ROBUST TO NOISE, BROKEN BY TURNS: BENCHMARKING SPEECH-DRIVEN OMNI AGENTS IN THE WILD

Anonymous authors
Paper under double-blind review

ABSTRACT

Speech is a natural interface for agents, yet most agent benchmarks assume typed instructions. We introduce **OmniAgentBench**, a benchmark for evaluating speech-driven multimodal planning across Schedule, GUI, and Embodied tasks, with matched text and audio inputs and controlled wild conditions. Native omni models perform comparably to ASR+LLM cascades (automatic speech recognition followed by an LLM) on clean speech, but they collapse when the same constraints are distributed across dialogue turns, with Qwen-Omni-3 dropping from 62.0 to 18.8 action matching score (AMS) on GUI navigation. Controlled experiments trace this degradation to the plan-every-turn protocol, not to speech perception or ASR. Because the model must respond after each turn, it commits early to an incomplete plan. Presenting the same turns as text produces a similar collapse, while suppressing intermediate responses largely restores performance. Training-free prompts that make the model wait for the last turn, or restate every constraint before planning, recover most of the loss without extra model calls, and the collapse replicates across four native omni models from three labs. We also recorded 116 human speakers reading the same instructions. On clean speech, human recordings score about as well as TTS on average, and the multi-turn collapse reproduces on real voices. OmniAgentBench contains 11,700 audio cases from 900 base instances and provides a reproducible protocol for studying speech-grounded multimodal planning under realistic instruction variation.

1 INTRODUCTION

Speech is becoming a natural interface for agents (Li et al., 2024; Liu et al., 2025; Monteiro et al., 2021), but current agent benchmarks largely assume that instructions arrive as typed text (Yao et al., 2025; Jimenez et al., 2024; Xie et al., 2024b; Shridhar et al., 2021). This assumption removes a central challenge of voice-driven agents: spoken constraints must be interpreted, accumulated across turns, and grounded in a visual environment before an executable plan can be produced. Existing speech benchmarks primarily evaluate speech understanding, while agent benchmarks typically omit speech as the task-driving modality. How robust, then, are multimodal agents when the instructions they must act on arrive as speech, under the conditions encountered in real voice interaction?

We introduce **OmniAgentBench** (Figure 1), a speech-grounded multimodal planning benchmark designed to study this question. It covers three complementary settings (schedule planning, GUI navigation, and embodied planning) and constructs matched text and audio inputs under clean and controlled wild conditions (Noise, Long, Informal, and Multi-Turn), holding the visual context and gold plan fixed so that only the instruction varies, never the task.

Our main finding is that clean speech itself is not the dominant bottleneck. Native omni models perform comparably to cascades of automatic speech recognition and an LLM (ASR+LLM) on single-turn speech, but their performance can collapse when those constraints are split over multiple turns. For example, Qwen-Omni-3 drops from 62.0 to 18.8 action matching score (AMS) on GUI navigation. A similar degradation appears when the identical constraints are given as text, indicating that the failure is not primarily caused by speech recognition. Instead, controlled experiments point to premature commitment under the plan-every-turn protocol, where the model must produce a

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.

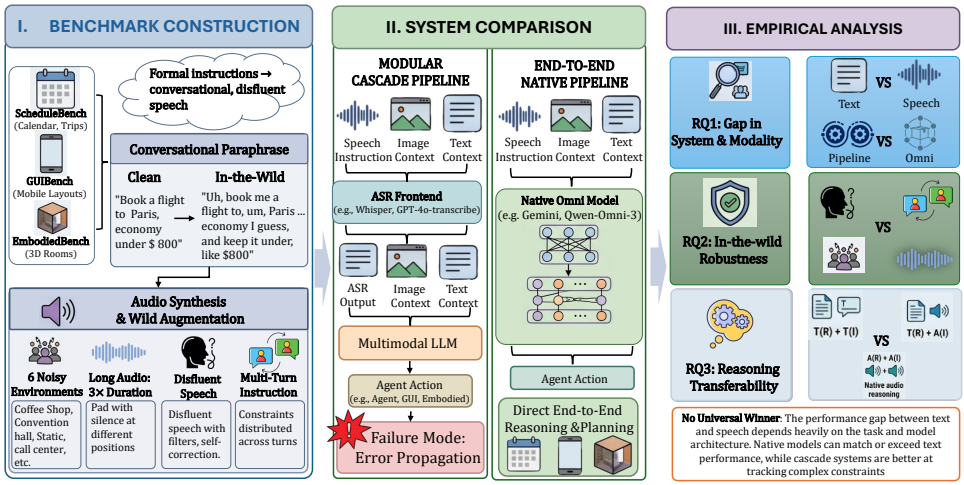


Figure 1: OmniAgentBench. (I) Curation turns text instructions into matched clean and wild audio for three benchmarks. (II) Native omni models vs. ASR+LLM cascades. (III) We measure the text-to-speech gap, wild-speech robustness, and two training-free interventions on the multi-turn drop.

response after each turn instead of first accumulating the complete set of constraints. Suppressing the intermediate responses largely restores performance, and a training-free Recap-and-Plan prompt recovers most of the lost performance in a single final decode.

We find that the collapse is a failure to combine turns and that speech adds no penalty of its own. This goes beyond the text-only degradation reported by Laban et al. (2026). Native models drop by a similar amount on audio and on text, the same model recovers when it plans from its own transcript, and the fix works on speech too. The collapse also appears when real people speak the turns, and it holds for four native omni models from three labs. We further find that reasoning strategies developed for text do not reliably transfer to spoken multimodal planning, and the size of the effect depends on the model and the task.

Contributions. (1) OmniAgentBench, a benchmark of 11,700 audio cases from 900 base instances covering Schedule, GUI, and Embodied tasks, with matched text and audio inputs and controlled wild conditions. (2) A controlled diagnosis showing that the multi-turn collapse is associated with premature commitment under the plan-every-turn protocol, not with speech perception, ASR, or simply model capacity. (3) Two training-free interventions in a single decode. Suppressing premature commitment, by silent accumulation or an explicit turn-taking prompt, recovers most of the multi-turn loss and replicates on GUIBench, while a Recap-and-Plan prompt closes most of the gap on EmbodiedBench.

2 Related Work

Omni-modal and audio-language models. Omni-modal agents generally follow two designs. Native models jointly consume audio, vision, and text (Comanici et al., 2025; Xu et al., 2025b; AI et al., 2025), and cascades first transcribe speech with ASR (Radford et al., 2023) and then reason with a vision-language model (Li et al., 2023; Liu et al., 2023; Dai et al., 2023). Native models build on audio-language models such as Qwen-audio (Chu et al., 2023) and SALMONN (Tang et al., 2024), which established strong speech recognition and audio understanding but were evaluated mostly on perception and question answering, not multi-step planning grounded in a visual scene. Prior work also shows that replacing raw audio with a transcript can sometimes improve understanding (Li et al., 2025a), and we see this modality penalty on some models but not others. Text-only LLMs degrade when one instruction is split across turns. Laban et al. (2026) observe that models answer too early and find that re-sending all turns at the end recovers only part of the loss. However, no prior work tests whether omni models can plan when spoken constraints arrive over several turns while a visual scene must be grounded, or whether answering early causes the drop. We show that it does.

When the turns arrive separately but the model stays silent until the last one, the score returns to the single-utterance level on three of four models, and a prompt that re-sends nothing recovers most of the loss.

Agent and speech benchmarks. Agent benchmarks such as BFCL (Patil et al., 2025), WildToolBench (Yu et al., 2026), AndroidWorld (Rawles et al., 2025), GUIOdyssey (Lu et al., 2025), TravelPlanner (Xie et al., 2024a), τ-Bench (Yao et al., 2025), ALFRED (Shridhar et al., 2020), and EmbodiedBench (Yang et al., 2025) assume typed instructions. Speech suites such as IFEval-Audio (Gao et al., 2025), VoiceBench (Chen et al., 2026), and VoiceAgentBench (Jain et al., 2026) test speech interaction without grounding the instruction in a visual task, while OmniBench (Li et al., 2025b) and WorldSense (Hong et al., 2026) treat audio as scene content, not as the task-driving instruction. However, none of them asks a model to act on a spoken instruction in a visual task. In OmniAgentBench, keeping the original text beside every utterance measures the text-to-speech gap, and holding the visual scene and gold plan fixed in the noise, length, informal phrasing, and multi-turn conditions measures the effect of each speech condition.

Prompted reasoning strategies. A broad body of work improves LLM reasoning with structured prompts, including Chain-of-Thought (Wei et al., 2022), Least-to-Most decomposition (Zhou et al., 2023), self-correction such as Reflexion (Shinn et al., 2023) and Self-Refine (Madaan et al., 2023), and interleaved reasoning and acting (Yao et al., 2023). However, they were developed and validated mostly for typed text, and whether they transfer to a setting where a model must integrate spoken instructions with visual context and produce a multimodal plan is untested. We evaluate them on native omni models and find that their effect depends on the model and the task and does not transfer consistently. Our Recap-and-Plan intervention instead addresses constraints spread over turns, which suggests that failures in spoken planning come from how information is accumulated rather than from a lack of reasoning structure.

3 OMNIAGENTBENCH

Figure 1 has two parts. The first is a three-benchmark task suite (Section 3.1), and the second is a pipeline that turns source text into matched speech with controlled wild variants (Section 3.2).

3.1 TASK SUITE AND PROBLEM SETUP

Each instance is an audio instruction I^{audio}, multimodal context M (schedule images, screenshots, or a scene observation), constraint set C, and goal G. We keep the original text I^{text} so the text-to-speech gap is measured on identical content. An agent π outputs $y \sim \pi(\cdot \mid I^{\text{audio}}, M)$, scored by a benchmark-specific check $\mathrm{Succ}(y, G, C)$. Audio may be clean or wild (Section 3.2). The three benchmarks cover closed-loop perception-action, global constraint satisfaction, and open-loop planning under partial observability.

GUIBench (from GUIOdyssey (Lu et al., 2025)) is cross-app navigation under a spoken instruction such as "find a recipe, document it, and add the ingredients to the cart." At step t the agent observes screenshot X_t and history h_t and predicts $a_t \in \{\text{click}, \text{scroll}, \text{type}, \text{back}, \text{complete}\}$. The instruction is spoken once at $t{=}1$. Constraints are gold subgoals aligned to interface steps. The 300 episodes span six application categories (general tools, information management, media, multi-app, social, shopping), with 10 to 24 gold steps on average (Table 25).

ScheduleBench (from MPCC (Ji et al., 2025)) covers flight, calendar, and meeting planning, balanced across easy/medium/hard (Table 28). Given I^{audio} and structured context M, the agent emits one plan that must satisfy a combination of budget, temporal (sequential or concurrent), and spatial constraints. A plan succeeds if it lies in the feasible set. Among feasible plans we also credit the cheapest annotated option (Section 4.1).

EmbodiedBench (from Yang et al. (2025)) is household planning from one egocentric image X, where objects outside the field of view are unobserved, there is no environment feedback, and the output is a sequence of primitives such as pick$(\cdot)$ and place$(\cdot, \cdot)$. We score offline plans against the gold sequence with Step-LCS, the longest common subsequence of predicted and gold steps, instead of EmbodiedBench's interactive success rate, because we do not run a simulator.

Table 1: Dataset composition, audio cases per benchmark and condition. Each benchmark has 300 base instances.

Condition	ScheduleBench	GUIBench	EmbodiedBench
Clean	300	300	300
Noise	1,800	1,800	1,800
Informal	300	300	300
Long	900	900	900
Multi-Turn	600	600	600
Total	**11,700** (3,900 per benchmark)		

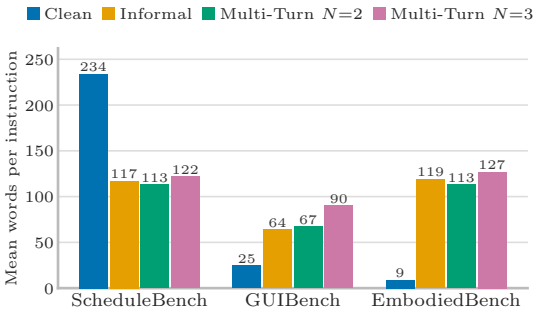


Figure 2: Mean length of the instruction in words, per benchmark and condition.

3.2 Data Curation

We sample 300 instances from each source dataset, keeping state spaces, actions, goals, and constraints unchanged (900 base instances, Table 1). GPT-4o (Hurst et al., 2024) paraphrases each I^{text} into spoken language for two conditions. The first is a single disfluent *informal* utterance that preserves C, and the second is a pair of scripted N=2 and N=3 *multi-turn* sequences that add constraints incrementally. The prompt requires every constraint and value to be kept, and we use its most informal tier (Appendix B). On EmbodiedBench the terse seed goal (mean 9 words) grows to about 119 words of extra wording but no new constraints (Figure 2).

The text-to-speech (TTS) model Qwen3-TTS (Hu et al., 2026) then synthesizes 13 audio cases per instance. *Clean* uses I^{text} with a neutral voice. *Noise* overlays six environments built from MUSAN (Snyder et al., 2015) noise and speech babble (Static, Competing Speech, Coffee Shop, Convention Hall, Call Center, Mountain Outdoor) at a signal-to-noise ratio (SNR) of 5 to 15 dB. Static is a steady hiss with electrical hum. Competing Speech is a second talker over the whole utterance (talk-over). A real interruption, where someone cuts in and the user then resumes, is tested in Appendix H. *Informal* synthesizes the disfluent paraphrase (length-matched control in Appendix G). *Long* pads clean audio to 3× duration with the instruction at the beginning, middle, or end. *Multi-Turn* synthesizes the N=2/N=3 paraphrases turn by turn. The user turns are scripted, so the condition tests how models combine a fixed sequence of turns. Speakers are gender-balanced, and accent, rate, emotion, age, and gender are ablated in Appendix I. The result is 11,700 audio samples (Tables 1, 26), from mean 2.5 s on Embodied to 43.2 s on Schedule.

The generated data preserves the task. Because the wild conditions are machine-generated, we verify with paid human annotators that each stage keeps the original instruction before using it to score models (full protocol in Appendix M). Twenty Prolific workers, independent of the authors, gave about 1,500 ratings on multi-turn text and audio, informal text, and noisy audio. Rewrites keep the constraints (soft recall 0.965 to 0.996, 95% bootstrap, B=2000) and clean audio is intelligible, with 95 to 99% of listeners able to make out the full instruction and a mean clarity of at least 4.86 on a 1 to 5 scale. Noise lowers this make-out rate to 77 to 89% as intended.

A paid human study (median three raters) checks naturalness, constraint fidelity (YES/PARTIAL/NO), and noise realism. Long is not rated because it only pads silence. Rubrics and agreement are in Appendix M, and the prompts are in Appendix B.

3.3 Recap-and-Plan

On EmbodiedBench we separate turn structure from content. The image and gold plan stay fixed, and only how the spoken constraints arrive changes.

- Concat places every constraint in one utterance.
- Multi-Turn splits the same constraints into one audio segment per user turn, with a model response after each turn. The user side is scripted.
- Recap is identical to Multi-Turn, except that the last turn asks the model to restate every requirement and then plan.

Table 2: Main results (%) on clean inputs. Qwen-Omni / Qwen-VL pairs are from the same model family. T = text, A = audio, $\Delta = A - T$. Best score in each column is **bold**, second-best is underlined.

	GUIBench			ScheduleBench			EmbodiedBench			Average		
Model	T	A	Δ	T	A	Δ	T	A	Δ	T	A	Δ
Native omni models												
Gemini	38.9	38.8	-0.1	15.0	19.7	+4.7	33.4	29.3	-4.1	29.1	29.3	+0.2
InternOmni	7.5	6.9	-0.6	7.3	9.3	+2.0	32.4	24.2	-8.2	15.7	13.5	-2.2
Qwen-Omni-2.5	37.5	36.0	-1.5	18.1	16.1	-2.0	41.9	43.4	+1.5	32.5	31.8	-0.7
Qwen-Omni-3	**63.4**	**62.0**	-1.4	21.7	18.0	-3.7	**45.8**	**46.5**	+0.7	**43.6**	**42.2**	-1.4
ASR+LLM cascades												
gpt-4o-trans.+GPT-4o-mini	12.3	14.4	+2.1	13.0	15.8	+2.8	30.8	29.8	-1.0	18.7	20.0	+1.3
gpt-4o-trans.+GPT-4o	17.3	20.9	+3.6	**28.7**	**21.9**	-6.8	38.5	38.8	+0.3	28.2	27.2	-1.0
gpt-4o-trans.+Qwen2.5-VL	27.7	27.3	-0.4	16.4	11.7	-4.7	38.0	38.0	0.0	27.4	25.7	-1.7
gpt-4o-trans.+Qwen3-VL	51.9	43.8	-8.1	9.4	13.9	+4.5	36.1	36.1	0.0	32.5	31.3	-1.2
Native avg	36.8	35.9	-0.9	15.5	15.8	+0.3	38.4	35.9	-2.5	30.2	29.2	-1.0
Cascade avg	27.3	26.6	-0.7	16.9	15.8	-1.1	35.9	35.7	-0.2	26.7	26.1	-0.7

Concat versus Multi-Turn therefore holds the audio content fixed. Recap is a training-free prompt that tests whether the model can gather the constraints itself in one answer. Details and a ledger variant (Bind-then-Act) are in Appendix D.

