configurations, as summarized in Table 3. The same base texts and underlying cue configurations are used across models, while each request is expressed through the native or documented control interface of the corresponding model. This results in 252 generated utterances across seven TTS models.

B.2 Models and Control Interfaces

We evaluate GPT-4o-mini-TTS, Qwen3-TTS, CosyVoice3, Seed-TTS, Fish Speech S2 Pro, Chatterbox-Turbo, and Dia. All models receive the

Table 4: Control interfaces and generation settings for the seven TTS models in the pilot audit. Identical underlying cue configurations are expressed through each model's supported control representation. Settings are fixed within each model throughout evaluation.

Model	Control Type	Control Representation	Temp.	Top-p
GPT-4o-mini-TTS	Instruction-based	Natural-language instruction	Default	–
Qwen3-TTS	Instruction-based	Natural-language instruction	0.8	0.95
CosyVoice3	Instruction-based	Natural-language instruction	0.8	0.95
Seed-TTS	Instruction-based	Natural-language instruction	Default	Default
Fish Speech S2 Pro	Explicit control	Native paralinguistic markers	0.8	0.95
Chatterbox-Turbo	Explicit control	Paralinguistic tags	Default	Default
Dia	Explicit control	Inline nonverbal tags	0.8	0.95

"–" indicates that the corresponding parameter is not exposed by the model interface. "Default" indicates the recommended setting of the released implementation.

same base texts and underlying cue configurations, while each configuration is translated into the native or documented control representation supported by the corresponding model.

GPT-4o-mini-TTS, Qwen3-TTS, CosyVoice3, and Seed-TTS use instruction-based control. Fish Speech S2 Pro, Chatterbox-Turbo, and Dia use their supported explicit paralinguistic representations. For repeated- and mixed-cue configurations, the requested cue types, counts, order, and lexical boundaries are specified whenever supported by the corresponding interface.

The control format and decoding configuration are fixed before evaluation. Each request is generated once, without sample-specific prompt tuning, repeated generation, or best-of-N selection.

Control templates. To ensure a consistent comparison, all systems receive the same underlying base text, cue types, cue counts, cue order, and requested lexical boundaries. Only the surface form of the control request is adapted to the interface supported by each model. For instruction-based systems, the instruction explicitly specifies the requested non-lexical vocalization and its position relative to the surrounding text. For systems with native paralinguistic control, the corresponding model-specific cue markers are inserted at the requested lexical boundaries.

Sampling and control parameters are fixed before evaluation and remain unchanged across requests within each model. We use the recommended generation configuration associated with each released system when a model does not expose an equivalent sampling parameter. No request-specific prompt tuning, repeated generation, or best-of-N selection is performed.

B.3 Evaluation Protocol

Qwen3-Omni is used as the primary evaluator for the pilot study. For each generated utterance, it identifies the realized cue sequence without access to the requested cue configuration or model identity and rates Speech Naturalness on a 1–5 scale using a fixed rubric.

Cue controllability is evaluated using the same definitions as in the main experiments. *Cue Match* reports macro-F1 over the three pilot cue categories, with repeated occurrences counted separately and cue order ignored. *Cue Count Accuracy* measures whether the total number of detected cues equals the requested number, regardless of cue type or order. *Cue Sequence Accuracy* requires exact agreement between the requested and detected cue sequences in cue type, count, and order. Speech Naturalness evaluates the overall fluency and perceptual naturalness of the complete utterance.

The reliability of Qwen3-Omni is assessed through independent human evaluation on a stratified subset of the pilot outputs. Human validation for both the pilot study and the main experiments is reported jointly in Appendix D.3.

C SymCue Construction and Verification

C.1 Cue Composition

Each cue-conditioned waveform is constructed from the corresponding base waveform and speaker-matched cue recordings according to its assigned cue configuration. These cue recordings come from the six-speaker pool described in Section 4.1. A multi-cue configuration is assigned to a speaker only when recordings are available for all requested cue types. Waveform composition is used only for data construction; HealthTTS directly generates complete utterances at inference time without external cue recordings or post-hoc cue insertion.

Table 5 summarizes cue-recording availability. Multi-cue construction requires all requested cue types to be available from the assigned speaker.

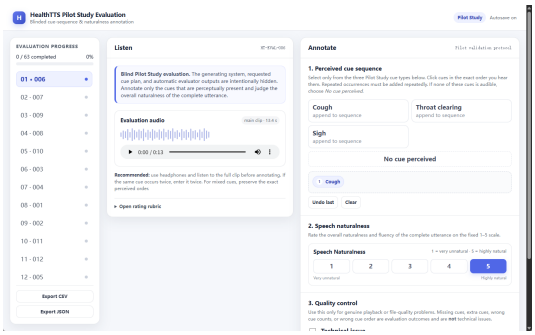


Figure 10: Human annotation interface for the pilot audit. Annotators record perceived cue sequences and rate Speech Naturalness (1–5), with system identity and requested cue configurations hidden.

Shared composition procedure. We use a stop–pause–cue–recovery–resume procedure. At each insertion point, the base waveform is faded out over a short edge window, followed by a pre-cue pause, a cropped cue recording, and a post-cue recovery pause. The remaining base waveform is then faded back in.

Inserted pauses contain a low-level room-tone residual with an RMS of approximately 1.0–1.2% of the base-speech RMS. Cue RMS is matched to the local base speech and scaled by a construction-time gain. Peak amplitude is limited to 0.95 during composition. Completed cue-conditioned waveforms are subsequently high-pass filtered at 80 Hz, low-pass filtered at 12 kHz, and loudness-normalized.

Table 5: Speaker-matched cue recordings available for each speaker. Checkmarks indicate availability; dashes indicate no recordings. C: cough; TC: throat clearing; Si: sigh; G: pain-related groan; Sn: sneeze; R: retch; H: hiccup.

Speaker	C	TC	Si	G	Sn	R	H	Types
A	✓	✓	–	–	–	–	–	2
B	✓	✓	–	–	–	–	–	2
C	–	–	✓	✓	–	✓	–	3
D	–	–	✓	–	–	✓	–	2
E	–	–	–	–	✓	–	✓	2
F	–	–	–	✓	✓	–	✓	3
Per cue	2	2	2	2	2	2	2	–

Requested cue positions and placement. Cue configurations specify an early (*head*), middle (*mid*), or late (*tail*) region of the base text. These regions correspond to character-position intervals of 10–30%, 35–65%, and 70–90%, respectively. Candidate lexical boundaries are selected at word or punctuation boundaries within the assigned region. These requested lexical boundaries determine the insertion points in the synthesized base speech, after which speaker-matched cue recordings are integrated using cue-specific composition strategies.

Insertion points are kept away from waveform boundaries by a margin of $\min(0.35\,\mathrm{s}, 0.05T)$, where T is the base-speech duration. Matched order variants share the base text, speaker reference, and requested lexical boundaries while varying the assignment and order of cue types.

Category-specific acoustic settings. Table 6 specifies the crop-duration ranges, pause lengths, edge fades, and minimum-gap settings for each cue category. Crop duration is capped by the available cue recording. Short cues such as hiccups use relatively brief interruptions, whereas sighs and retches permit longer cue and recovery intervals. Minimum-gap settings constrain spacing between adjacent insertions in both repeated-cue and mixed-type configurations.

Construction-time acoustic variation. We vary cue gain and pause length using three levels: light, medium, and heavy. For both pre- and post-cue pauses, these levels select the lower bound, midpoint, and upper bound of the corresponding range in Table 6. The associated gain factors are listed in Table 7.

Table 7: Construction-time cue gains. Values multiply the cue RMS after matching to the local base-speech RMS.

Cue	Light	Medium	Heavy
Cough	0.90	1.15	1.40
Throat clearing	0.85	1.10	1.30
Sigh	0.85	1.10	1.30
Pain-related groan	0.90	1.15	1.35
Sneeze	0.95	1.20	1.45
Retch	0.90	1.15	1.40
Hiccup	0.90	1.15	1.35

Table 6: Cue-specific timing and transition settings for data construction. Crop durations cannot exceed the available cue recording; pause lengths vary with the construction level. Room tone is a fraction of the base-speech root-mean-square (RMS) amplitude.