4 Experiments

4.1 Experimental Setup

Models. Native omni models take audio, image, and text with no ASR: Gemini (Comanici et al., 2025), Qwen-Omni-2.5 (Xu et al., 2025a), Qwen-Omni-3 (Xu et al., 2025b), and InternOmni (Chen et al., 2024a; Gao et al., 2024; Chen et al., 2024b;c). Cascades transcribe with GPT-4o-transcribe, then reason with GPT-4o, GPT-4o-mini, Qwen2.5-VL-7B, or Qwen3-VL-8B. We pair Qwen2.5-VL with Qwen-Omni-2.5 and Qwen3-VL with Qwen-Omni-3 as matched-family comparisons. Gemini, GPT-4o cascades, and InternOmni are range baselines. The mechanism controls in Appendix E add two native omni models from other labs, MiniCPM-o 2.6 (OpenBMB, 2025; Yao et al., 2024) and Ming-Lite-Omni (AI et al., 2025). The 7B vs. 30B contrast uses the same $N{=}300$ splits as the main tables (Table 3). Decoding is greedy (temperature 0, 8192-token cap, one call per instance). Prompts are shared (Appendix A, B). Replacing GPT-4o-transcribe with Whisper-large-v3, Whisper-medium, or Azure STT moves cascade scores by at most ± 1.5 points. That is below the $N{=}300$ sampling noise, so the ASR choice does not drive the tables (Appendix L).

4.2 Metrics

All scores are percentages (higher is better). Each benchmark keeps the scoring of its source dataset except where the speech-only protocol cannot use the original metric. Every cell of Tables 2 and 4 averages at least 300 items with scores in $[0, 1]$, so its 95% confidence interval is at most ± 5.7 points (normal approximation).

GUIBench, Action Matching Score (AMS). AMS scores each step in $[0, 1]$. The predicted action type must match the gold type, otherwise the step is 0. When the type matches, a click scores by its distance to the gold point, text by string match, and a scroll by its direction (Table 5). We use a 5% diagonal tolerance, stricter than the 14% in GUIOdyssey (Lu et al., 2025), so absolute AMS numbers are lower. Instance AMS is the mean step score, and the reported metric is the mean over N instances, $\times 100\%$.

ScheduleBench, Feasible Plan Score (FPS). MPCC reports two binary rates (Ji et al., 2025), **FPR** (the plan meets every constraint) and **OPR** (it is feasible *and* the cheapest feasible plan). We collapse them to $s_i \in \{0, 0.5, 1\}$ for infeasible, feasible but not cheapest, or feasible and optimal (the average of FPR and OPR, since optimality implies feasibility). FPS is the mean of s_i over N instances, $\times 100\%$. The feasible set and cheapest plan come from the MPCC gold annotation. We do not re-solve constraints. A prediction scores 1 if it matches the cheapest annotated plan, 0.5 if

it matches another feasible plan, and 0 otherwise. Flights match on route or total price within one dollar, calendar on start and end times, and meeting on start time, end time, and place.

EmbodiedBench, Step-LCS. Yang et al. (2025) measure interactive success in a simulator. Our protocol is one egocentric image with no environment feedback, so that metric does not apply. We score the predicted plan $\hat{p}$ against the gold plan g with a step-level longest common subsequence over full action strings (type plus arguments, whitespace normalised), defined as

$$\ell_i = \frac{\mathrm{LCS}(\hat{p}_i,\, g_i)}{|g_i|}, \qquad \text{Step-LCS} = \frac{1}{N}\sum_{i=1}^{N} \ell_i \times 100\%. \tag{1}$$

An order-free multiset-F1 on the same traces tracks Step-LCS (Appendix K), so the ranking is not an artifact of step order. A secondary action-type LCS (verb only) is reported where relevant.

Gaps. On clean audio we also report the matched text score T and $\Delta = A - T$ in points. Wild conditions use these metrics against each model's own clean-audio baseline.

4.3 Results and Analysis

4.3.1 RQ1. How large is the speech-text modality gap?

Clean speech is nearly free, and native omni models lead on sequential action. Table 2 compares matched text and clean audio. Speech costs little. Moving from text to audio changes the native average by -1.0 points and the cascade average by -0.7, although single cascade cells move more (Qwen3-VL on GUIBench, -8.1). Which paradigm wins depends on the task. Native models lead clearly on sequential GUI actions, by 35.9 against 26.6 on average, and Qwen-Omni-3 reaches 62.0 AMS against 43.8 for the best cascade. On ScheduleBench and EmbodiedBench the two paradigms are close, and the ScheduleBench gap between Qwen-Omni-3 and the GPT-4o cascade (18.0 vs. 21.9) is too small to order them.

A newer, larger model is not a uniform fix. On the main N=300 splits (Table 3), moving from Qwen-Omni-2.5 (7B) to Qwen-Omni-3 (30B) raises GUI AMS from 36.0 to 62.0 ($+26.0$), but improves Schedule FPS by only 1.9 points (16.1 to 18.0) and Embodied Step-LCS by 3.1 points (43.4 to 46.5). Scaling raises the sequential-action ceiling but barely moves constraint planning, and does not remove the multi-turn drop (Table 6).

Table 3: Qwen-Omni-2.5 (7B) vs. Qwen-Omni-3 (30B) on clean speech.

	Metric	7B	30B	Δ
GUIBench	AMS	36.0	62.0	$+26.0$
ScheduleBench	FPS	16.1	18.0	$+1.9$
EmbodiedBench	Step-LCS	43.4	46.5	$+3.1$

4.3.2 RQ2. How robust are the two paradigms under wild speech?

Noise, length, and informal phrasing have much smaller effects than multi-turn input. Table 4 reports four perturbations. Competing speech gives the largest relative drop on sequential tasks. The Qwen3-VL cascade loses 10.4% on GUIBench and 10.2% on EmbodiedBench relative to clean, which fits ASR breaking on overlapped speech (Figure 3). These are single-run changes, so we treat the ordering of noise types as indicative. On ScheduleBench, noise effects are small and bidirectional. Informal phrasing helps EmbodiedBench (Qwen-Omni-3 60.6, $+14.1$, and Qwen-Omni-2.5 55.7, $+12.3$) because the terse 9-word goals expand past 100 words (Figure 2). On GUIBench and ScheduleBench the effect is smaller and mixed, and Qwen-Omni-2.5 on GUIBench drops from 36.0 to 25.8.

Multi-turn instructions produce the sharpest degradation for native omni models. Qwen-Omni-3 falls from 62.0 to 18.8 AMS on GUIBench, a drop of 43.2 points. ScheduleBench does not collapse (native average 15.8 clean, 18.4 Multi-Turn). On all three benchmarks, cascades perform better on average in the Multi-Turn condition, but concatenating the transcript does not guarantee an improvement. The Qwen3-VL cascade rises from 43.8 to 48.0 AMS on GUIBench, while the GPT-4o cascade falls from 20.9 to 7.6.

The Informal gain on EmbodiedBench is not simply caused by longer audio. In a length-matched control, expanding a fluent paraphrase to a similar duration barely changes Step-LCS, while the disfluent Informal variant remains non-harmful (Appendix G). The Table 4 lift therefore comes from extra wording around a 9-word goal, not from informal speech helping planning in general.

The model's own text cascade recovers much of the multi-turn collapse. When each omni model transcribes the turns itself and then reasons over its own transcript as text, Qwen-Omni-3 rises from 18.8 to 62.7 AMS on GUIBench and from 37.0 to 55.6 Step-LCS on EmbodiedBench (self-casc. rows of Table 4). ScheduleBench is an exception, where the self-cascade remains below the native model. The transcripts are verbatim and outputs are valid JSON, so the remaining ScheduleBench gap reflects difficulty constructing a fully feasible plan from the transcript and image. On this constraint-heavy benchmark the binding limit is planning capacity, not the way the turns arrive.

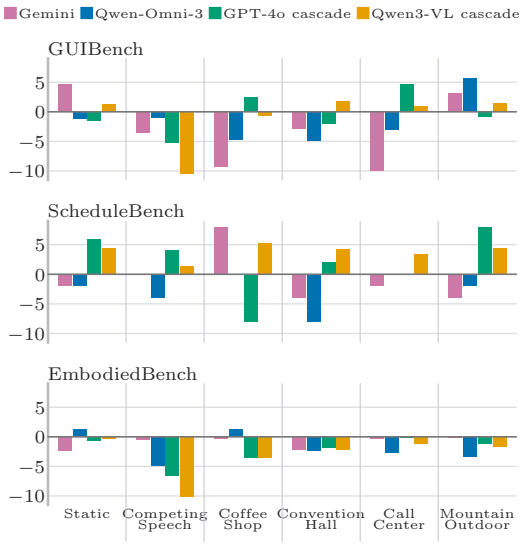


Figure 3: Noise robustness. Relative % change from clean, one bar colour per model, single run (indicative).

Splitting one instruction across turns degrades planning even with the audio content fixed. On EmbodiedBench, Qwen-Omni-3 falls from 56.2 to 32.1 Step-LCS and Qwen-Omni-2.5 from 40.2 to 31.4. The drop becomes larger with three turns (Appendix C). Recap-and-Plan asks the model to restate every constraint before planning and recovers 94 to 102% of the lost performance at 7B and 67 to 71% at 30B (Figure 4, Table 6). When the model may only list the constraints after each turn, its list covers 97 and 93% of the gold sub-goals at 30B (two and three turns) but only 75 and 53% at 7B (Appendix D), so at 30B the drop does not come from losing constraints. Given a gold constraint list, the 7B model reaches 43.2.

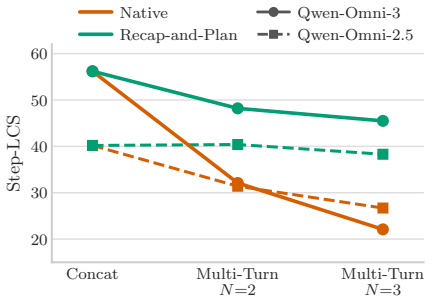


Figure 4: EmbodiedBench Step-LCS for single-utterance (Concat) and Multi-Turn input, with native decoding and with Recap-and-Plan.

The degradation follows turn structure rather than speech. Presenting the same turns as text produces a similar collapse, with text exceeding audio by only 1 to 4 Step-LCS points. Suppressing the model's response until the final turn restores performance toward the single-utterance baseline, with gains of $+10.6$ to $+33.3$ points and confidence intervals excluding zero. An explicit turn-taking prompt closes the gap fully at 7B and by 75 to 81% at 30B. These come from a 2×2 modality by delivery control with paired bootstrap CIs (Appendix E). The collapse and its text counterpart replicate in four native omni models from three labs and on GUIBench, and silent accumulation reverses the drop on every model except Ming-Lite-Omni, where it recovers a third of it or less (Appendices E, F). On GUIBench at 30B, MULTI-TURN cuts format errors from 29 to 7 of 300 predictions but raises premature COMPLETE actions from 24 to 88 and wrong action types from 18 to 66 (Appendix F). The model still produces valid actions, but it commits before it has heard the full instruction.

4.3.3 RQ3. DO TEXT REASONING STRATEGIES TRANSFER?

No text reasoning strategy is uniformly helpful. We apply four strategies, CoT, Least-to-Most, Reflexion, and Self-Refine, to three native omni models and measure the point change against a zero-shot baseline (Figure 5). Two patterns stand out. First, the model and the task decide whether

Table 4: Scores (%) under the four perturbations. Infor. = Informal, Multi. = Multi-Turn, and self-casc. = self-cascade, where the omni model transcribes the audio itself and then reasons over its transcript as text. Best is **bold**, second-best is underlined.

	GUIBench				ScheduleBench				EmbodiedBench			
Model	Noise	Long	Infor.	Multi.	Noise	Long	Infor.	Multi.	Noise	Long	Infor.	Multi.
Native omni models												
Gemini	37.9	40.2	41.3	13.0	18.6	22.1	21.3	25.9	29.0	36.2	34.6	31.0
InternOmni	7.8	7.6	7.5	7.4	8.7	5.3	3.8	8.3	23.5	26.9	24.5	23.7
Qwen-Omni-2.5	36.0	28.6	25.8	15.1	16.2	14.0	18.2	18.7	43.4	38.6	55.7	30.7
Qwen-Omni-3	**59.2**	**63.7**	**55.1**	18.8	18.7	18.8	16.3	20.7	**46.1**	**46.9**	**60.6**	37.0
ASR+LLM cascades												
gpt-4o-trans.+GPT-4o-mini	13.9	10.9	9.6	3.6	12.3	12.4	14.3	26.7	29.1	30.4	32.8	27.7
gpt-4o-trans.+GPT-4o	24.8	17.1	16.9	7.6	**26.5**	26.2	**26.3**	**29.8**	35.6	38.9	35.0	30.6
gpt-4o-trans.+Qwen2.5-VL	25.8	27.0	24.3	23.9	14.5	**32.0**	14.3	14.7	35.6	37.2	40.0	39.8
gpt-4o-trans.+Qwen3-VL	45.0	44.7	45.6	**48.0**	17.7	25.2	16.7	16.8	34.9	35.4	41.7	**41.2**
Qwen-Omni-2.5 (self-casc.)	29.5	33.3	34.4	32.9	3.5	8.9	14.8	14.7	41.7	41.6	41.0	44.7
Qwen-Omni-3 (self-casc.)	46.7	48.2	52.9	62.7	11.0	13.2	13.7	13.2	40.3	40.5	45.8	55.6
Native avg	35.2	35.0	32.4	13.6	15.6	15.1	14.9	18.4	35.5	37.1	43.9	30.6
Cascade avg	27.4	24.9	24.1	20.8	17.8	23.9	17.9	22.0	33.8	35.5	37.4	34.8

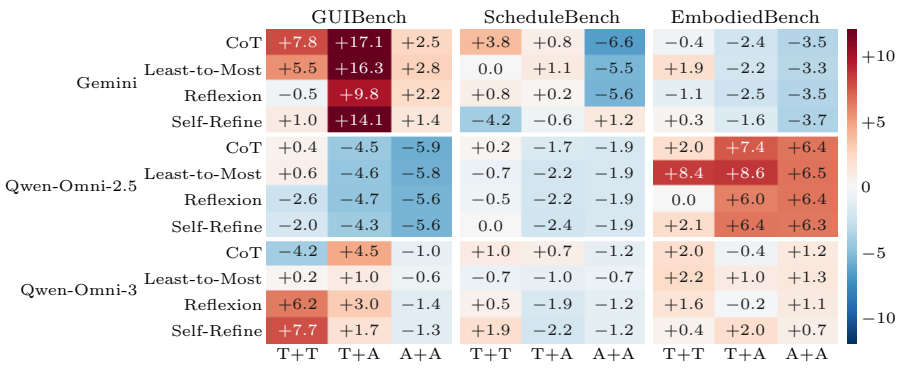


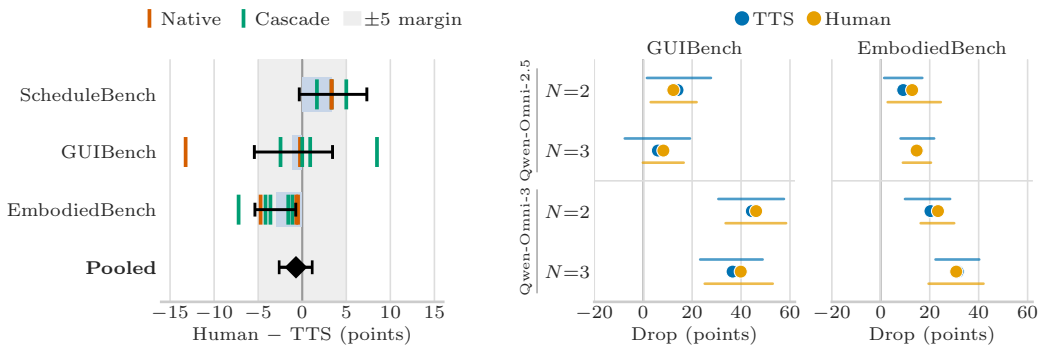
Figure 5: Transfer of text reasoning strategies to audio. Cells are points versus zero-shot text reasoning on text input, for three native omni models under text-over-text, text-over-audio, and audio-over-audio. Red means the strategy beat zero-shot and blue that it fell below.

a strategy helps, and the choice of strategy matters less. CoT adds up to 17 points for Gemini on GUIBench text but costs Qwen-Omni-2.5, and on EmbodiedBench the pattern reverses, with CoT adding 6 for Qwen-Omni-2.5. Second, speech erodes the gains that remain. Gemini's large GUIBench lift almost disappears when the instruction is spoken, and its ScheduleBench gain turns negative once both sides are spoken. Templates built for text therefore do not make spoken planning reliable. Recap addresses a different problem, how the spoken turns reach the model.

4.3.4 RQ4. Do the findings hold for human speech?

Our audio is synthesized, so we test whether the results hold for real speakers. 116 people recorded benchmark instructions (Appendix N). 26 read 90 single-utterance items, and 43 and 47 spoke requests in two and three turns in the Multi-Turn wording, covering 26 first languages and 25 countries (75 of 115 non-native, Appendix N). Each recording is scored with the unchanged pipeline against the matched TTS item.