Cue	Crop (s)	Min. gap (s)	Pause before (s)	Pause after (s)	Edge fade (ms)	Room tone
Cough	0.28–0.90	2.0	0.22–0.42	0.40–0.75	55	0.012
Throat clearing	0.20–0.70	1.8	0.48–0.65	0.30–0.55	50	0.012
Sigh	0.40–1.40	2.4	0.28–0.50	0.45–0.85	65	0.012
Pain-related groan	0.30–1.00	2.2	0.25–0.45	0.40–0.70	60	0.012
Sneeze	0.35–1.00	2.2	0.65–0.80	0.45–0.80	45	0.012
Retch	0.35–1.20	2.6	0.28–0.50	0.50–0.90	60	0.012
Hiccup	0.12–0.45	1.4	0.10–0.22	0.25–0.45	40	0.010

These levels introduce bounded acoustic variation during data construction. They are not included in the cue configuration provided to HealthTTS and are not evaluated as a separate control dimension.

C.2 Construction Verification

Automatic verification. All constructed cue-conditioned waveforms undergo automatic verification for linguistic consistency, cue-sequence compliance, and signal quality. Qwen3-Omni identifies the perceptually realized cue sequence, which is compared with the requested cue sequence to detect missing, substituted, duplicated, or incorrectly ordered cues. Transcript- and signal-level checks additionally identify clear linguistic errors and invalid waveforms.

Automatic cue-sequence verification assesses cue type, count, and order. Requested lexical boundaries are retained as construction conditions but are not used as an automatic acceptance criterion, consistent with the primary controllability metrics in Section 5.

Reconstruction and replacement. Samples that fail automatic verification are reconstructed at most twice using revised cue composition strategies and then verified again. A total of 1,128 samples undergo a second construction round, and 270 require a third. The 54 remaining failures are discarded and replaced within the same cue-configuration stratum and data partition. Replacement samples undergo the same construction and verification procedure to preserve the target allocation.

Human construction audit. As an additional quality check, two annotators independently inspect a stratified subset of 350 retained cue-conditioned utterances, with 50 samples from each of the seven cue categories. Sampling further covers one-, two-, and three-cue configurations and repeated- and mixed-type configurations.

The audit assesses cue recognizability, placement appropriateness, local timing, and natural cue–speech integration. Annotators judge whether the intended cues are perceptually recognizable, whether their placement, duration, and surrounding pauses are appropriate in the local speech context, and whether transitions to surrounding speech are coherent. A sample is accepted when both annotators judge that it satisfies all construction criteria. Disagreements are resolved through joint re-inspection to obtain a consensus decision. When an audited sample does not satisfy these criteria, the identified issue is used to revise the corresponding cue composition strategy, and the sample is reconstructed before finalization.

Audit results. Of the 350 audited cue-conditioned utterances, 333 (95.1%) satisfy all construction criteria after the initial audit, including consensus resolution where needed. The remaining 17 samples are returned for reconstruction. The identified issues primarily involve pause or cue-duration patterns (6 samples), placement relative to the surrounding speech context (5 samples), cue–speech transition quality (4 samples), and insufficient cue salience or inappropriate local gain (2 samples). After reconstruction and repeated inspection, all audited samples satisfy the construction criteria and are retained in the finalized corpus.

Table 8: Cue-count distribution of SymCue and SymCueBench. The full-cue evaluation uses the complete SymCueBench, while the shared-cue setting is a 292-request subset.

Split / Setting	1 cue	2 cues	3 cues	Total
SymCue (train)	1,644	2,002	1,312	4,958
SymCueBench (eval.; full-cue)	412	500	328	1,240
Shared-cue subset	172	90	30	292
Full corpus	2,056	2,502	1,640	6,198

The purpose of this audit is not to replace the automatic verification stage, but to identify systematic artifacts that may not be captured reliably by cue-sequence checks alone. Observed failure patterns are used to revise the corresponding cue-specific composition settings before corpus finalization.

C.3 Temporal Annotation and Verification

Automatic temporal annotation. Only cue-conditioned waveforms that pass construction verification proceed to temporal annotation. Qwen3-Omni automatically identifies the realized cue regions in each retained waveform. Each annotation records the cue type, onset and offset times, and its associated requested lexical boundary. These realized temporal intervals are mapped to speech-token and acoustic-frame resolutions to construct the cue-region masks used in Section 3.