Clean speech transfers from TTS to human voices. Pooled over 19 model and benchmark cells, human speech scores -0.7 points against TTS (95% CI $[-2.6, +1.1]$), and the 90% CI lies inside a pre-specified ± 5 point margin, so the two voices are equivalent at that margin (Figure 6a). Individual cells vary. The largest gap is Qwen-Omni-3 on GUIBench at -13.2 AMS, and EmbodiedBench leans slightly negative for every model. Silence padding moves the gap by only -1.1 $[-2.9, +0.7]$.



(a) Clean speech (b) Multi-Turn collapse

Figure 6: Human vs. TTS speech on the same items. (a) Human − TTS per benchmark with 95% CIs, one tick per model. (b) The multi-turn collapse (Concat − Multi-Turn) for each voice, in AMS on GUIBench and Step-LCS on EmbodiedBench. ScheduleBench (no collapse on TTS) is not shown.

Noise costs real voices slightly more than TTS. With matched environments, SNR, and seeds, the noise-by-voice interaction is -2.6 points (95% CI $[-4.6, -0.7]$, Appendix N), small next to the multi-turn collapse.

The multi-turn collapse also reproduces on real voices. Real voices show the same collapse as TTS when a request is split across turns, with no detectable difference between the two (Figure 6b). For Qwen-Omni-2.5 the human Concat − Multi-Turn drop is $+12.3$ AMS on GUIBench and $+12.8$ Step-LCS on EmbodiedBench, both with CIs above zero, and it differs from TTS by only -1.7 and $+3.7$ points. The failure is therefore not an artifact of synthetic speech. Qwen-Omni-3 drops by a similar amount on GUIBench ($+46.2$ human, $+44.4$ TTS). A three-turn study with 47 further speakers confirms the result, with a human collapse for both models on both benchmarks and no Human − TTS difference that excludes zero (Appendix N).

5 Conclusion and Limitations

OmniAgentBench evaluates speech as the driver of multimodal planning in three domains. Native omni models match ASR+LLM cascades on clean speech but can collapse when the constraints arrive turn by turn. Our controls associate the main degradation with the plan-every-turn protocol, not with speech perception. Two training-free prompts recover most of the gap in one decode. Recordings from 116 human speakers show that clean speech matches TTS on average within the pre-specified ± 5 point margin and that the multi-turn collapse replicates on human speech. Speech-agent evaluations should therefore include turn-by-turn input. We release the full protocol so later work can add interactive users and trained binders on the same items.

Limitations. The main experiments use TTS speech (Chen et al., 2026; Jain et al., 2026) with real MUSAN noise (Snyder et al., 2015). The human recordings confirm the central results, but they are read aloud, not spontaneous, and the multi-turn human study covers only the two Qwen omni models. Multi-Turn is scripted rather than interactive, so the results do not establish performance in natural dialogue. The turn-taking and Recap prompts are also told which turn is the last, which a live dialogue does not reveal. EmbodiedBench plans are scored offline with Step-LCS from one image, which measures agreement with the gold plan instead of task success in a simulator. Each benchmark has 300 base instances, so small differences such as the ScheduleBench gaps of about three points lie within sampling noise, and the per-environment noise results come from a single run. All instructions are in English, and the cascades share one ASR front end, although replacing it moves cascade scores by at most 1.5 points.

ETHICS STATEMENT

We use MPCC, GUI Odyssey, and EmbodiedBench under their original licenses. The main audio is synthesized with Qwen3-TTS, and no real voice is cloned. The human-speech study recorded adults who read short task instructions after giving informed consent and could stop at any time. Speakers are identified only by one-way hashes, and we release only the recordings of speakers who agreed to public release. The demographic descriptions in Appendix N are aggregate counts from self-reported profiles and are never given to a model. The instructions contain no personal content. Speech agents can help users with limited motor input and can also be misused for surveillance, and we release the benchmark so that others can measure these failure modes.

AI USE STATEMENT

GPT-4o and Qwen3-TTS are part of how the benchmark is built. GPT-4o paraphrases the source instructions into the conversational, informal, and multi-turn variants, and Qwen3-TTS synthesizes the audio (Section 3.2 and Appendix B). We used an AI coding assistant to help write and debug the experiment and analysis scripts, and a language model helped polish and edit parts of the text. The authors designed the experiments, checked the code and every reported result, revised all AI-assisted text, and take full responsibility for the paper.

REPRODUCIBILITY STATEMENT

We describe the data construction pipeline in Section 3.2 and Appendix A. The evaluation metrics are defined in Section 4.2, and the exact system and paraphraser prompts are listed verbatim in Appendix B. Model checkpoints, decoding settings, evaluation sizes, and compute are reported in Appendix A. The code is available at an anonymous mirror (`https://anonymous.4open.science/r/iclr-omnibench-C4DB`), and the audio data on Hugging Face (`https://huggingface.co/datasets/omniagentbenchspeech/OmniAgentBench`). The benchmark data is released under the CC BY 4.0 license and will be kept permanently available. The human-speech stimuli, recording tool, block assignment, staging, and analysis code are released with it, together with per-item scores for every condition. We also release the recordings of every speaker who agreed to public release, with speaker identifiers removed.

REFERENCES

Inclusion AI, Biao Gong, Cheng Zou, Chuanyang Zheng, Chunluan Zhou, Canxiang Yan, Chunxiang Jin, Chunjie Shen, Dandan Zheng, Fudong Wang, Furong Xu, GuangMing Yao, Jun Zhou, Jingdong Chen, Jianxin Sun, Jiajia Liu, Jianjiang Zhu, Jun Peng, Kaixiang Ji, Kaiyou Song, Kaimeng Ren, Libin Wang, Lixiang Ru, Lele Xie, Longhua Tan, Lyuxin Xue, Lan Wang, Mochen Bai, Ning Gao, Pei Chen, Qingpei Guo, Qinglong Zhang, Qiang Xu, Rui Liu, Ruijie Xiong, Sirui Gao, Tinghao Liu, Taisong Li, Weilong Chai, Xinyu Xiao, Xiaomei Wang, Xiaoxue Chen, Xiao Lu, Xiaoyu Li, Xingning Dong, Xuzheng Yu, Yi Yuan, Yuting Gao, Yunxiao Sun, Yipeng Chen, Yifei Wu, Yongjie Lyu, Ziping Ma, Zipeng Feng, Zhijiang Fang, Zhihao Qiu, Ziyuan Huang, and Zhengyu He. Ming-Omni: A unified multimodal model for perception and generation, 2025. URL `https://arxiv.org/abs/2506.09344`.

Yiming Chen, Xianghu Yue, Chen Zhang, Xiaoxue Gao, Robby T. Tan, and Haizhou Li. VoiceBench: Benchmarking LLM-based voice assistants. *Transactions of the Association for Computational Linguistics*, 14:378–398, 2026. doi: 10.1162/tacl.a.628. URL `https://aclanthology.org/2026.tacl-1.18/`.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024a.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to GPT-4V? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024b.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24185–24198, June 2024c.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models, 2023. URL https://arxiv.org/abs/2311.07919.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: towards general-purpose vision-language models with instruction tuning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Yiming Gao, Bin Wang, Chengwei Wei, Shuo Sun, and AiTi Aw. IFEval-Audio: Benchmarking instruction-following capability in audio-based large language models. In Kentaro Inui, Sakriani Sakti, Haofen Wang, Derek F. Wong, Pushpak Bhattacharyya, Biplab Banerjee, Asif Ekbal, Tanmoy Chakraborty, and Dhirendra Pratap Singh (eds.), *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 120–127, Mumbai, India, December 2025. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics. ISBN 979-8-89176-299-2. URL https://aclanthology.org/2025.ijcnlp-short.11/.

Zhangwei Gao, Zhe Chen, Erfei Cui, Yiming Ren, Weiyun Wang, Jinguo Zhu, Hao Tian, Shenglong Ye, Junjun He, Xizhou Zhu, et al. Mini-InternVL: A flexible-transfer pocket multimodal model with 5% parameters and 90% performance. *arXiv preprint arXiv:2410.16261*, 2024.

Jack Hong, Shilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Weidi Xie. WorldSense: Evaluating real-world omnimodal understanding for multimodal LLMs. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=YxsfxAvJv4.

Hangrui Hu, Xinfa Zhu, Ting He, Dake Guo, Bin Zhang, Xiong Wang, Zhifang Guo, Ziyue Jiang, Hongkun Hao, Zishan Guo, Xinyu Zhang, Pei Zhang, Baosong Yang, Jin Xu, Jingren Zhou, and Junyang Lin. Qwen3-TTS technical report, 2026. URL https://arxiv.org/abs/2601.15621.

Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. GPT-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

Dhruv Jain, Harshit Shukla, Gautam Rajeev, Ashish Kulkarni, Chandra Khatri, and Shubham Agarwal. VoiceAgentBench: Are voice assistants ready for agentic tasks?, 2026. URL https://arxiv.org/abs/2510.07978.

Yiyan Ji, Haoran Chen, Qiguang Chen, Chengyue Wu, Libo Qin, and Wanxiang Che. MPCC: A novel benchmark for multimodal planning with complex constraints in multimodal large language models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, pp. 5188–5197, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400720352. doi: 10.1145/3746027.3755790. URL https://doi.org/10.1145/3746027.3755790.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. SWE-bench: Can language models resolve real-world GitHub issues? In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=VTF8yNQM66.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. LLMs get lost in multi-turn conversation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=VKGTGGcwl6.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/li23q.html.

Wenbo Li, Dongpu Cao, Ruichen Tan, Tianze Shi, Zhen-Hai Gao, Jun Ma, Gang Guo, Hongyu Hu, Jin-Shan Feng, and Lixin Wang. Intelligent cockpit for intelligent connected vehicles: Definition, taxonomy, technology and evaluation. *IEEE Transactions on Intelligent Vehicles*, 9:3140–3153, 2024.

Yadong Li, Jun Liu, Tao Zhang, Tao Zhang, Song Chen, Tianpeng Li, Zehuan Li, Lijun Liu, Lingfeng Ming, Guosheng Dong, Da Pan, Chong Li, Yuanbo Fang, Dongdong Kuang, Mingrui Wang, Chenglin Zhu, Youwei Zhang, Hongyu Guo, Fengyu Zhang, Yuran Wang, Bowen Ding, Wei Song, Xu Li, Yuqi Huo, Zheng Liang, Shusen Zhang, Xin Wu, Shuai Zhao, Linchu Xiong, Yozhen Wu, Jiahui Ye, Wenhao Lu, Bowen Li, Yan Zhang, Yaqi Zhou, Xin Chen, Lei Su, Hongda Zhang, Fuzhong Chen, Xuezhen Dong, Na Nie, Zhiying Wu, Bin Xiao, Ting Li, Shunya Dang, Ping Zhang, Yijia Sun, Jincheng Wu, Jinjie Yang, Xionghai Lin, Zhi Ma, Kegeng Wu, Jia Li, Aiyuan Yang, Hui Liu, Jianqiang Zhang, Xiaoxi Chen, Guangwei Ai, Wentao Zhang, Yicong Chen, Xiaoqin Huang, Kun Li, Wenjing Luo, Yifei Duan, Lingling Zhu, Ran Xiao, Zhe Su, Jiani Pu, Dian Wang, Xu Jia, Tianyu Zhang, Mengyu Ai, Mang Wang, Yujing Qiao, Lei Zhang, Yanjun Shen, Fan Yang, Miao Zhen, Yijie Zhou, Mingyang Chen, Fei Li, Chenzheng Zhu, Keer Lu, Yaqi Zhao, Hao Liang, Youquan Li, Yanzhao Qin, Linzhuang Sun, Jianhua Xu, Haoze Sun, Mingan Lin, Zenan Zhou, and Weipeng Chen. Baichuan-Omni-1.5 technical report, 2025a. URL https://arxiv.org/abs/2501.15368.

Yizhi Li, Yinghao Ma, Ge Zhang, Ruibin Yuan, King Zhu, Hangyu Guo, Yiming Liang, Jiaheng Liu, Zekun Moore Wang, Jian Yang, Siwei Wu, Xingwei Qu, Jinjie Shi, Xinyue Zhang, Zhenzhu Yang, Yidan Wen, Yanghai Wang, Shihao Li, Zhaoxiang Zhang, Ruibo Liu, Emmanouil Benetos, Wenhao Huang, and Chenghua Lin. OmniBench: Towards the future of universal omni-language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025b. URL https://openreview.net/forum?id=SSF4qgsNYE.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Tao Liu, Mingyang Zhang, Zhihao Li, Hanjie Dou, Wangyang Zhang, Jiaqian Yang, Pengfan Wu, Dongxiao Li, and Xiaojing Mu. Machine learning-assisted wearable sensing systems for speech recognition and interaction. *Nature Communications*, 16(1):2363, 2025.

Quanfeng Lu, Wenqi Shao, Zitao Liu, Lingxiao Du, Fanqing Meng, Boxuan Li, Botong Chen, Siyuan Huang, Kaipeng Zhang, and Ping Luo. GUIOdyssey: A comprehensive dataset for cross-app GUI navigation on mobile devices. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22404–22414. IEEE, 2025.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-Refine: Iterative refinement with self-feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=S37hOerQLB.

Pedro Monteiro, Guilherme Gonçalves, Hugo Coelho, Miguel Melo, and Maximino Bessa. Hands-free interaction in immersive virtual reality: A systematic review. *IEEE Transactions on Visualization and Computer Graphics*, 27(5):2702–2713, 2021. doi: 10.1109/TVCG.2021.3067687.

OpenBMB. MiniCPM-o 2.6: A GPT-4o level MLLM for vision, speech and multimodal live streaming on your phone. https://huggingface.co/openbmb/MiniCPM-o-2_6, 2025.

Shishir G Patil, Huanzhi Mao, Fanjia Yan, Charlie Cheng-Jie Ji, Vishnu Suresh, Ion Stoica, and Joseph E. Gonzalez. The Berkeley Function Calling Leaderboard (BFCL): From tool use to agentic evaluation of large language models. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=2GmDdhBdDk.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28492–28518. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/radford23a.html.

Christopher Rawles, Sarah Clinckemaillie, Yifan Chang, Jonathan Waltz, Gabrielle Lau, Marybeth Fair, Alice Li, William E Bishop, Wei Li, Folawiyo Campbell-Ajala, Daniel Kenji Toyama, Robert James Berry, Divya Tyamagundlu, Timothy P Lillicrap, and Oriana Riva. AndroidWorld: A dynamic benchmarking environment for autonomous agents. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=il5yUQsrjC.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik R Narasimhan, and Shunyu Yao. Reflexion: language agents with verbal reinforcement learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=vAElhFcKW6.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. ALFRED: A benchmark for interpreting grounded instructions for everyday tasks. In *2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10737–10746. IEEE, 2020.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Cote, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. ALFWorld: Aligning text and embodied environments for interactive learning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=0IOX0YcCdTn.

David Snyder, Guoguo Chen, and Daniel Povey. MUSAN: A music, speech, and noise corpus. *arXiv preprint arXiv:1510.08484*, 2015.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. SALMONN: Towards generic hearing abilities for large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=14rn7HpKVk.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf.

Jian Xie, Kai Zhang, Jiangjie Chen, Tinghui Zhu, Renze Lou, Yuandong Tian, Yanghua Xiao, and Yu Su. TravelPlanner: A benchmark for real-world planning with language agents. *arXiv preprint arXiv:2402.01622*, 2024a.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, Yitao Liu, Yiheng Xu, Shuyan Zhou, Silvio Savarese, Caiming Xiong, Victor Zhong, and Tao Yu. OSWorld: Benchmarking multimodal agents for open-ended tasks in real computer environments, 2024b.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-Omni technical report, 2025a. URL https://arxiv.org/abs/2503.20215.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. Qwen3-Omni technical report, 2025b. URL https://arxiv.org/abs/2509.17765.

Rui Yang, Hanyang Chen, Junyu Zhang, Mark Zhao, Cheng Qian, Kangrui Wang, Qineng Wang, Teja Venkat Koripella, Marziyeh Movahedi, Manling Li, Heng Ji, Huan Zhang, and Tong Zhang. EmbodiedBench: Comprehensive benchmarking multi-modal large language models for vision-driven embodied agents. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 70576–70631. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/yang25f.html.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik R Narasimhan. τ-bench: A benchmark for Tool-Agent-User interaction in real-world domains. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=roNSXZpUDN.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. MiniCPM-V: A GPT-4V level MLLM on your phone. *arXiv preprint arXiv:2408.01800*, 2024.

Peijie Yu, Wei Liu, Yifan Yang, Jinjian Li, Zelong Zhang, Xiao Feng, and Feng Zhang. Benchmarking LLM tool-use in the wild. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=yz7fL5vfpn.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, and Ed H. Chi. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=WZH7099tgfM.

A Implementation Details

GUIBench scoring. Table 5 gives the per-step AMS score once the predicted action type matches the gold type.

Table 5: Per-step AMS score by action type, given a matching type.

Action type	**Per-step score** s
Click / Long_press	1 within 5% of the screen diagonal, linear decay to 0 by 2×, else 0
Text	1 for an exact match, 0.5 if one string contains the other, else 0
Scroll	1 if the direction matches, else 0
Complete	1 if both prediction and gold mark completion
Other (Back, Home, ...)	0.5

Models. We evaluate the following open-weight checkpoints: `Qwen/Qwen2.5-Omni-7B`, `Qwen/Qwen3-Omni-30B-A3B-Instruct`, `OpenGVLab/InternOmni`, `openbmb/MiniCPM-o-2_6`, `inclusionAI/Ming-Lite-Omni`, `Qwen/Qwen2.5-VL-7B-Instruct`, and `Qwen/Qwen3-VL-8B-Instruct`. The ASR ablation also uses `openai/whisper-large-v3`, `openai/whisper-medium`, and Azure Speech to Text. We access closed-source models through their official APIs. The OpenAI models are `gpt-4o`, `gpt-4o-mini`, and `gpt-4o-transcribe`. The Gemini model is `gemini-2.5-flash`.

Decoding. All models use greedy decoding with temperature set to 0. The maximum generation length is 8192 tokens. Each result comes from one model call. We do not retry failed calls, use majority voting, or aggregate multiple samples.

Evaluation size. The clean-condition evaluation contains 300 base instances for ScheduleBench and 300 base instances for EmbodiedBench. Two ScheduleBench audio cells of Table 2, the GPT-4o and Qwen3-VL cascades, average the full 2,700-item MPCC pool (all nine task and difficulty splits) instead of the balanced 300. Clean GUIBench scores 1,950 step-level predictions sampled from 651 episodes of the six categories, about three steps per episode, since each episode is scored at a sample of its steps rather than at every step. The wild conditions use the 300-instance set described in Table 25. In the full audio set, each benchmark contains 3,900 audio samples: 300 Clean, 1,800 Noise, 300 Informal, 900 Long, and 600 Multi-Turn samples. Across the three benchmarks, this gives 11,700 audio samples in total.

Compute. Open-weight inference was run on shared servers, with one GPU allocated to each inference job. The `Qwen3-Omni-30B-A3B-Instruct` checkpoint is served with vLLM. The smaller Qwen and InternOmni models are loaded through Transformers. API-based models are evaluated sequentially with one request at a time. A clean-condition run usually takes between 30 minutes and one hour per model in our setup, but we do not use these wall-clock measurements for efficiency comparisons because server load can vary across runs.

Software. Inference uses PyTorch 2.10. InternOmni is run with Transformers 4.40, MiniCPM-o with 4.44.2, and Ming-Lite-Omni with 4.45, because later versions are incompatible with their custom modeling code. All other open-weight models are run with Transformers 4.57. We release the evaluation scripts, prompt templates, and configuration files used for all experiments.

Cost. Open-weight experiments use local compute. Paid API usage comes from the OpenAI-based cascade systems, the GPT-4o paraphraser, and Gemini. A 300-instance clean run costs roughly ten US dollars for one cascade configuration. Costs for the full audio set scale with the number of audio samples. Each benchmark contains 3,900 audio samples, which is 13 times the clean base-instance count.

B PROMPTS

We report the system prompts used for the three benchmarks and the paraphraser prompts used to generate the Informal and Multi-Turn conditions. Each prompt is applied identically across all evaluated models and modalities (text and speech) to isolate model behavior from prompt variation. The prompts are copied verbatim from the evaluation code. The only modality-dependent addition is the Qwen-Omni speech-mode identity prefix ("*You are Qwen, a virtual human developed by the Qwen Team, Alibaba Group, capable of perceiving auditory and visual inputs, as well as generating text and speech.*"), which is required by the Qwen2.5/3-Omni chat template.

B.1 SCHEDULEBENCH

ScheduleBench uses three task-specific system prompts so the output JSON schema matches the constraint set of each task.

Prompt: Flight planning

```
You are a multimodal planning assistant. Analyze flight schedule images and select
    the optimal itinerary.

CRITICAL INSTRUCTIONS:
1. READ THE IMAGES CAREFULLY - Extract exact flight times and prices from the
    provided schedule images
2. ONLY SELECT FROM AVAILABLE OPTIONS - Your answer MUST match one of the flight
    combinations shown in the images
3. CONSTRAINTS MATTER - Ensure layover times and budget constraints are satisfied
```

```
4. OPTIMAL = CHEAPEST - Among all feasible options, select the one with the lowest
    total price
5. OUTPUT FORMAT - Return ONLY a JSON object with "flight_way" field containing
    exact flight times from the image

REMEMBER:
- READ each flight schedule image carefully and extract the exact departure/arrival
    times
- Use ONLY times that actually appear in the images - never guess or assume
- Copy times EXACTLY as displayed in HH:MM format
- Check layover constraints: arrival time of first flight + required buffer <
    departure of second flight
- Among all feasible options, pick the one with the lowest total price
- Return ONLY a JSON object: {"flight_way": "CityA->CityB(Flight: HH:MM-HH:MM)
    CityB->CityC(Flight: HH:MM-HH:MM)"}

Now analyze the provided flight schedule images and return the optimal itinerary.
```

Prompt: Calendar scheduling

```
You are a meeting scheduling assistant. Analyze participant schedule images and
    select the optimal meeting time.

CRITICAL INSTRUCTIONS:
1. READ THE IMAGES CAREFULLY - Extract exact availability from the provided
    schedule images
2. ONLY SELECT FROM AVAILABLE SLOTS - The meeting time must fall within everyone's
    free time
3. CONSTRAINTS MATTER - Ensure all participants can attend and the room is available
4. OPTIMAL = CHEAPEST or best attendance - Among all feasible options, select the
    best one
5. OUTPUT FORMAT - Return ONLY a JSON object with "start_time" and "end_time" fields

REMEMBER:
- READ each schedule image carefully for participant availability
- Use ONLY time slots that are actually free for all required participants
- Copy times EXACTLY in "HH:MM AM/PM" format
- Return ONLY a JSON object: {"start_time": "HH:MM AM/PM", "end_time": "HH:MM
    AM/PM"}

Now analyze the provided schedule images and return the optimal meeting time.
```

Prompt: Meeting planning

```
You are a meeting planning assistant. Analyze participant schedules and meeting
    place options to select the optimal meeting arrangement.

CRITICAL INSTRUCTIONS:
1. READ THE IMAGES CAREFULLY - Extract participant schedules and meeting place
    details from the images
2. ONLY SELECT FROM AVAILABLE OPTIONS - Choose a meeting place and time that works
    for all participants
3. CONSTRAINTS MATTER - Consider travel distances, schedules, and availability
4. OPTIMAL = LOWEST COST - Among all feasible options, select the one with the
    lowest total cost
5. OUTPUT FORMAT - Return ONLY a JSON object with "meeting_place", "start_time",
    and "end_time" fields

REMEMBER:
- READ each image carefully for schedules and distances
- Use ONLY times and places shown in the images
- Copy times EXACTLY in "HH:MM AM/PM" format
```

```
- Return ONLY a JSON object: {"meeting_place": "Meeting Place X", "start_time":
    "HH:MM AM/PM", "end_time": "HH:MM AM/PM"}

Now analyze the provided images and return the optimal meeting arrangement.
```

B.2 GUIBENCH

GUIBench uses a single system prompt across all six task categories (*General Tool, Information Management, Media Entertainment, Multi-Apps, Social Sharing, Web Shopping*).

Prompt: GUIBench

```
You are a mobile device automation agent. You see a screenshot of a phone/tablet
    and must predict the NEXT action to perform.

AVAILABLE ACTIONS (use EXACTLY one):
- CLICK x y           - Tap at pixel coordinates (x, y)
- TEXT "some text"    - Type the given text into the focused field
- SCROLL x1 y1 x2 y2  - Swipe from (x1, y1) to (x2, y2)
- LONG_PRESS x y      - Long press at pixel coordinates (x, y)
- COMPLETE            - Task is finished, no more actions needed

RESPONSE FORMAT - return ONLY a JSON object, nothing else:
{"action": "CLICK", "x": 503, "y": 907}
{"action": "TEXT", "text": "chocolate chip cookies"}
{"action": "SCROLL", "x1": 576, "y1": 743, "x2": 569, "y2": 477}
{"action": "LONG_PRESS", "x": 200, "y": 300}
{"action": "COMPLETE"}

RULES:
1. Return ONLY the JSON object. No explanation, no markdown, no extra text.
2. Coordinates must be integers within the screen resolution.
3. Pick exactly one action per response.
4. If the task appears done, return {"action": "COMPLETE"}.
5. Do NOT guess coordinates. Look at the screenshot and identify the exact UI
    element to tap.
```

B.3 EMBODIEDBENCH

EmbodiedBench uses one system prompt defining the action vocabulary (11 primitives) and a JSON-array output schema. Every plan must end with NoOp, matching the ALFRED gold plans (Shridhar et al., 2020) that EmbodiedBench (Yang et al., 2025) builds on.

Prompt: EmbodiedBench

```
You are an embodied household-task planner. Given a high-level goal (delivered via
    text or speech) and the image of the current scene, predict the sequence of
    high-level actions the agent must take to achieve the goal.

Each action MUST have the form: ActionType ['arg1', 'arg2']
Available ActionType values: GotoLocation, PickupObject, PutObject, ToggleObject,
    HeatObject, CoolObject, CleanObject, SliceObject, OpenObject, CloseObject,
    NoOp.

Always end the plan with a NoOp.

Output ONLY a JSON array of action strings. Example output:
[
  "GotoLocation ['dresser']",
  "PickupObject ['cellphone']",
  "GotoLocation ['floorlamp']",
  "ToggleObject ['floorlamp']",
```

```
  "NoOp"
]

Do not output anything other than the JSON array.
```

B.4 WILD PARAPHRASER PROMPTS (INFORMAL AND MULTI-TURN)

The Informal and Multi-Turn conditions are produced by prompting GPT-4o with a system message that combines a base template (`DENSE_SYSTEM_TEMPLATE` for single-utterance Informal, or `SCATTER_SYSTEM_TEMPLATE` for Multi-Turn) with an informality tier instruction (`low`, `medium`, or `high`). The Informal condition uses the `high` tier on a single utterance. The Multi-Turn condition uses the `high` tier scattered across $N{=}2$ or $N{=}3$ turns.

Prompt: Informal

```
You are rewriting a list of planning requirements into a single natural-sounding
    spoken utterance.

Rephrase the requirements into a very informal, spontaneous-sounding spoken request
    - the way someone would actually talk, with frequent self-corrections ("wait
    no, I mean..."), hedging ("I guess", "maybe"), filler words ("um", "like"),
    and mid-sentence tangents that circle back. It should sound like real
    unscripted speech, not polished text.

Rules:
- Preserve ALL numeric values, city/place names, dates, times, and budget amounts
    EXACTLY as given.
- Do NOT add requirements that are not in the list.
- Do NOT drop any requirement - every listed constraint must appear.
- Output ONLY the rewritten request. No commentary, no prefixes, no quotation marks
    around the output.
```

Prompt: Multi-Turn

```
You are simulating a user who gives planning instructions across multiple
    conversation turns - the way someone would gradually reveal their needs in a
    real dialogue.

Rephrase the requirements into a very informal, spontaneous-sounding spoken request
    - the way someone would actually talk, with frequent self-corrections ("wait
    no, I mean..."), hedging ("I guess", "maybe"), filler words ("um", "like"),
    and mid-sentence tangents that circle back. It should sound like real
    unscripted speech, not polished text.

Additional rules:
- The first turn should set the overall topic/goal.
- Later turns add, refine, or elaborate on earlier details.
- Use natural turn-opening phrases that reference what came before ("oh and also",
    "actually", "one more thing", "I forgot to mention").
- Preserve ALL numeric values, city/place names, dates, times, and budget amounts
    EXACTLY as given.
- Do NOT add requirements that are not in the list.
- Do NOT drop any requirement.
- Output EXACTLY the requested number of turns.
- Format each turn on its own line as: T1: <text>
- Output ONLY the turns. No commentary.
```

Informality tier definitions. The three tiers control how much filler, hedging, and self-correction the paraphraser adds:

- **low.** Clear, slightly informal spoken request. Simple natural phrasing, minimal hedging or filler.

- **medium.** Moderately informal spoken request, as if chatting with a helpful assistant. Some hedging, informal connectives, occasional asides.
- **high.** Very informal, spontaneous-sounding speech. Frequent self-corrections, hedging, filler words, mid-sentence tangents.

B.5 MECHANISM-CONTROL PROMPTS (RECAP, WAIT, END-MARKER)

These prompts define the interventions in Table 6 (RECAP) and Table 8 (WAIT, END-MARKER). RECAP appends one instruction to the final turn of the default multi-turn protocol. WAIT replaces the system prompt with a turn-taking preamble (prepended to the same format and action-space specification) and closes with a final-turn instruction. END-MARKER keeps the default system prompt and only marks the final turn. All three reuse the same released audio and items as the Multi-Turn condition.

Prompt: Recap (final-turn instruction)

```
The task was given across several separate turns. Before answering, first combine
    every requirement mentioned across all turns into one complete task, then
    output the plan that satisfies all of them.
```

Prompt: Wait (system-prompt preamble)

```
The user will describe one household task across several separate turns. Do not
    plan or output any actions yet. After each turn, reply with only the single
    word OK. Only when the user says the task is complete should you output the
    plan.
```

Prompt: Wait (final-turn instruction)

```
That is everything. Now output ONLY the complete plan as a JSON array that
    satisfies every requirement mentioned across all turns.
```

Prompt: End-marker (final-turn instruction)

```
This is the final turn. Output ONLY the complete plan as a JSON array that
    satisfies every requirement mentioned across all turns.
```

C MORE RESULTS ANALYSIS

More turns deepen the drop. The concat-to-multi-turn drop grows from two to three turns for all four native models (Figure 7). On EmbodiedBench Qwen-Omni-3 falls from 56.2 to 32.1 Step-LCS at N=2 and from 55.0 to 22.1 at N=3, and Qwen-Omni-2.5 from 40.2 to 31.4 then 39.0 to 26.7 (Table 6). Spreading the same constraints over more turns makes them harder to aggregate.

D RECAP-AND-PLAN AND BIND-THEN-ACT

The main text shows that native omni models drop when spoken constraints are split across turns. Here we compare two training-free interventions on EmbodiedBench only.

Protocol. EmbodiedBench gold plans decompose into ordered sub-goals, so Step-LCS gives a fine-grained plan-quality signal. From the clean split we use all n=300 instances. We hold the visual context and gold plan fixed and vary only how the instruction is presented:

- CONCAT: all constraints in one spoken utterance, sent as a single audio clip in one user message.
- MULTI-TURN: the same constraints split across N dialogue turns. Each turn is a separate user message with one audio segment. The model responds at every turn, and we score the plan it emits at the final turn. The user side is still scripted. This is stricter than packing several clips into one message, but it is not an open interactive user.

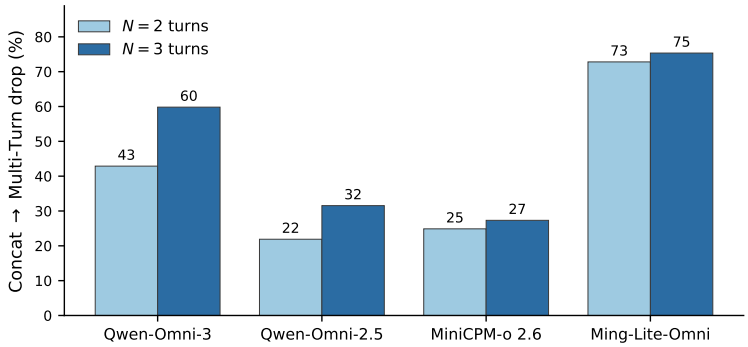


Figure 7: **More turns deepen the drop.** Relative Concat-to-Multi-Turn drop on EmbodiedBench (n=300) at N=2 and N=3, for four native omni models, from the controlled harness (Tables 6, 7). Every model drops more at three turns than at two.

Table 6: **Recap-and-Plan, Bind-then-Act, and Oracle Ledger** on EmbodiedBench (Step-LCS %, n=300). CONCAT is one utterance, MT a turn-by-turn dialogue with a response at each turn, RECAP restates then plans, BIND plans from a model-extracted ledger with no audio, and ORACLE from a gold ledger. Rec. is the Concat-to-MT gap restored by Recap, Cov. is ledger coverage of gold sub-goals. All cells come from one canonical harness (native decoding with the shared system prompt), so they are directly comparable with the controls in Table 7.

Model	N	CONCAT	MT	RECAP	Rec.	BIND	Cov.	ORACLE
Qwen-Omni-2.5	2	40.2	31.4	**40.4**	102%	22.4	74.5	43.2
	3	39.0	26.7	**38.3**	94%	16.4	52.7	43.2
Qwen-Omni-3	2	56.2	32.1	48.2	67%	**51.1**	96.9	58.1
	3	55.0	22.1	**45.5**	71%	45.4	92.8	58.1

- RECAP: identical to MULTI-TURN, except that the final turn appends one instruction, *"The task was given across several separate turns. Before answering, first combine every requirement mentioned across all turns into one complete task, then output the plan that satisfies all of them."*
- BIND: after each spoken turn the model may only update a running numbered constraint ledger and is forbidden from planning. After the last turn we start a fresh decode whose only inputs are the scene image and that ledger. The raw audio history is discarded. We also score *ledger coverage*: the fraction of gold sub-goal descriptions with at least 40% of their content words in the final ledger.

Because CONCAT and MULTI-TURN carry the same audio and differ only in turn structure, a drop here is hard to attribute to not hearing the words. RECAP and BIND then test two ways of putting constraints back together: a single shared decode versus an externalized ledger. All conditions use greedy decoding and the same Step-LCS scorer as the main experiments, on two native omni models, Qwen-Omni-2.5 (Qwen2.5-Omni-7B) and Qwen-Omni-3 (Qwen3-Omni-30B-A3B). This harness fixes a single prompt and item split, so its absolute levels differ from the official Table 2 and 4 cells. We read the numbers here as within-harness contrasts, not the released scores. Numbers are in Table 6.