Requested lexical boundaries guide cue placement during construction, whereas the realized onset–offset intervals define the localized training masks. This distinction accounts for timing changes introduced by cue duration, pauses, and cue–speech transitions. Base waveforms contain no cue regions and therefore receive zero masks.

Human annotation audit. A separate stratified subset of 350 automatically annotated cue-conditioned utterances is independently inspected by two annotators, with 50 samples from each of the seven cue categories. Sampling further covers one-, two-, and three-cue configurations and repeated- and mixed-type configurations.

Annotators verify the perceived cue type, count, and order together with the corresponding onset–offset intervals and requested lexical-boundary associations. An annotation is accepted when both annotators judge it to be correct. Disagreements are resolved through joint re-inspection to obtain a consensus decision.

Of the 350 audited annotations, 338 (96.6%) satisfy all annotation criteria after the initial audit. The remaining 12 contain incorrect or incomplete cue-region annotations and are returned for re-annotation. After correction and repeated inspection, all audited annotations satisfy the annotation criteria.

C.4 Leakage-Controlled Partitioning

Training and evaluation resource pools are defined before data construction. Base texts, speaker-matched cue recordings, and speaker-reference recordings are assigned to disjoint pools for SymCue and SymCueBench. Speech construction, automatic verification, reconstruction, temporal annotation, human auditing, and replacement remain within the assigned partition. Each paired unit and all cue-order variants derived from the same base text are kept together.

The final verified corpus contains 4,958 training pairs in **SymCue** and 1,240 evaluation pairs in **SymCueBench**. Speaker identities are shared across partitions, so evaluation measures generalization to held-out base texts, cue-recording instances, and speaker-reference recordings for the same six speaker identities.

Data distribution. Table 8 summarizes the cue-count distribution of SymCue and SymCueBench. The training and evaluation partitions preserve similar proportions of one-, two-, and three-cue configurations while using disjoint base texts, cue-recording instances, and speaker-reference recordings.

Table 9: Composition of the preference data used for Stage-III preference optimization.

Preference-data statistic	Count
Training inputs considered	4,958
Candidates per input	4
Candidate utterances	19,832
Sequence-based preference pairs	2,846
Perceptual preference pairs	1,370
Retained preference pairs	4,216
Inputs without a retained pair	742

Multi-cue configurations include both same-type repetitions and mixed-type sequences. In the full corpus, approximately balanced coverage is maintained between these two composition families. All cue-order variants derived from the same base text remain within the same partition.

Evaluation-request distribution. The full-cue setting uses the complete 1,240-request SymCueBench, whereas the shared-cue setting uses a 292-request subset restricted to cough, throat clearing, and sigh so that all compared models are evaluated on the same cue inventory (Table 8).

The shared-cue subset preserves the original base texts, speaker references, cue configurations, and requested lexical boundaries from SymCueBench rather than forming a separately constructed benchmark. Repeated and mixed configurations are both represented, and mixed-cue examples include distinct cue orders.

D Additional Experimental Details

D.1 Training and Generation Details

HealthTTS training. HealthTTS is initialized from CosyVoice3. The speech tokenizer and vocoder remain frozen throughout training. Supervised adaptation follows the sequential procedure described in Section 3. In Stage I, only the autoregressive speech language model is updated while the flow decoder remains fixed. In Stage II, the adapted language model is frozen and only the flow decoder is updated. Supervised adaptation is performed for three epochs with an effective batch size of 32. We use AdamW with learning rates of 2×10^{-5} for the language model and 1×10^{-5} for the flow decoder, a 5% warm-up ratio, weight decay of 0.01, and gradient clipping at 1.0. The cue-weighting coefficients are set to $\alpha = 2.0$ and $\beta = 2.0$.

Preference-pair construction. Preference data are constructed from the cue-conditioned portion of the SymCue training split. For each training input, we generate four candidate utterances using the Stage-I language model together with the Stage-II flow decoder under identical text, speaker reference, cue configuration, and decoding settings. This produces 19,832 candidate utterances from 4,958 cue-conditioned inputs.

Preference construction follows a two-level criterion. First, Qwen3-Omni identifies the realized cue sequence for each candidate. When candidates differ in cue-sequence compliance, a candidate that exactly matches the requested cue type, count, and order is preferred over one containing a missing, substituted, duplicated, or incorrectly ordered cue. Second, when multiple candidates correctly reproduce the requested cue sequence, Qwen3-Omni performs model-blinded perceptual comparison based on cue realism and the naturalness of cue–speech integration. Comparisons for which no clear preference can be established are discarded.

We retain at most one chosen–rejected pair per training input. This results in 4,216 preference pairs, of which 2,846 are determined by cue-sequence compliance and 1,370 by perceptual quality; 742 inputs are discarded because no sufficiently clear preference can be established.

Preference pairs are constructed once before DPO training and remain fixed throughout optimization. Qwen3-Omni is used only for offline preference construction and is not queried during DPO training.

Table 10: Models and control interfaces used in evaluation.

Model	Model Release	Cue Control
CosyVoice3	Official 0.5B checkpoint	Instruction / inline markers
Fish Speech S2 Pro	Official S2 Pro	Native text control
Chatterbox-Turbo	Official Turbo checkpoint	Paralinguistic tags
Dia	Official 1.6B checkpoint	Inline nonverbal tags
HealthTTS	CosyVoice3-based	Non-lexical control tokens

Generation settings. Each request is generated three times using independent random seeds, and evaluation metrics are computed over the pooled generations from all three seeds. Sampling parameters are fixed within each model throughout evaluation. We use the recommended decoding configuration of each released model rather than imposing a single sampling configuration across model architectures. For HealthTTS, autoregressive decoding uses temperature 0.8 and top-p 0.95. All models receive the same base text, speaker reference, and underlying cue configuration, with the configuration translated into the control representation supported by each model.

A retry is permitted only for technical failures such as an empty, corrupted, or prematurely terminated waveform. Semantic failures, including omitted, duplicated, or incorrectly ordered cues, are not regenerated. If a technical failure remains after one retry, it is treated as incorrect for cue-control metrics and excluded from waveform-dependent quality metrics.

Cue Insertion. Cue Insertion is a post-hoc cue-insertion reference rather than a generative model. We first synthesize base speech using CosyVoice3 and align the requested lexical boundaries to the generated waveform. A speaker-matched cue recording is then inserted at each requested boundary. Cue loudness is normalized to the local speech level, a short local pause of at most 80 ms is permitted when necessary, and 40 ms cross-fades are applied at cue–speech transitions.

The procedure follows the same general composition principles used during SymCue construction, but evaluation uses held-out cue recordings and does not reuse the exact cue waveform used to construct the corresponding SymCueBench example. Cue Insertion therefore serves as a post-hoc cue-insertion reference rather than having access to the evaluation reference.

D.2 Evaluation Metrics

Qwen3-Omni serves as the primary evaluator for cue realization and perceptual speech quality over the complete evaluation sets. Human evaluation is used only for independent validation on stratified subsets, as described in Appendix D.3. For each generated utterance, Qwen3-Omni receives the audio without access to the requested cue configuration or generation condition. For Speaker Similarity, the corresponding speaker reference is additionally provided.

Cue controllability. Cue Match measures cue realization independent of order. We compute F1 for each cue category included in the evaluation setting and report macro-F1 across categories. Repeated occurrences are counted separately, such that a request containing two coughs requires two detected cough instances for both occurrences to be matched.

Cue Count Accuracy measures whether the total number of detected cues equals the requested number, regardless of cue type or order. Cue Sequence Accuracy requires exact agreement between the requested and detected cue sequences in cue type, count, and order. Thus, cough–sigh and sigh–cough are treated as different sequences, and repeated cues must occur the requested number of times.

The controllability evaluation therefore focuses on cue type, count, and order. Requested lexical boundaries are provided as conditioning information but are not scored separately. Cue Count Accuracy and Cue Sequence Accuracy are reported as the percentage of pooled generated utterances satisfying the corresponding criterion.

Table 11: Five-point perceptual rubrics used for Qwen3-Omni and human evaluation. Intermediate scores represent gradations between the anchor descriptions.