The drop deepens with turns. Splitting identical constraints across turns lowers Step-LCS (Table 6). Qwen-Omni-2.5 falls from 40.2 under CONCAT to 31.4 at two turns (-8.8) and from 39.0 to 26.7 at three (-12.3). Qwen-Omni-3 falls from 56.2 to 32.1 at two turns (-24.1) and from 55.0 to 22.1 at three (-32.9). Both models handle the same constraints better when they arrive together. The drop deepens from two to three turns on both models, and is markedly larger at 30B.

A training-free prompt recovers a large fraction of the loss. RECAP restores Qwen-Omni-2.5 to 40.4 at two turns (102% of the gap) and to 38.3 at three turns (94%), essentially erasing the collapse. It restores Qwen-Omni-3 to 48.2 at two turns and 45.5 at three (67% and 71%), so at 30B Recap does

Table 7: 2×2 **(modality $\times$ delivery) and silent-accumulation controls** on EmbodiedBench (Step-LCS %, n=300, one harness) across four native omni models from three labs. Subscripts a/t are audio/text. SILENT$_a$ sends all audio turns with no model response between them. MT$_a$ falls below CONCAT$_a$ for every model.

Model	N	CONCAT$_a$	CONCAT$_t$	MT$_a$	MT$_t$	SILENT$_a$
Qwen-Omni-2.5	2	40.2	42.7	31.4	33.9	**42.0**
	3	39.0	42.9	26.7	28.0	**42.1**
Qwen-Omni-3	2	56.2	54.3	32.1	33.2	**55.6**
	3	55.0	53.0	22.1	25.7	**55.4**
MiniCPM-o 2.6	2	20.1	31.1	15.1	22.0	22.5
	3	19.4	32.7	14.1	20.4	25.4
Ming-Lite-Omni	2	35.3	35.1	9.6	19.2	17.8
	3	36.1	33.9	8.9	15.3	12.8

not fully close the gap. The same instruction helps a 7B dense model and a 30B mixture-of-experts model, with no training.

Bind-then-Act does not replace Recap. RECAP still lets restatement and planning share hidden state. BIND does not: the planner cannot attend to audio that was never written down. On Qwen-Omni-2.5 the ledger covers 74.5% of gold sub-goals at two turns and 52.7% at three, and Bind scores 22.4 and 16.4, below Multi-Turn. On Qwen-Omni-3 coverage is 96.9% and 92.8%. Two-turn Bind reaches 51.1, beating Recap (48.2) and 5.1 points below Concat (56.2). Three-turn Bind is 45.4 against Recap 45.5. A nearly complete ledger is not enough at three turns. At 7B, where the ledger is incomplete, Recap is stronger. At 30B, Bind edges Recap at two turns (51.1 vs 48.2) and ties at three (45.4 vs 45.5).

Takeaway. On this EmbodiedBench protocol the multi-turn drop is consistent with failing to carry constraints across turns. Recap-and-Plan recovers 67 to 102% of the Concat-to-Multi-Turn gap at 7B and 30B. Bind-then-Act matches or beats Recap only at 30B, where ledger coverage is high. 7B Bind underperforms Multi-Turn, 30B two-turn Bind is 51.1 against Concat 56.2, and 30B three-turn Bind ties Recap (45.4 vs 45.5).

E TEXT AND SILENT CONTROLS: A 2×2 DECOMPOSITION

To find out *why* the multi-turn drop happens, we cross modality (TEXT vs AUDIO) with delivery (CONCAT vs MULTI-TURN) on the same n=300 EmbodiedBench items, and add a SILENT accumulation condition. TEXT turns use the exact transcript of the released audio turns (the `wild_user_message_high` field the audio was synthesized from), so TEXT and AUDIO carry identical words. SILENT presents all turns as separate user messages, but the model does not respond until the final turn. This isolates the effect of the model's own intermediate responses. These reruns use the same single canonical harness as Table 6, so the CONCAT and MT cells match it exactly. We report paired, item-matched contrasts with 95% bootstrap CIs (B=2000). Table 7 gives the cells.

Modality has almost no effect, but intermediate responses have a large one. The Text $-$ Audio multi-turn contrast is $+2.5/+1.3$ at 7B and $+1.1/+3.6$ at 30B (two/three turns), all within a few points of zero. Splitting the instruction across turns hurts about equally whether it is heard or read, so the drop comes from aggregating turns, not from speech perception. The Silent $-$ Multi-Turn contrast is $+10.6\,[+7.9, +13.1]$ and $+15.4\,[+12.9, +17.9]$ at 7B and $+23.4\,[+20.5, +26.4]$ and $+33.3\,[+30.7, +36.0]$ at 30B. Every CI excludes zero, and Step-LCS returns to the single-utterance baseline. Removing the model's own intermediate plans removes the collapse. The failure is the model committing to a plan at each turn, not perception or sequence length alone. Bind pulls the same lever by deferring the plan, and Recap by restating every constraint before the final plan.

An explicit turn-taking prompt recovers most of the drop. If the collapse is the model committing at each turn, then a prompt that tells the model to hold its plan until the last turn should recover

Table 8: **Prompt-confound controls** on EmbodiedBench (Step-LCS %, n=300, same items and harness as Table 7). CONCAT and MT repeat Table 7. WAIT tells the model to acknowledge each turn and plan only on the last. END keeps the default prompt but flags the final turn. Rec. is the fraction of the CONCAT−MT gap restored.

Model	N	CONCAT	MT	WAIT	Rec.	END	Rec.
Qwen-Omni-2.5	2	40.2	31.4	**42.9**	131%	38.5	81%
	3	39.0	26.7	**41.9**	124%	31.8	41%
Qwen-Omni-3	2	56.2	32.1	**50.1**	75%	47.4	63%
	3	55.0	22.1	**48.9**	81%	43.3	64%

it without any change to the audio. We test two such prompts on the same items (Table 8). WAIT instructs the model to acknowledge each intermediate turn and plan only on the final turn. END-MARKER keeps the default prompt but flags the last turn explicitly. WAIT recovers the entire drop at 7B (124 to 131% of the concat-to-multi-turn gap) and 75 to 81% at 30B. That leaves a residual of about 6 Step-LCS at 30B (WAIT 50.1/48.9 vs CONCAT 56.2/55.0 at two/three turns). END-MARKER helps less (41 to 81%). Marking the final turn is not enough, and the model must be told to defer. Compliance depends on the model. Qwen-Omni-3 replies only OK on every intermediate turn (300/300 items). Qwen-Omni-2.5 does so on about 40% of turns and still over-recovers, so even partial deferral removes most of the collapse. The 30B residual persists under full compliance. The plan-every-turn protocol is thus the main cause of the collapse. A small aggregation residual survives at 30B even when premature commitment is removed.

The collapse replicates across four models. We ran the identical harness on two further native omni models from other labs, MiniCPM-o 2.6 from OpenBMB (OpenBMB, 2025; Yao et al., 2024) and Ming-Lite-Omni from inclusionAI (AI et al., 2025). The audio concat-to-multi-turn drop is present, with a CI excluding zero, for every model. It is $+5.0/+5.3$ for MiniCPM-o and $+25.8/+27.2$ for Ming at two and three turns, against $+8.8/+12.3$ for Qwen-Omni-2.5 and $+24.1/+32.9$ for Qwen-Omni-3. Multi-turn text collapses as well on both non-Qwen models ($+9.1$ to $+18.6$), so the failure is not audio-specific in any of the four. Two differences set the non-Qwen models apart. Their modality gap is larger, $+6.3$ to $+9.6$ Step-LCS with the CI excluding zero (text multi-turn above audio multi-turn). A modest audio penalty therefore adds to the aggregation failure, where Qwen shows a near-zero gap. Silent accumulation lifts MiniCPM-o above its Concat score, but on Ming it recovers a third of the drop or less, against a near-full recovery on Qwen. Incomplete aggregation across turns is the main driver on all four models, with a smaller extra audio cost on MiniCPM-o and Ming.

The Recap-and-Plan fix generalizes. We run the restate-then-plan prompt in the same harness (RECAP, identical to multi-turn with one added final-turn instruction). It recovers the audio collapse fully on Qwen-Omni-2.5 (102 and 94% of the concat-to-multi-turn gap at two and three turns) and on MiniCPM-o 2.6 (180 and 240%, above the single-utterance ceiling). It recovers most of it on Qwen-Omni-3 (67 and 71%). On the text side it recovers 61 to 106% across these three models. On Ming-Lite-Omni it recovers the text collapse (76 to 92%) but only part of the much deeper audio collapse (16 to 38%). There, audio multi-turn is about 9 Step-LCS against a 35 concat baseline. One training-free prompt therefore closes most of the collapse on three of the four models and the text collapse on all four. Ming's audio is the one case that a single recap only partly repairs.

The collapse is robust to accent and to a child voice. We resynthesize the N=2 EmbodiedBench items at n=300 with edge-tts in six English accents (British, Indian, Nigerian, Singaporean, Filipino, Australian) and a child voice. The visual scene, gold plan, and turn structure stay fixed, and we measure the concat-to-multi-turn collapse within each voice for all four native omni models (Table 9). The collapse is significant, with the 95% paired bootstrap CI excluding zero, in all 28 model-by-voice cells. Its size barely changes with the voice. Qwen-Omni-3 spans only $+21.8$ to $+23.5$ across the seven voices, and Ming-Lite-Omni $+22.0$ to $+25.1$. The per-model means are $+22.6$ (Qwen-Omni-3), $+23.1$ (Ming-Lite-Omni), $+9.3$ (Qwen-Omni-2.5), and $+6.2$ (MiniCPM-o), each matching that model's released-voice baseline. The collapse is therefore not an artifact of one synthesized voice, accent, or speaker age. This ablation does not change the released 11,700-sample benchmark.

Table 9: **Accent and child-voice robustness** of the concat-to-multi-turn collapse on EmbodiedBench (Step-LCS %, N=2, n=300, edge-tts voices), for four native omni models. Each cell is the paired item-matched collapse (CONCAT−MT) with a 95% bootstrap CI (B=2000). The collapse is significant (CI excludes zero) in all 28 cells.

Voice	Qwen-Omni-3 (30B)	Qwen-Omni-2.5 (7B)	MiniCPM-o 2.6	Ming-Lite-Omni
British	+22.9 [+20.1, +25.8]	+8.4 [+5.7, +11.1]	+5.5 [+3.0, +8.3]	+23.3 [+20.8, +26.1]
Indian	+22.3 [+19.5, +24.9]	+12.4 [+9.6, +15.3]	+7.4 [+4.5, +10.4]	+22.2 [+19.8, +24.8]
Nigerian	+21.9 [+19.1, +24.6]	+8.5 [+6.0, +11.1]	+5.7 [+2.8, +8.5]	+22.2 [+19.6, +24.8]
Singaporean	+21.8 [+19.1, +24.7]	+8.0 [+5.4, +10.8]	+7.8 [+5.0, +10.7]	+22.0 [+19.3, +24.6]
Filipino	+22.2 [+19.3, +25.2]	+9.8 [+7.2, +12.3]	+3.6 [+0.7, +6.4]	+25.1 [+22.5, +27.8]
Australian	+23.5 [+20.6, +26.3]	+7.7 [+4.9, +10.4]	+8.5 [+5.6, +11.3]	+23.6 [+21.1, +26.4]
Child	+23.5 [+20.7, +26.4]	+10.4 [+7.7, +13.2]	+4.8 [+2.2, +7.4]	+23.3 [+20.6, +26.0]
Mean	+22.6	+9.3	+6.2	+23.1

Table 10: **GUI multi-turn controls** on GUIBench (AMS %, n=300 paired steps, one harness). Subscripts a/t are audio/text. SILENT$_a$ sends all audio turns with no model reply between them. Contrasts (lower block) are paired item-matched with 95% bootstrap CIs (B=2000), and all four exclude zero for both models.

Model	CONCAT$_a$	MT$_a$	SILENT$_a$	RECAP$_a$	CONCAT$_t$	MT$_t$
Qwen-Omni-3 (30B)	61.0	18.6	60.8	23.3	62.4	24.5
Qwen-Omni-2.5 (7B)	29.4	16.0	27.1	20.3	35.4	21.7

Contrast (audio)	Qwen-Omni-3 (30B)	Qwen-Omni-2.5 (7B)
Collapse (CONCAT−MT)	+42.4 [+35.7, +48.9]	+13.4 [+8.6, +18.1]
Silent recovery (SILENT−MT)	+42.2 [+35.8, +48.4]	+11.2 [+6.2, +16.0]
Recap recovery (RECAP−MT)	+4.7 [+1.8, +7.8]	+4.3 [+1.5, +7.3]
Text collapse (CONCAT−MT)	+37.9 [+31.1, +44.8]	+13.8 [+8.3, +19.0]

F GUI CONTROLS: THE MECHANISM REPLICATES ON THE HEADLINE TASK

Our mechanism experiments so far use EmbodiedBench, where Step-LCS gives a graded plan-quality signal. Because the paper's headline drop is on GUI navigation, we replicate the same controls there. We take 100 GUIBench episodes and expand each into its first three steps (n=300 paired items). The screenshot and gold action stay fixed, and only the turn structure varies, as in Appendix E. The conditions are CONCAT (one utterance), MULTI-TURN (a reply at each turn), SILENT (all turns, no reply until the last), RECAP, and a TEXT mirror of each. We score with the Action Matching Score (AMS) of the main GUI experiments and report paired bootstrap CIs (B=2000). Table 10 gives the cells and contrasts.

The collapse and its silent recovery replicate on GUI. The audio collapse is large and its CI excludes zero for both models (+42.4 at 30B, +13.4 at 7B). Presenting the same turns as TEXT collapses almost as much (+37.9 at 30B, +13.8 at 7B), so on GUI too the drop is turn structure, not perception. Silent accumulation recovers it essentially in full at 30B (+42.2 of the +42.4 collapse, returning AMS from 18.6 to 60.8) and most of it at 7B (+11.2 of +13.4). The mechanism matches EmbodiedBench. The model commits to an action at each turn, and removing that intermediate commitment removes the collapse. The bootstrap here treats steps as the unit. Clustering by episode (100 episodes) shifts each interval by at most about a point (the 30B collapse becomes [+34.4, +49.7]). The contrasts are therefore not an artifact of non-independent steps.

One difference from EmbodiedBench: Recap transfers poorly. On GUI the one-shot RECAP prompt recovers only +4.7 at 30B (11% of the collapse) and +4.3 at 7B, far below the 67 to 102% it recovers on EmbodiedBench. Silent accumulation, which imposes the accumulate-then-act structure directly, still works, but a single restate-then-act instruction does not. The confound diagnosis is robust across tasks, though the one-shot textual fix is not.

The failures are premature commitment, not format errors. A failure taxonomy on the audio cells confirms the mechanism. Moving from CONCAT to MULTI-TURN at 30B, format-parse errors

Table 11: **Matched-model ASR+LLM cascade.** Each Qwen-Omni model is run three ways on the same items: native multi-turn audio, its own ASR+LLM cascade (Whisper transcribes the audio and the same model plans from the concatenated transcript), and a perfect-transcript upper bound (gold text). n=300 Embodied items and 100 GUI episodes, all $300/300$ scored. The cascade recovers the native multi-turn collapse on every cell.

Model	Task (metric)	Native audio (MT)	ASR+LLM cascade	Perfect transcript
Qwen-Omni-2.5	Embodied (Step-LCS)	31.4	**42.3**	42.7
Qwen-Omni-3	Embodied (Step-LCS)	32.1	**53.7**	54.3
Qwen-Omni-2.5	GUI (AMS)	16.0	**36.6**	35.4
Qwen-Omni-3	GUI (AMS)	18.6	**62.6**	62.4

fall from 29 to 7 of 300, while premature-COMPLETE actions rise from 24 to 88 and wrong-action-type errors from 18 to 66. The model is not failing to produce valid actions under multi-turn. It is committing to (often terminal) actions before it has heard the full instruction. This is the GUI analogue of the premature-planning failure on EmbodiedBench.

The same model as its own ASR+LLM cascade does not collapse. One might ask whether native omni models collapse because of the architecture or simply because they are weaker reasoners than the VL models used in the cascades. We remove that confound by running each Qwen-Omni model as its own ASR+LLM cascade. Whisper transcribes the multi-turn audio, and the same omni model plans from the concatenated transcript as text (Table 11). The same model that collapses on native multi-turn audio recovers almost completely as a cascade, on both tasks and both sizes. Qwen-Omni-3 goes from 32.1 to 53.7 Step-LCS on Embodied and from 18.6 to 62.6 AMS on GUI, within about one point of feeding it the gold transcript. The deficit therefore lies in the turn structure, not in the model's ability to plan. It is also not specific to the VL reasoners of the main cascades. These cascade scores nearly reproduce the concat and text-concat cells of Tables 7 and 10. This is expected, since a cascade joins the transcript into one turn. The self-cascade rows of Table 4 are a separate run in which the omni model also transcribes the audio itself, and they recover by a similar amount (62.7 AMS and 55.6 Step-LCS for Qwen-Omni-3).