Dimension	Score	Low anchor	High anchor
Speech Naturalness	1 / 5	Clearly unnatural, with severe artifacts, discontinuities, or implausible cue–speech transitions	Highly natural and fluent, with coherent integration of cues into the surrounding speech
Cue Realism	1 / 5	The perceived cue sounds clearly artificial, distorted, or unlike a plausible human vocalization	The perceived cue sounds highly realistic and consistent with a natural human vocalization
Speaker Similarity	1 / 5	Generated speech sounds clearly different from the reference speaker	Generated speech strongly matches the vocal identity of the reference speaker

Speech preservation and perceptual quality. WER is computed using Whisper-large-v3 Radford et al. (2023) after lowercasing, punctuation removal, and removal of cue-control symbols from the transcript.

Speaker Similarity, Speech Naturalness, and Cue Realism are evaluated by Qwen3-Omni on 1–5 scales using fixed rubrics. Speaker Similarity scores are divided by 5 before reporting. Speaker Similarity measures how closely the generated voice matches the corresponding speaker reference. Speech Naturalness assesses the overall fluency and perceptual naturalness of the complete utterance. Cue Realism assesses how realistic the audible symptom-related acoustic cues sound. Qwen3-Omni evaluates cue realism from the generated audio alone, and Cue Realism is reported only for successfully realized requested cues.

A stratified subset of generated utterances is independently evaluated by human annotators using the same perceptual rubrics. The correspondence between Qwen3-Omni and human evaluation is reported in Appendix D.3.

Evaluator input and blinding. Qwen3-Omni receives only the generated audio and the evaluation instruction. It is not given the requested cue configuration, generation method, benchmark condition, or reference cue sequence. For Speaker Similarity, the corresponding speaker-reference audio is additionally provided. This design prevents the evaluator from inferring correctness from the conditioning request itself.

We use a fixed evaluator prompt and output schema for all systems. Evaluator decoding uses temperature 0.1 and top-p 0.8. Malformed outputs are retried once using the identical audio and instruction; the evaluator is not resampled to select among multiple valid judgments.

The returned sequence is subsequently compared with the requested configuration to compute Cue Match, Cue Count Accuracy, and Cue Sequence Accuracy. The requested cue configuration is used only during this metric computation and is never provided to the evaluator.

Perceptual evaluation prompt. For perceptual evaluation, Qwen3-Omni is instructed to judge the complete utterance using the fixed criteria in Table 11. Speech Naturalness is rated for the complete utterance. Cue Realism is judged from the audible symptom-related acoustic cues in the generated audio, with scores retained for evaluation only for successfully realized requested cues. Speaker Similarity is evaluated with access to the corresponding speaker-reference recording.

For scores between the two anchors, the evaluator uses the following interpretation: 2 indicates substantial perceptual mismatch or unnaturalness, 3 indicates acceptable but clearly noticeable imperfections, and 4 indicates minor imperfections while remaining largely natural, realistic, or speaker-consistent.

Table 12: Comparison between Qwen3-Omni and human judgments on 63 pilot and 140 main-experiment utterances. Cue sequences are evaluated using exact agreement (%), while perceptual ratings are compared using Pearson's correlation coefficient (r).

Setting	Dimension	Human–Omni Correspondence
Pilot	Cue Sequence	87.3%
Pilot	Speech Naturalness	$r = 0.83$
Main	Cue Sequence	86.79%
Main	Speaker Similarity	$r = 0.80$
Main	Speech Naturalness	$r = 0.77$
Main	Cue Realism	$r = 0.81$

D.3 Human Validation of Qwen3-Omni Evaluation

Qwen3-Omni serves as the primary evaluator for cue realization and perceptual quality over the complete evaluation sets. Human evaluation is conducted only on stratified subsets to validate the reliability of the automatic evaluator and does not contribute to the reported benchmark scores.

For the pilot study, 63 utterances are sampled across all seven TTS models and the single-cue, repeated-cue, and mixed-cue configurations. For the main experiments, 140 utterances are sampled across the seven cue categories and balanced as closely as possible across evaluated models, cue counts, and repeated versus mixed-cue configurations.

Two third-year medical students independently evaluate the sampled utterances without access to the generation method or Qwen3-Omni outputs. For both the pilot and main-experiment subsets, annotators record the perceived cue sequence and rate Speech Naturalness on a 1–5 scale. For the main-experiment subset, they additionally rate Speaker Similarity and Cue Realism using the same perceptual rubrics provided to Qwen3-Omni. Speaker Similarity is evaluated with access to the corresponding speaker reference.