G INFORMAL SPEECH: A LENGTH-MATCHED CONTROL

The Informal condition replaces the clean instruction with a single disfluent spoken utterance carrying filler words, self-corrections, and mild rambling (Appendix B). Because disfluency adds words, informal utterances are longer than their clean counterparts. This raises a question about the Informal results: is any change caused by the informal style itself, or simply by the greater length of the audio? We run a controlled experiment that separates the two.

On the 50 EmbodiedBench instances of Appendix I we synthesize three versions of each instruction with the same neutral voice, so that voice and engine are held constant and only the wording varies. *Clean* is the original instruction (mean 8.9 words, 2.5 s of audio). *Informal* is the disfluent paraphrase (mean 25.1 words, 9.1 s). *Control* is a fluent, formal paraphrase that preserves every constraint but contains no filler and no disfluency, expanded to a comparable length (mean 21.2 words, 7.2 s). Both long conditions are about three times the clean duration. Clean versus Control isolates length, and Control versus Informal isolates informal style at a comparable length. One caveat applies. This control's Informal paraphrase (mean 25.1 words) is a separate, shorter one-shot rewrite. The released Informal condition is far longer (mean about 119 words on EmbodiedBench, Figure 2) because it comes from a denser expansion pipeline. This control therefore bounds the length effect only for short utterances and does not length-match the released condition. The lift of that condition in Table 4 is better read as redundant re-wording of a 9-word goal than as slang helping planning.

Table 12 reports two native omni models as Step-LCS, and the pattern is the same for both. Moving from Clean to the length-matched Control barely changes the score (Qwen-Omni-2.5 $+0.010$, Qwen-Omni-3 -0.008), so the added length alone does nothing. Moving from Control to Informal is likewise small and, if anything, positive ($+0.053$ and $+0.060$), so the informal style does not hurt either. Because the length step is essentially zero for both models, the Informal behavior cannot be a hidden length effect. The order-invariant multiset F1 of Appendix K gives the same ordering, and the

Condition	Words	Dur. (s)	Step-LCS	Δ vs Clean	Δ vs Control
Qwen-Omni-2.5 (7B)					
Clean	8.9	2.5	0.417		
Control (formal)	21.2	7.2	0.427	+0.010	
Informal	25.1	9.1	0.480	+0.063	+0.053
Qwen-Omni-3 (30B)					
Clean	8.9	2.5	0.342		
Control (formal)	21.2	7.2	0.333	−0.008	
Informal	25.1	9.1	0.393	+0.052	+0.060

Table 12: Length-matched informal control on EmbodiedBench (Step-LCS, 0 to 1 scale). Clean to Control isolates length, and Control to Informal isolates informal style at a comparable length.

	7B	30B
Clean Step-LCS	35.0	38.5
Interrupt Step-LCS	29.8	29.5
Δ (interrupt − clean)	−5.2	−9.0

Table 13: Concrete spoken interruption (20 templates) on EmbodiedBench, n=50. Scores are Step-LCS ($\times$100). Each interrupt clip keeps the same constraints as clean and splices in one off-topic second-speaker line with an answer and resume. Decoding uses the official Qwen speech chat-template prefix.

7B clean score (0.417) matches its ablation clean (0.410). We conclude that the Informal condition does not degrade either model on EmbodiedBench. The small movements that remain reflect the coarse-metric noise noted for the wild conditions, not audio length.

H Natural Interruptions as a Side Sequence

The Competing Speech condition in the main tables (the folder released as `noise_audio/interruption/`) overlays other-task speech on a continuous user utterance. That is competing-speaker noise, not a conversation break, so we call it competing-speech overlay. Here we use a more concrete interruption protocol. A second speaker cuts in with an off-topic line, the user answers, and then resumes the same instruction. The side sequence is spliced, not overlaid, so no constraint is deleted. The extra work is to keep the task across the break. Length is not matched (interrupt clips are longer), and the 250 ms gaps are sequential turns rather than overlapping barge-in.

We use 20 spoken templates, cycled over the same 50 EmbodiedBench instances as the informal-length control. Every template is a person speaking, e.g. "Excuse me, your coffee is ready at the counter," followed by a short answer and a resume cue. There is no self-repair or sound-effect type. The user voice is fixed. The interrupter is a second TTS voice. Evaluation uses the same Qwen speech-mode identity prefix as the main tables. The informal-length control and the speaker ablation use the plain system prompt without this prefix, so their clean baselines on the same 50 instances differ from the ones here, and each comparison stays within one prompt setting.

The interrupted instruction stays intelligible. Whisper recovers the full task on the interrupted clips. A coffee barge-in renders as "Pick up laptop. Excuse me, your coffee is ready at the counter. Oh, thanks, I'll grab it in a second. Sorry about that. Okay, where was I. And turn on lamp": the aside is audible and the original constraints are still in the transcript.

Effect on planning. Table 13 is a matched clean vs. interrupt contrast on those 50 instances. Both models lose Step-LCS (7B −5.2, 30B −9.0). It does *not* reuse the main-table competing-speech overlay numbers and does not show that a break without a second speaker is harmless. On instances the model can already plan (clean Step-LCS ≥ 0.5, selected on an independent clean run to avoid regression to the mean), the same clips cost about 17 to 18 points (Appendix J).

Axis	Mean WER
Gender	1.6%
Rate	1.9%
Emotion	1.9%
Accent	2.5%
Age	2.5%
All (1,440 clips)	2.2%

Table 14: Intelligibility of the ablation audio: ASR word error rate by axis.

Axis	Condition	7B Δ	30B Δ
Accent (vs American)	Indian	-2.2	$+2.5$
	Nigerian	-6.0	-1.4
	Irish	-4.4	$+1.5$
Rate	slow	$+1.8$	$+0.2$
	fast	$+1.2$	$+0.8$
Emotion	hurried	$+0.5$	$+2.2$
	angry	$+1.8$	$+1.7$
	whispering	$+3.3$	-0.3
Age	child	$+1.0$	-1.3
	elderly	$+1.5$	-1.5
Gender	male	$+0.8$	$+2.0$

Table 15: Speaker-variation deltas (change in Step-LCS, points) for both native omni models on EmbodiedBench. Accent is compared to the same-engine American voice, the other axes to the main clean audio. Clean Step-LCS is 41.0 (7B) and 35.5 (30B). The American-accent baseline is 45.7 (7B) and 37.2 (30B).

I Speaker and Speech Variation (Ablation)

The main benchmark holds the speaker fixed and varies the acoustic and linguistic condition. This ablation does the opposite. It holds the instruction content and the clean condition fixed and changes one speaker or speech attribute at a time. Any change in a model score can then be attributed to that attribute. We vary five attributes on a fixed subset of the data: accent (American, Indian, Nigerian, Irish), speaking rate (slow, fast), emotion (hurried, angry, whispering), apparent age (child, elderly), and gender (male). Accent variants use per-locale neural voices (edge-tts), with American rendered on the same engine so the accent axis is internally consistent. Rate, emotion, age, and gender are set through the Qwen3-TTS instruction. Accent cannot be controlled reliably that way. The subset is 120 instances across all three benchmarks (50 EmbodiedBench, 50 GUIBench, 20 ScheduleBench), giving 1,440 clips at 12 conditions each.

Content is preserved across all variations. Before comparing model behavior we confirm the variation does not degrade intelligibility. Transcribing every clip with Whisper-large-v3 gives a mean word error rate of 2.2 percent, with no axis above 2.5 percent (Table 14). Even the hardest cases stay low (whispering 2.0, fast 1.8, child 2.7 percent). So a change in model score reflects the speaker, not missing words.

Planning is largely invariant to the speaker. Table 15 reports both native omni models on EmbodiedBench as the change in Step-LCS from the clean baseline for each condition. Every change is small for both models. Rate, emotion, age, and gender move performance by at most 3.3 points, within the noise of a graded metric on 50 instances. Accent, read against the same-engine American voice, is the only axis with a visible effect, and only for the smaller model. Qwen-Omni-2.5 (7B) loses up to 6 points (Nigerian -6.0, Irish -4.4, Indian -2.2), while Qwen-Omni-3 (30B) stays within ± 2.5 on every accent (Indian $+2.5$, Irish $+1.5$, Nigerian -1.4). Both models are therefore largely invariant to who is speaking and how, and the larger model is, if anything, more robust to accent.

Condition	Model	Full Δ	Qualified Δ
Interrupt	7B	-5.2	-20.1
	30B	-9.0	-20.7
Informal	7B	$+6.3$	-1.3
	30B	$+5.2$	-1.9

Table 16: Wild-condition effects on EmbodiedBench (change in Step-LCS, points) on the full 50 instances versus the baseline-qualified subset (clean Step-LCS ≥ 0.5, $n \approx 22$ to 27). Qualification removes low-baseline dilution. The Qualified Δ column selects on the same run. For interruption, selecting on an independent clean run corrects for regression to the mean and gives -17.0/-18.4 (see text).

J Baseline Qualification Sharpens the Wild-Condition Effects

EmbodiedBench clean scores are low and Step-LCS is graded. Instances the model cannot plan even on clean input pull every condition's delta toward zero and mask the real effect of a perturbation. As a robustness check we re-score the natural speech conditions on the instances whose clean Step-LCS is at least 0.5. A delta then reflects a change on tasks the model can otherwise do.

Table 16 reports the effect. The concrete-interruption drop (Appendix H) is large on qualified instances. Selecting the subset on the *same* run that we then measure inflates it through regression to the mean, so we select on an *independent* clean run instead. That gives -17.0 points for Qwen-Omni-2.5 (n=24, 95% CI $[-26.7, -6.6]$) and -18.4 for Qwen-Omni-3 (n=24, $[-28.1, -10.1]$). These are about 2 to 3 points smaller than the same-run values of -20.1 and -20.7 in Table 16, but still large and excluding zero. The informal effect moves the other way. Its small apparent gain on the full set (Appendix G) disappears under qualification (7B -1.3, 30B -1.9). Informal phrasing therefore neither helps nor hurts, and its small full-set gain does not survive qualification. The mild accent sensitivity of the 7B also sharpens (Nigerian -8.7 on qualified instances). Apart from this speaker effect, interruption is the one condition that degrades planning under this qualified view.

K Robustness of Step-LCS to Step Ordering

Our EmbodiedBench planning metric, Step-LCS (`plan_step_full_lcs_acc`), is the length-normalized longest common subsequence between the predicted and gold step sequences, so it rewards steps that appear in the gold order. A fair question is whether our conclusions come from penalizing re-ordering rather than from genuine planning differences. We therefore re-score the exact same saved predictions with an order-invariant metric and check whether the picture changes.

The order-invariant metric is a multiset F1 over parsed `(action, arguments)` steps. For a gold plan g and a prediction p, the overlap is the size of the multiset intersection of their steps, precision is the overlap over $|p|$, recall is the overlap over $|g|$, and the score is their harmonic mean. Order is ignored entirely, so a step present in both plans counts regardless of position.

We re-score two models with no additional inference: Qwen-Omni-2.5 on the clean audio and the twelve speaker-variation conditions of Appendix I (650 plans), and InternOmni on the EmbodiedBench clean-speech and text conditions (600 plans). The re-score parses every plan with one shared parser that drops the terminal NoOp step, so its absolute values sit below the main tables (InternOmni 14.1 and 26.6 here against 24.2 and 32.4 in Table 2), but both metrics are computed on the same parsed plans. Across all 1,250 scored plans the two metrics agree closely, with a Pearson correlation of 0.95 and a Spearman rank correlation of 0.96. Condition means differ by at most 0.06 (Table 17). Among the 55 condition pairs whose Step-LCS gap exceeds a 0.02 noise floor, the order-invariant metric keeps the sign of the gap in 53. It is slightly higher on average, as expected for a metric that re-ordering cannot hurt. It never reverses a substantive comparison, including the text-over-speech advantage for InternOmni. Step-LCS therefore tracks planning quality rather than an ordering artifact.

Model	Condition	Step-LCS	Multiset-F1
InternOmni	Text	0.266	0.325
InternOmni	Clean speech	0.141	0.166
Qwen-Omni-2.5	Clean	0.410	0.412
Qwen-Omni-2.5	Accent (Nig.)	0.397	0.425
All 1,250 plans: Pearson $r = 0.95$, Spearman $\rho = 0.96$			

Table 17: Step-LCS versus an order-invariant multiset F1 on the same saved predictions. The two metrics are strongly correlated and preserve every substantive comparison.

L Ablation Studies of ASR models

The cascade pipelines in our main results (Table 2) all use **GPT-4o-transcribe** as the ASR frontend. We picked it for reproducibility and a controlled comparison, not because we claim it is the best ASR. To check that our conclusions do not depend on this choice, we re-ran the cascade with three alternative ASR backends.

We compare four ASRs. **GPT-4o-transcribe** is a proprietary multilingual model trained on conversational speech, used as the frontend in the main results. **Whisper-large-v3** (Radford et al., 2023) is an open-weight model widely used in academic work for its zero-shot performance across accents and noise. **Whisper-medium** is a smaller-capacity variant with faster inference. **Azure Speech-to-Text** is a commercial cloud API. Everything else is held constant: the same LLM models (Qwen3-VL-8B-Instruct and GPT-4o), the same prompt templates (Appendix B), and greedy decoding.

Figure 8 reports the result across all three benchmarks under clean audio. The metric per benchmark matches the main table: **AMS** (Action Matching Score) on GUIBench, **FPS** (Feasible Plan Score) on ScheduleBench, and **Step-LCS** on EmbodiedBench. Each marker is one alternative ASR's score deviation from GPT-4o-transcribe in percentage points, for each benchmark crossed with each of the two LLM backbones.

Across every benchmark and every LLM combination, the three alternative ASRs fall within ± 1.5 percentage points of GPT-4o-transcribe. By comparison, switching the LLM model between Qwen3-VL and GPT-4o changes each metric by 3 to 23 points (Table 2), an order of magnitude more than swapping ASRs. With N=300 per cell, the bootstrap 95% confidence interval on each score is roughly ± 2 to ± 3 points. The deviations in the figure are at or below this sampling-noise floor. We therefore use GPT-4o-transcribe as the single ASR frontend in the main results.

M Human Validation of the Wild-Condition Data (Ablation)

As an ablation on data quality, we ran a short human study confirming that our generated wild text and audio still convey the original task. Annotators reviewed four conditions: multi-turn text (Stage A), multi-turn audio (Stage B), informal text (Stage C), and noisy audio (Stage D). For the text stages they marked each constraint as expressed (YES, PARTIAL, or NO) and rated naturalness. For the audio stages they reported whether they could make out the full instruction and rated clarity on a 1 to 5 scale. Stages A and B cover 150 tasks each (50 per benchmark, on the same 150 tasks) and Stages C and D cover 90 each, drawn from the released wild conditions and hosted on Label Studio. After cleaning we collected about 1,500 ratings from 20 paid annotators, a median of 3 per task. Text stages report soft recall (YES=1, PARTIAL=0.5) and audio stages report make-out rate and mean clarity, all with 95% bootstrap confidence intervals (B=2000).

Results. The generated data preserves the task (Table 18, Figure 9). Text rewrites keep constraints with soft recall between 0.965 and 0.996. Clean multi-turn audio is highly intelligible, with make-out between 95% and 99% and clarity of at least 4.86, so the TTS does not degrade the rewrite. Noise lowers intelligibility, with make-out between 77% and 89% and clarity of 4.36 or higher. The Step 0 coverage check flags EmbodiedBench under noise (0.69), where the trimmed constraint list sometimes omits source content, and we read those cells with that caveat. Appendix N tests whether the results hold on human-recorded speech.