For cue-sequence annotation, a human label is accepted directly when both annotators report the same perceived sequence. Disagreements are resolved through joint re-evaluation to obtain a consensus sequence. The resulting human sequence is compared with the Qwen3-Omni prediction using exact agreement. For perceptual ratings, the two human scores are averaged for each utterance, and Pearson's correlation coefficient is computed between the mean human score and the corresponding Qwen3-Omni score.

The human evaluation shows strong correspondence with Qwen3-Omni across cue realization and perceptual speech quality. We therefore use Qwen3-Omni to evaluate the complete evaluation sets, while the human annotations serve only as an independent validation of the automatic evaluation protocol.

D.4 Ablation Training Protocol

All ablation models start from the same CosyVoice3 checkpoint and use the same SymCue training split, optimizer, batch size, number of updates, and random-seed configuration as HealthTTS. They are evaluated in the full-cue setting of SymCueBench using the same three generation seeds.

For *w/o LM Weighting*, the additional language-model weight on cue-aligned speech tokens is removed by setting $\alpha = 0$, while flow cue weighting and DPO are retained. For *w/o Flow Weighting*, $\beta = 0$, while LM cue weighting and DPO are retained. For *w/o DPO*, the model is evaluated directly after Stage II without preference-based refinement.

To prevent preference-data differences from confounding the comparison, variants that retain DPO use the same chosen–rejected preference pairs. The frozen reference model is the corresponding Stage-I language model for each variant. Thus, each ablation removes only the targeted component while preserving the remaining training procedure as closely as possible.

Table 13: Composition of the 14 VSP evaluation scenarios. Two scenarios are included for each symptom-related acoustic cue category. Scenario wording is held fixed across the three synthesis conditions.

ID	Scenario focus	Target cue
S1	Respiratory complaint A	Cough
S2	Respiratory complaint B	Cough
S3	Throat discomfort A	Throat clearing
S4	Throat discomfort B	Throat clearing
S5	Fatigue / breath discomfort A	Sigh
S6	Fatigue / breath discomfort B	Sigh
S7	Persistent hiccup presentation A	Hiccup
S8	Persistent hiccup presentation B	Hiccup
S9	Nausea presentation A	Retch
S10	Nausea presentation B	Retch
S11	Upper-respiratory presentation A	Sneeze
S12	Upper-respiratory presentation B	Sneeze
S13	Pain presentation A	Pain-related groan
S14	Pain presentation B	Pain-related groan

D.5 Virtual Standardized Patient Evaluation

Participants and scenarios. Thirty medical students from Years 3–5 participate in the evaluation. The study contains 14 VSP scenarios covering the seven symptom-related acoustic cue categories, with two scenarios for each category. Each scenario is rendered under three conditions: base speech generated by CosyVoice3, Cue Insertion, and HealthTTS. Linguistic content and intended speaker identity are kept identical across conditions.

Assignment. Each participant evaluates all 14 scenarios but hears only one condition for each scenario, avoiding repeated exposure to different versions of the same dialogue. Conditions are counterbalanced across participants such that every scenario–condition pair receives ten ratings. This results in 140 judgments for each condition. Scenario order is randomized for each participant, and condition identity is hidden.

After listening to each utterance, participants select the perceived symptom-related acoustic cue from the seven cue categories and rate Patient Realism and Perceived Usefulness on 1–5 scales. Patient Realism measures how convincing the utterance sounds as a patient presentation, while Perceived Usefulness reflects its perceived suitability for clinical training.

Participant instructions and rating scales. After each audio response, participants first answer a forced-choice symptom-cue recognition question by selecting one of the seven cue categories: cough, throat clearing, sigh, hiccup, retch, sneeze, or pain-related groan. No information about the target cue or generating system is shown.

Participants then answer the following two questions:

- **Patient Realism:** "How convincing does this audio response sound as speech produced by a patient in the presented scenario?"
- **Perceived Usefulness:** "How useful would this type of audio response be for practicing interaction with a virtual standardized patient?"

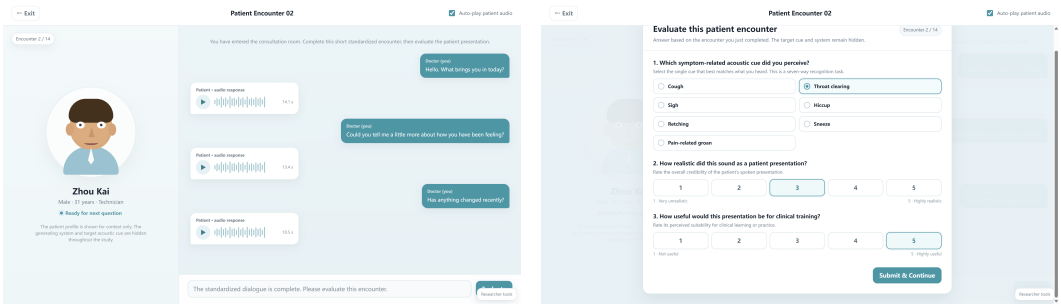
Both dimensions are rated on a five-point Likert scale. For Patient Realism, 1 denotes *not convincing at all* and 5 denotes *highly convincing*. For Perceived Usefulness, 1 denotes *not useful at all* and 5 denotes *highly useful*. Intermediate values indicate increasing degrees of realism or usefulness.

Scoring. Symptom Cue Recognition is computed as the proportion of correct selections among the 140 judgments for each condition:

$$\text{Recognition} = \frac{N_{\text{correct}}}{140} \times 100.$$

Patient Realism and Perceived Usefulness are reported as the mean ratings over the same 140 judgments. The seven-way recognition task has a chance level of $1/7 = 14.29\%$.

CosyVoice3 base speech obtains 21 correct responses (15.00%), close to the 14.29% chance level. Cue Insertion obtains 134 correct responses (95.71%), while HealthTTS obtains 131 (93.57%). The near-chance result for base speech suggests that cue recognition is primarily driven by the acoustic cue rather than linguistic content. HealthTTS further receives higher mean ratings than Cue Insertion for Patient Realism (4.40 vs. 3.50) and Perceived Usefulness (4.50 vs. 3.35), suggesting that cue recognizability alone is insufficient for convincing patient presentation and that natural cue–speech integration also contributes to perceived realism and usefulness.



(a) VSP encounter interface.

(b) Post-encounter questionnaire.

Figure 11: Illustration of the virtual standardized patient (VSP) evaluation interface. (a) The VSP encounter interface shows the patient profile, clinician prompts, and audio responses, while hiding the generating system and target cue from participants. (b) The post-encounter questionnaire collects seven-way Symptom Cue Recognition and 1–5 ratings of Patient Realism and Clinical Usefulness.

Evaluator Prompt: Preference Comparison

Compare the two speech samples.
Both samples are generated with the same text, speaker reference, and requested cue configuration.
First, determine whether either sample better reproduces the requested symptom-related acoustic cue sequence, including cue type, count, and order.
If both samples correctly reproduce the requested cue sequence, select the sample with more realistic symptom-related cues and more natural cue–speech integration.
Return only one of: A, B, or Tie.

Evaluator Prompt: Cue Sequence Verification

Listen to the provided waveform and identify all audible symptom-related acoustic cues.
Use only the following cue categories: cough, throat_clear, sigh, hiccup, retch, sneeze, and pain_related_groan.
Return the detected cue sequence in temporal order.
Do not infer cues from the transcript or any provided target configuration. If no listed cue is audible, return an empty sequence.

Evaluator Prompt: Preference Comparison

Compare the two speech samples.
Both samples are generated with the same text, speaker reference, and requested cue configuration.
First, determine whether either sample better reproduces the requested symptom-related acoustic cue sequence, including cue type, count, and order.

If both samples correctly reproduce the requested cue sequence, select the sample with more realistic symptom-related cues and more natural cue–speech integration.
Return only one of: A, B, or Tie.

Evaluator Prompt: Temporal Cue Annotation

Listen to the provided waveform and identify all audible symptom-related acoustic cues.
For each detected cue, output: cue type, onset time, and offset time.
Use only the following cue categories: cough, throat_clear, sigh, hiccup, retch, sneeze, and pain_related_groan.
Do not use the expected cue configuration when identifying cues.