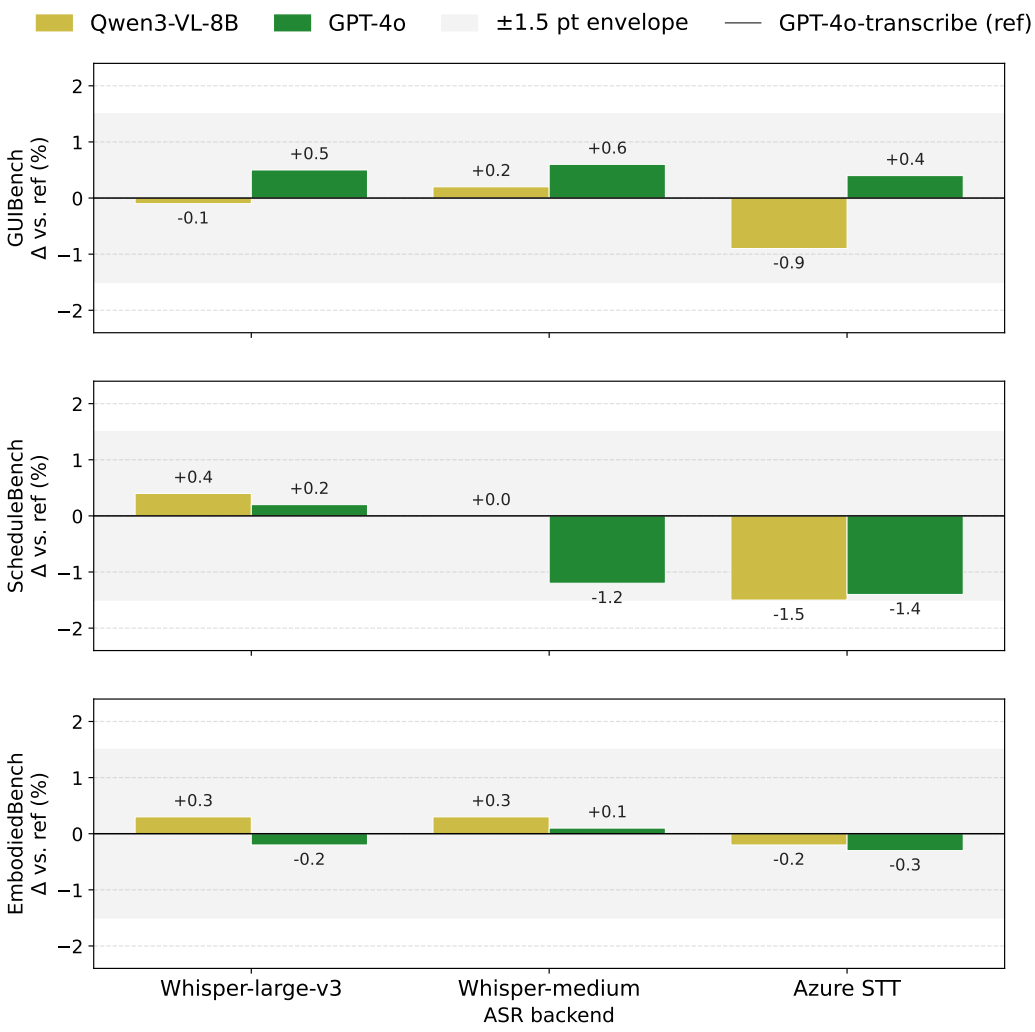


Figure 8: **ASR ablation across the three benchmarks (clean audio, N=300 per cell).** Each marker is one alternative ASR's deviation from the GPT-4o-transcribe reference (vertical gray line at 0), in percentage points. Metrics match Table 2: AMS on GUIBench, FPS on ScheduleBench, Step-LCS on EmbodiedBench. All 18 cells sit inside the ±1.5 percentage-point envelope. With N=300, bootstrap 95% confidence intervals on each cell are roughly ±2 to ±3 points, so the cell-level Δ values shown here are at or below the sampling-noise floor.

N Human-Speech Validation Study

Design. We ask whether the conclusions drawn from TTS audio hold when real people speak the instructions. Three studies record human speech for items that already have TTS audio. The single-utterance study covers 90 clean instructions, 30 per benchmark. The two-turn study covers 60 requests, 20 per benchmark, each split across two turns in the informal wording of the Multi-Turn condition. The three-turn study covers another 60 requests, split across three turns in the same wording. Every human recording is scored against the TTS audio of the same item, with the same pipeline, prompts, greedy decoding, and scorers. Gaps are therefore item-matched, and the speaker effect is not confounded with the item sample. The headline N=300 results in the main text use a larger item set, so their numbers are not comparable to the ones below.

Participants. Speakers were recruited online from a crowdsourcing pool. They had to be adults, report fluent English, and use a desktop computer with a microphone. People who had taken part in an earlier study of this project could not join a later one, and no speaker appears in more than one

Table 18: Human validation across four wild conditions. Recall/Make-out: text stages report soft recall (YES=1, PARTIAL=0.5). Audio stages report the proportion of annotators who heard the full instruction. Coverage is the Step 0 check (annotators who agreed the numbered constraint list covers the original instruction), and is not collected on Stages A and C. Fleiss κ on the binary expressed-or-not call. Ceiling effects deflate κ when answers cluster at YES. "—" marks cells where the metric does not apply or where κ is undefined under single-annotator coverage. CIs are 95% bootstrap (B=2000).

Stage	Benchmark	n_{ann}	Recall / Make-out	Coverage	Audio clarity	Fleiss κ
A. MT text	ScheduleBench	151	0.965 [0.935, 0.986]	—	—	−0.087
	GUIBench	151	0.996 [0.990, 1.000]	—	—	1.000
	EmbodiedBench	150	0.989 [0.980, 0.996]	—	—	1.000
B. MT audio	ScheduleBench	155	0.95 [0.92, 0.98]	0.91 [0.86, 0.95]	4.86 [4.8, 4.9]	—
	GUIBench	173	0.99 [0.97, 1.00]	0.98 [0.96, 1.00]	4.92 [4.9, 5.0]	—
	EmbodiedBench	142	0.99 [0.96, 1.00]	0.90 [0.85, 0.95]	4.98 [4.9, 5.0]	—
C. Inf text	ScheduleBench	90	0.996 [0.991, 1.000]	—	—	0.000
	GUIBench	91	0.984 [0.964, 1.000]	—	—	0.324
	EmbodiedBench	93	0.990 [0.978, 1.000]	—	—	1.000
D. Noisy audio	ScheduleBench	94	0.89 [0.83, 0.95]	0.84 [0.77, 0.90]	4.61 [4.5, 4.7]	—
	GUIBench	91	0.78 [0.69, 0.86]	0.96 [0.91, 0.99]	4.44 [4.3, 4.6]	—
	EmbodiedBench	90	0.77 [0.68, 0.86]	**0.69 [0.59, 0.79]**	4.36 [4.2, 4.5]	—

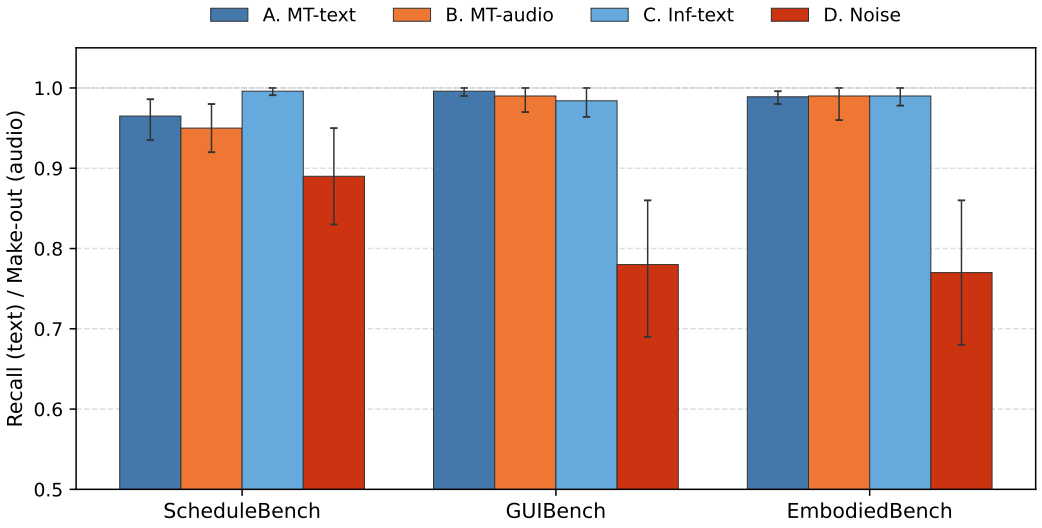


Figure 9: Per-stage, per-benchmark recall (text) and make-out (audio) with 95% bootstrap CIs. Text rewrites and clean multi-turn audio cluster near ceiling.

study. Researchers and team members, who wrote the stimuli and know the hypotheses, are excluded from every result. Speakers who withdrew before submitting are excluded as well. In the two-turn study we excluded three further submissions before looking at any model output. Two speakers had their browser translate the page and read the requests in Portuguese, and one sent six silent recordings. In the three-turn study one speaker sent a single recording, so none of their requests is complete. Three other speakers read the recorder's instruction line aloud before each turn. We cut that line from their recordings at the word boundary found by Whisper and transcribed every cut recording again to confirm that only the turn remains. This leaves 26 speakers in the single-utterance study, 43 in the two-turn study, and 47 in the three-turn study, 116 in total. Four of the 43 come from a pilot of the two-turn study that used a clean rather than informal wording and covered ScheduleBench and EmbodiedBench only. Four of the 47 come from a pilot of the three-turn study that used the same wording and recorder as the main run. Table 19 describes the 115 speakers with a profile. The sample

covers 26 first languages and 25 countries of residence, and 75 of the 115 are non-native English speakers.

Table 19: Speakers with a profile in each study. Profile fields are self-reported and are never given to a model.

	Single-utterance study	**Two-turn study**	**Three-turn study**
Speakers with a profile	25	43	47
Non-native English	16	32	27
First languages	10: English 9, Spanish 4, Portuguese 4, Tamil 2, Kannada 1, Latvian 1, remaining 4	15: English 11, Portuguese 6, Arabic 6, Spanish 4, Vietnamese 3, Hindi 2, remaining 11	16: English 20, Portuguese 5, Telugu 3, Swahili 2, Spanish 2, Arabic 2, remaining 13
Countries of residence	13: South Africa 5, Mexico 4, India 3, Portugal 2, Brazil 2, remaining 9	15: Brazil 6, Canada 6, Egypt 6, Chile 4, South Africa 4, remaining 17	18: United Kingdom 8, India 8, South Africa 7, Brazil 4, Poland 4, remaining 16
Ethnicity	White 9, Black 7, Asian 4, Other 3, Mixed 2	White 18, Asian 10, Black 7, Mixed 6, Other 2	White 16, Black 16, Asian 12, Prefer not to say 2, Other 1
Gender	18 male, 7 female	25 male, 18 female	28 male, 19 female
Age	20 to 64 (median 29)	22 to 67 (median 33)	19 to 62 (median 31)

Stimuli. The single-utterance items are clean instructions sampled across the task types of each benchmark. ScheduleBench is templated. Its 700-instruction pool reduces to nine sentence frames once numbers, dates, and place names are removed, and our 30 items cover eight of them. Its effective item count is lower than 30, so its intervals are optimistic. GUIBench and EmbodiedBench have one frame per item. The multi-turn requests use the exact turn texts from which the released Multi-Turn audio was synthesized. Transcribing that TTS audio recovers all of the content words of the median turn and at least 78% of every turn, so the human and TTS turns carry the same words. Requests are dealt across task types so that every type of each benchmark appears.

Procedure. Speakers use a browser recorder. It shows a consent notice, collects a short profile (first language, gender, age band, English proficiency), and then presents one item at a time with record, playback, and re-record. The recording server hands out blocks of items in sequence, so successive speakers tile the item set. In the single-utterance study each speaker reads nine items, three per benchmark, interleaved so the long ScheduleBench paragraphs are spread through the session. Every item is read one to four times, and 81 of the 90 items at least twice. In the two-turn and three-turn studies each speaker records three requests, one per benchmark, and says the turns of a request one after the other. One two-turn speaker recorded two GUIBench requests. Speakers are asked to read each turn exactly as written, fillers included. A two-turn session takes about seven minutes. Speaker identifiers are stored only as one-way hashes.

Quality control. Every recording is decoded, checked for silence and clipping, and converted to 16 kHz mono. We check reading faithfulness with an ASR whose language is fixed to English, as the share of each stimulus's content words found in the transcript. The median recording keeps all of them. Of the 231 single-utterance recordings, 224 keep at least half, and so do 259 of the 272 two-turn recordings. We re-transcribed the two-turn cases below half with Whisper large-v3, which recovers at least 61% of the content words in every one of them, so they are faithful readings that the first ASR truncated. In the three-turn study 390 of the 410 recordings keep at least half, and Whisper recovers at least 56% of the content words in each of the other 20. We do not filter recordings on transcript quality, because an ASR-based filter removes accented speakers and makes human speech look more like TTS than it is. When a single-utterance item has several readings we evaluate one, chosen by a fixed hash of the recording id.

Analysis. Metrics are AMS on GUIBench, FPS on ScheduleBench, and Step-LCS on EmbodiedBench. For each model and benchmark we report $\Delta = \text{human} - \text{TTS}$ with a 95% paired bootstrap interval over items (B=2000). A single cell of 30 items resolves large differences only. Our main estimate therefore pools all cells. The pooled bootstrap resamples items within each benchmark and uses the same draw for every model, so shared item difficulty is not counted as independent evidence. Cells where a model scores below 10 on both voices are left out of the pool. At that floor both voices fail the same items, and the cell adds zeros that say nothing about the voice. We pre-specified an equivalence margin of ± 5 points, somewhat wider than the ± 3 point bootstrap uncertainty of a single cell at N=300, and test it with two one-sided tests, which amount to checking that the 90% interval

lies inside the margin. For the two multi-turn studies the unit is the request. The collapse is the concatenated score minus the turn-by-turn score. For a human reading, the concatenated clip joins that speaker's own turns with the 1.5 s gap of the released concat audio, so any word a speaker drops is dropped on both sides. GUI steps are averaged within an episode first. A request enters only when at least one speaker recorded all of its turns, so a benchmark can have fewer than 20 requests. We compare the human collapse with the TTS collapse on the same requests, and fit a linear mixed model on every reading, the pilot included, with random intercepts for speaker and request.

Clean speech. Table 20 gives every cell. Pooled over the 19 informative cells, human speech scores -0.7 points against TTS (95% CI $[-2.6, +1.1]$). The 90% interval $[-2.3, +0.9]$ lies inside the ± 5 margin, so the two voices are equivalent at this margin. 4 of the 24 cells exclude zero, all of them human penalties: Qwen-Omni-3 on GUIBench, gpt-4o-trans.+Qwen2.5-VL on EmbodiedBench, Qwen-Omni-2.5 (self-casc.) on EmbodiedBench and Qwen-Omni-3 (self-casc.) on EmbodiedBench. The largest is Qwen-Omni-3 on GUIBench (-13.2 AMS, 95% CI $[-26.6, -3.2]$). EmbodiedBench leans negative for every model, and all three of its significant cells are cascades that transcribe first. Its instructions are short, which may make a single misheard object name costly.

Table 20: Human versus TTS speech per model and benchmark, on the same items. Δ = human − TTS in points with a 95% paired bootstrap CI. † At the floor on both voices and left out of the pooled estimate.

Benchmark	Model	Text	TTS	Human	Δ [95% CI]
GUIBench	Qwen-Omni-2.5	32.1	32.3	32.1	−0.3 [−11.1, +10.0]
	Qwen-Omni-3	76.6	63.2	50.0	−13.2 [−26.6, −3.2]
	InternOmni†	5.8	9.3	9.3	0.0 [0.0, 0.0]
	gpt-4o-trans.+GPT-4o-mini†	6.7	6.7	8.9	+2.2 [0.0, +6.6]
	gpt-4o-trans.+GPT-4o	10.8	17.4	18.3	+0.9 [−6.1, +10.0]
	gpt-4o-trans.+Qwen2.5-VL	21.2	14.6	14.6	0.0 [0.0, 0.0]
	Qwen-Omni-2.5 (self-casc.)	28.9	40.9	38.5	−2.5 [−10.0, +2.6]
	Qwen-Omni-3 (self-casc.)	69.9	39.9	48.4	+8.5 [−1.5, +20.0]
ScheduleBench	Qwen-Omni-2.5†	6.7	5.0	5.0	0.0 [0.0, 0.0]
	Qwen-Omni-3	15.0	16.7	20.0	+3.3 [0.0, +8.3]
	InternOmni	13.3	8.3	11.7	+3.3 [0.0, +10.0]
	gpt-4o-trans.+GPT-4o-mini	20.0	26.7	28.3	+1.7 [−10.0, +13.3]
	gpt-4o-trans.+GPT-4o	46.7	35.0	38.3	+3.3 [−11.7, +18.3]
	gpt-4o-trans.+Qwen2.5-VL†	6.7	6.7	5.0	−1.7 [−5.0, 0.0]
	Qwen-Omni-2.5 (self-casc.)†	3.3	5.0	5.0	0.0 [0.0, 0.0]
	Qwen-Omni-3 (self-casc.)	13.3	16.7	21.7	+5.0 [−3.3, +13.3]
EmbodiedBench	Qwen-Omni-2.5	29.1	35.4	30.7	−4.7 [−13.1, +2.6]
	Qwen-Omni-3	38.9	37.8	37.3	−0.5 [−6.9, +6.2]
	InternOmni	25.4	10.5	9.8	−0.7 [−4.8, +3.2]
	gpt-4o-trans.+GPT-4o-mini	29.3	22.7	21.6	−1.1 [−5.0, +2.3]
	gpt-4o-trans.+GPT-4o	36.7	31.8	30.2	−1.6 [−6.8, +3.4]
	gpt-4o-trans.+Qwen2.5-VL	26.8	27.5	23.8	−3.6 [−6.9, −0.7]
	Qwen-Omni-2.5 (self-casc.)	26.0	29.6	25.4	−4.2 [−8.1, −0.8]
	Qwen-Omni-3 (self-casc.)	37.3	39.8	32.6	−7.2 [−13.6, −2.2]
Pooled over the 19 cells without †, 90% CI [−2.3, +0.9]					−0.7 [−2.6, +1.1]

Noise. We apply the six MUSAN environments of the main Noise condition to the human recordings with the recipe used for the TTS audio. That is 15 dB SNR on ScheduleBench and EmbodiedBench, the 5, 10, and 15 dB cycle on GUIBench, and babble built from clean TTS speech. We regenerate the noisy TTS with the same recipe on the same items, so a TTS clip and its human counterpart share the environment, the SNR, and the random seed. Measured SNRs match to within 0.1 dB. The quantity of interest is the interaction $(A^{\text{noisy}}_{\text{hum}} - A_{\text{hum}}) - (A^{\text{noisy}}_{\text{TTS}} - A_{\text{TTS}})$, averaged over the six environments. Pooled over 19 cells it is -2.6 points (95% CI $[-4.6, -0.7]$, Table 21). Noise costs human recordings slightly more than TTS, and the interval excludes zero. The benchmark's noisy TTS may therefore understate the noise penalty for real speakers by a few points.

Table 21: Clean versus noisy speech for each voice. Noise is the mean over the six environments. † At the floor on both voices and left out of the pooled estimate. The interaction is the human noise cost minus the TTS noise cost, with a 95% paired bootstrap CI. Negative means noise hurts human speech more.

		TTS		Human		
Benchmark	**Model**	Clean	Noise	Clean	Noise	**Interaction** [95% CI]
GUIBench	Qwen-Omni-2.5	32.3	35.8	32.1	32.1	-3.5 [-13.6, $+5.9$]
	Qwen-Omni-3	63.2	57.7	50.0	51.8	$+7.3$ [-2.5, $+18.7$]
	InternOmni†	9.3	9.6	9.3	9.3	-0.3 [-0.8, 0.0]
	gpt-4o-trans.+GPT-4o-mini†	6.7	9.0	8.9	8.8	-2.3 [-8.0, $+2.2$]
	gpt-4o-trans.+GPT-4o	17.4	23.2	18.3	16.8	-7.3 [-22.1, $+4.0$]
	gpt-4o-trans.+Qwen2.5-VL	14.6	15.7	14.6	9.8	-5.9 [-14.0, -0.5]
	Qwen-Omni-2.5 (self-casc.)	40.9	38.8	38.5	35.6	-0.7 [-6.6, $+5.8$]
	Qwen-Omni-3 (self-casc.)	39.9	46.0	48.4	47.7	-6.8 [-18.2, $+3.3$]
ScheduleBench	Qwen-Omni-2.5†	5.0	4.2	5.0	2.2	-1.9 [-5.0, 0.0]
	Qwen-Omni-3	16.7	18.9	20.0	19.2	-3.1 [-11.1, $+3.1$]
	InternOmni	8.3	10.0	11.7	9.2	-4.2 [-10.8, 0.0]
	gpt-4o-trans.+GPT-4o-mini	26.7	25.3	28.3	28.9	$+1.9$ [-6.4, $+10.8$]
	gpt-4o-trans.+GPT-4o	35.0	36.7	38.3	36.7	-3.3 [-15.3, $+8.3$]
	gpt-4o-trans.+Qwen2.5-VL†	6.7	6.7	5.0	5.8	$+0.8$ [0.0, $+2.5$]
	Qwen-Omni-2.5 (self-casc.)†	5.0	5.0	5.0	5.8	$+0.8$ [-0.6, $+3.1$]
	Qwen-Omni-3 (self-casc.)	16.7	16.1	21.7	21.1	0.0 [-3.1, $+2.8$]
EmbodiedBench	Qwen-Omni-2.5	35.4	33.3	30.7	28.9	$+0.3$ [-6.3, $+6.9$]
	Qwen-Omni-3	37.8	39.1	37.3	33.0	-5.7 [-12.0, $+0.1$]
	InternOmni	10.5	10.1	9.8	9.1	-0.3 [-4.3, $+3.8$]
	gpt-4o-trans.+GPT-4o-mini	22.7	24.1	21.6	20.9	-2.1 [-6.5, $+3.1$]
	gpt-4o-trans.+GPT-4o	31.8	32.9	30.2	25.9	-5.3 [-9.8, -1.1]
	gpt-4o-trans.+Qwen2.5-VL	27.5	26.4	23.8	17.1	-5.7 [-10.4, -1.0]
	Qwen-Omni-2.5 (self-casc.)	29.6	29.7	25.4	23.5	-2.0 [-4.6, $+0.4$]
	Qwen-Omni-3 (self-casc.)	39.8	38.8	32.6	28.8	-2.7 [-6.9, $+0.9$]
Pooled over 19 cells						-2.6 [-4.6, -0.7]

Long audio. We pad each human recording and its TTS counterpart with silence to three times its length, with the speech at the beginning, middle, or end, as in the main Long condition. Averaged over the three positions, the interaction with the voice is -1.1 points (95% CI [-2.9, $+0.7$], Table 22).

Informal wording. The single-utterance speakers read the clean instructions. Informal speech comes from the two-turn study instead. Its concatenated condition is an informal request said in one go. There, for Qwen-Omni-2.5, Human $-$ TTS is -0.4 Step-LCS on EmbodiedBench and -5.1 AMS on GUIBench (Table 23), so informal human speech costs at most about five points.

Multi-turn requests. Table 23 gives the collapse for each voice on the same requests. On GUIBench, Qwen-Omni-2.5 collapses by $+14.0$ AMS on TTS and by $+12.3$ on human speech (95% CI [$+3.1$, $+21.8$]). On EmbodiedBench the TTS collapse is $+9.2$ Step-LCS and the human collapse $+12.8$ (95% CI [$+2.8$, $+24.5$]). The Human $-$ TTS difference is -1.7 [-17.9, $+11.7$] on GUIBench and $+3.7$ [-6.8, $+16.1$] on EmbodiedBench. Neither difference excludes zero. The collapse has the same sign on both voices, and its size agrees within the precision that 20 requests allow. The mixed model over every reading gives a collapse of $+12.0$ [$+4.7$, $+19.3$] ($p = 0.001$) on GUIBench, from 40 readings by 39 speakers, and $+13.5$ [$+6.1$, $+20.9$] ($p < 0.001$) on EmbodiedBench, from 48 readings by 43 speakers with the clean-wording pilot included. Qwen-Omni-3 collapses more on both voices. Its collapse is $+44.4$ AMS on TTS and $+46.2$ on human speech on GUIBench, and $+20.3$ and $+23.4$ Step-LCS on EmbodiedBench. The Human $-$ TTS differences are again small, $+1.8$ [-7.1, $+9.6$] and $+3.1$ [-2.6, $+10.4$], and the mixed model gives $+44.9$ [$+36.0$, $+53.8$] and $+24.5$ [$+19.3$, $+29.8$] (both $p < 0.001$). ScheduleBench shows no multi-turn collapse even on TTS (Table 4), so its requests were recorded but are not scored here.

Three turns. The three-turn study repeats the comparison with every request split across three turns (Table 23). Human speech again shows the collapse, like TTS. The Human $-$ TTS difference is $+2.2$ [-10.0, $+15.4$] on GUIBench and 0.0 [-7.1, $+7.1$] on EmbodiedBench for Qwen-Omni-2.5, and $+3.3$ [-9.5, $+16.3$] and -0.6 [-7.2, $+5.8$] for Qwen-Omni-3, so no difference excludes zero. The mixed model finds a human collapse in all four cells. It ranges from $+9.3$ [$+3.3$, $+15.4$] ($p = 0.003$) for Qwen-Omni-2.5 on GUIBench to $+37.4$ [$+28.0$, $+46.8$] ($p < 0.001$) for Qwen-Omni-3 on

Table 22: Clean versus Long (silence-padded) speech for each voice. Long is the mean over the three positions. The interaction is the human Long cost minus the TTS Long cost, with a 95% paired bootstrap CI. † At the floor on both voices and left out of the pooled estimate.

		TTS		Human		
Benchmark	**Model**	Clean	Long	Clean	Long	**Interaction** [95% CI]
GUIBench	Qwen-Omni-2.5	32.3	34.2	32.1	29.7	-4.2 [-16.1, $+10.2$]
	Qwen-Omni-3	63.2	58.8	50.0	56.7	$+11.1$ [$+1.1$, $+24.4$]
	InternOmni†	9.3	9.3	9.3	9.3	0.0 [0.0, 0.0]
	gpt-4o-trans.+GPT-4o-mini†	6.7	8.6	8.9	8.6	-2.2 [-6.6, 0.0]
	gpt-4o-trans.+GPT-4o	17.4	16.5	18.3	17.7	$+0.3$ [-6.7, $+7.5$]
	gpt-4o-trans.+Qwen2.5-VL	14.6	14.6	14.6	13.5	-1.1 [-3.3, 0.0]
	Qwen-Omni-2.5 (self-casc.)	40.9	34.7	38.5	29.3	-2.9 [-9.3, $+3.0$]
	Qwen-Omni-3 (self-casc.)	39.9	47.7	48.4	51.1	-5.1 [-17.2, $+6.8$]
ScheduleBench	Qwen-Omni-2.5†	5.0	3.3	5.0	6.1	$+2.8$ [$+0.6$, $+6.1$]
	Qwen-Omni-3	16.7	21.1	20.0	20.6	-3.9 [-10.6, $+1.7$]
	InternOmni	8.3	4.4	11.7	3.3	-4.4 [-11.7, 0.0]
	gpt-4o-trans.+GPT-4o-mini	26.7	25.6	28.3	26.1	-1.1 [-10.0, $+7.8$]
	gpt-4o-trans.+GPT-4o	35.0	36.1	38.3	40.6	$+1.1$ [-13.3, $+15.0$]
	gpt-4o-trans.+Qwen2.5-VL†	6.7	6.7	5.0	5.6	$+0.6$ [0.0, $+1.7$]
	Qwen-Omni-2.5 (self-casc.)†	5.0	4.4	5.0	6.1	$+1.7$ [0.0, $+3.9$]
	Qwen-Omni-3 (self-casc.)	16.7	16.7	21.7	21.1	-0.6 [-1.7, 0.0]
EmbodiedBench	Qwen-Omni-2.5	35.4	33.8	30.7	32.3	$+3.2$ [-5.3, $+12.5$]
	Qwen-Omni-3	37.8	43.1	37.3	35.8	-6.8 [-14.4, -0.2]
	InternOmni	10.5	15.3	9.8	10.7	-3.9 [-9.4, $+1.7$]
	gpt-4o-trans.+GPT-4o-mini	22.7	27.0	21.6	23.9	-2.0 [-7.1, $+2.8$]
	gpt-4o-trans.+GPT-4o	31.8	34.7	30.2	31.4	-1.7 [-9.3, $+4.5$]
	gpt-4o-trans.+Qwen2.5-VL	27.5	27.8	23.8	25.6	$+1.5$ [-0.8, $+4.2$]
	Qwen-Omni-2.5 (self-casc.)	29.6	30.4	25.4	25.1	-1.1 [-3.5, $+1.0$]
	Qwen-Omni-3 (self-casc.)	39.8	39.8	32.6	33.9	$+1.3$ [-0.6, $+3.3$]
Pooled over 19 cells						-1.1 [-2.9, $+0.7$]

GUIBench. On EmbodiedBench the collapse is larger at three turns than at two for both models and both voices. On human speech Qwen-Omni-3 collapses by $+30.8$ Step-LCS against $+23.4$ at two turns, and Qwen-Omni-2.5 by $+14.7$ against $+12.8$. GUIBench does not show this. On its 20 episodes both models collapse less at three turns than at two on both voices, and the TTS collapse of Qwen-Omni-2.5 is only $+6.0$ AMS (95% CI [-7.5, $+19.1$]). The two-turn and three-turn requests are different variants, so these comparisons across turn counts are descriptive. Within each turn count, human speech follows TTS in both benchmarks.

Table 23: The multi-turn collapse (concat − multi-turn) on TTS and on human speech, on the same requests. Human numbers average each request over its speakers, and each speaker's collapse uses their own concatenated turns. Turns is the number of turns a request is split into, and n is the number of requests.

				TTS			Human			
Benchmark	**Model**	Turns	n	Concat	Multi	Collapse	Concat	Multi	Collapse	**Human − TTS** [95% CI]
GUIBench	Qwen-Omni-2.5	2	20	35.0	21.0	$+14.0$	29.9	17.6	$+12.3$	-1.7 [-17.9, $+11.7$]
GUIBench	Qwen-Omni-2.5	3	20	30.4	24.4	$+6.0$	31.5	23.3	$+8.3$	$+2.2$ [-10.0, $+15.4$]
GUIBench	Qwen-Omni-3	2	20	59.9	15.5	$+44.4$	58.2	12.0	$+46.2$	$+1.8$ [-7.1, $+9.6$]
GUIBench	Qwen-Omni-3	3	20	60.9	24.3	$+36.5$	61.5	21.6	$+39.9$	$+3.3$ [-9.5, $+16.3$]
EmbodiedBench	Qwen-Omni-2.5	2	20	41.6	32.4	$+9.2$	41.2	28.4	$+12.8$	$+3.7$ [-6.8, $+16.1$]
EmbodiedBench	Qwen-Omni-2.5	3	19	39.3	24.6	$+14.7$	38.5	23.8	$+14.7$	0.0 [-7.1, $+7.1$]
EmbodiedBench	Qwen-Omni-3	2	20	57.6	37.3	$+20.3$	57.2	33.8	$+23.4$	$+3.1$ [-2.6, $+10.4$]
EmbodiedBench	Qwen-Omni-3	3	19	57.1	25.7	$+31.5$	55.3	24.5	$+30.8$	-0.6 [-7.2, $+5.8$]

An ASR observation. Left to detect the language, the cascades' ASR turns correct English into another language or script on about 2% of human recordings. Examples are Cyrillic for a Latvian speaker and Afrikaans for a South African speaker. The failure is random, since the same clip usually transcribes correctly on a second try, and it hits native and non-native speakers at similar rates. In a separate check over 14 edge-tts English accents, no synthetic voice triggered it. It is too rare to move the scores above, but a synthesized benchmark cannot surface it.

Limitations. Speakers read the instructions aloud rather than phrasing them on their own. This holds the content fixed, which a matched-item design needs, but read speech is more fluent than a user improvising a request. The study therefore bounds the gap between voices, not the full gap to

natural use. The two multi-turn studies script the dialogue in the same way as the benchmark. Men outnumber women in all three studies. We evaluate one reading per single-utterance item.

O DATA STATISTICS

This appendix lists audio lengths and text lengths. Table 1 in the main text gives the 11,700-sample breakdown. Table 27 reports text statistics. Wild conditions consistently expand vocabulary across all three benchmarks (e.g., GUIBench grows from 1,330 unique words on clean to about 2,000 under Informal and Multi-Turn). Average instruction length increases sharply for GUIBench and EmbodiedBench, while ScheduleBench stays near its spoken clean length. Its 234-word written instruction includes a role preamble that is not spoken, and the spoken clean instruction has about 120 words. Tables 28, 24, and 25 give per-benchmark breakdowns: ScheduleBench by task and difficulty, EmbodiedBench by task type, and GUIBench by application category.

Table 24: EmbodiedBench statistics by task type.

Task Type	# Tasks	Avg. Actions	Avg. Objects	# Scenes
Look at object in light	74	4.0	2.9	2
Pick, clean, then place	54	6.7	3.8	2
Pick, cool, then place	46	7.2	3.9	1
Pick, heat, then place	57	7.4	3.9	1
Pick two objects and place	69	5.2	3.1	4
Total / Avg.	**300**	**5.9**	**3.5**	**10**

Table 25: GUIBench statistics by category.

Category	Episodes	Avg Steps	Unique Apps
General Tool	50	12.0	44
Information Management	50	16.2	41
Media Entertainment	50	10.2	30
Multi Apps	50	24.4	49
Social Sharing	50	13.0	47
Web Shopping	50	17.8	41

Table 26: Audio duration statistics (seconds) across benchmarks and perturbation conditions. Infor. = Informal. Multi. = Multi-Turn, where the reported duration is the sum across multiple audio segments within one instance. Long is averaged over begin/mid/end insertion positions, and Noise is averaged over the six acoustic environments and matches the clean-audio duration by construction.

Benchmark	**Condition**	**Mean**	**Median**	**Std**	**Min**	**Max**
ScheduleBench	Clean	43.2	46.6	11.3	13.2	66.4
	Noise	43.2	46.6	11.3	13.2	66.4
	Long	129.7	139.7	34.0	39.6	199.2
	Infor.	55.6	53.3	19.8	15.7	134.5
	Multi.	55.8	55.1	13.2	30.5	109.2
GUIBench	Clean	9.3	8.8	3.0	3.0	20.8
	Noise	9.3	8.8	3.0	3.0	20.8
	Long	28.0	26.4	9.1	8.9	62.4
	Infor.	30.9	29.0	9.2	15.1	66.9
	Multi.	36.8	35.8	8.1	20.9	67.7
EmbodiedBench	Clean	2.5	2.3	0.7	1.2	5.1
	Noise	2.5	2.3	0.7	1.2	5.1
	Long	7.4	7.0	2.0	3.6	15.4
	Infor.	49.8	46.0	20.2	17.0	132.2
	Multi.	46.8	44.9	13.7	21.1	102.1

Table 27: Text instruction statistics (same numbers as Figure 2).

Benchmark	Condition	Mean Words	Std Words	Vocab
ScheduleBench	Clean	234	36	464
	Inf.	117	30	1608
	MT (N=2)	113	24	1366
	MT (N=3)	122	22	1451
GUIBench	Clean	25	7	1330
	Inf.	64	16	2071
	MT (N=2)	67	15	1994
	MT (N=3)	90	18	2323
EmbodiedBench	Clean	9	3	287
	Inf.	119	39	1235
	MT (N=2)	113	35	1136
	MT (N=3)	127	36	1203

Table 28: ScheduleBench statistics by task and difficulty.

Task	Difficulty	# Tasks	Avg. Words	Avg. Duration
Flight	Easy	34	157	27.1
	Medium	33	239	55.7
	Hard	33	240	56.7
Calendar	Easy	34	198	32.5
	Medium	33	222	32.2
	Hard	33	247	36.3
Meeting	Easy	34	250	48.7
	Medium	33	274	49.7
	Hard	33	279	50.9