AUDITING AUDIO–VISUAL INTEGRATION IN OMNI-MODAL MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Video can change speech recognition without showing whether a model uses the visible articulation. We evaluate eight omni-modal models using silent word recognition, speech in noise, repeated and permuted frames, and cross-dubs that hold the auditory waveform fixed across visual donors. The six open-weight models recognise silent command words at near-chance accuracy and show little average movement toward the donor's word, although their output distributions vary across donors. The two closed-weight models benefit from native video in selected noisy conditions. A follow-up tests one of these models with matched waveforms across native, reprocessed and dubbed clips and finds no clear benefit or preference for compatible video. A masker-only tail changes its audio-only accuracy by about eleven percentage points. A specialist audiovisual recogniser shows donor-directed effects on the same files; its benefit survives reprocessing and depends on the donor's word. We report word and onset outcomes separately, including transcripts that the onset rule cannot classify. Exploratory probes recover label information after linear nuisance adjustment, but the tested interventions do not establish how that information affects recognition. These results distinguish changes caused by adding video from evidence that a model uses visual speech, while leaving their architectural and training causes unresolved.

1 INTRODUCTION

Does a model use visible speech, or does adding video simply change its answer? Audiovisual accuracy alone cannot resolve this question. A repeated image can alter the input sequence and the response without supplying the articulation needed to recognise a word. Conversely, useful visual information need not produce a large change in accuracy when the audio is already clear.

Human speech perception suggests several complementary tests. Seeing a talker improves recognition in noise (Sumby & Pollack, 1954), especially when the audio is degraded but remains intelligible (Ross et al., 2007; Ma et al., 2009). Audiovisual offsets reduce this benefit over a task-dependent range, often around 100 to 300 ms (van Wassenhove et al., 2007; Wallace & Stevenson, 2014). Conflicting syllables can also elicit a response present in neither stream: the McGurk effect (Mcgurk & Macdonald, 1976; Magnotti & Beauchamp, 2015; 2017). Its value as a general measure of audiovisual speech perception is disputed (Alsius et al., 2018; Getz & Toscano, 2021; Van Engen et al., 2022; Ahn et al., 2024; Van Engen et al., 2017). We therefore use several tests and ask which visual manipulation changes each response.

Omni-modal large language models accept audio and video in one context (Xu et al., 2025a;b; Cui et al., 2026; Gemma Team, 2025; Abdin et al., 2024; Team, 2025). In scene and sound-event tasks, visual information can override audio (Jung et al., 2026; Nazi et al., 2026; Chen et al., 2025). Speech benchmarks examine alignment, speaker attribution and content matching on natural video (Zhou et al., 2026; Nguyen et al., 2026; Zhang et al., 2026). HumanOmni-Speaker adds a motion encoder and speech training to a Qwen2.5-Omni backbone and reports visual and audiovisual speech recognition (Bai et al., 2026). Our study complements these evaluations with a controlled question: when the auditory waveform is identical, does changing the visible word move the answer toward that word?

We test Qwen2.5-Omni, Qwen3-Omni, MiniCPM-o 2.6 and 4.5, Gemma 3n, Phi-4-multimodal, and Gemini 2.5 Flash and Pro. The main experiments use four GRID command words from seven

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.

talkers (Cooke et al., 2006). Repeated-frame and permuted-frame controls preserve the number of frames while changing their content or order. Cross-dubs pair a fixed waveform with different words spoken by the same talker. A follow-up experiment compares native video, reprocessed native video, another utterance of the same word, and an incompatible donor. A specialist audiovisual recogniser provides a positive comparison on the same clips. We also evaluate responses to human-normed McGurk stimuli (Magnotti et al., 2020) and conduct exploratory analyses of hidden states.

The measurements yield different conclusions. Adding video changes noisy-speech accuracy substantially in some open-weight models, but repeated or permuted frames produce similar changes. Average donor-directed effects are small in all six open-weight models on clean cross-dubs and in the three tested in noise. Their outputs nevertheless vary across donors. The hosted models benefit from native video in selected conditions, but Flash's benefit does not replicate in the follow-up construction; the specialist's benefit does and depends on visual compatibility. These results motivate three contributions:

1. A waveform-matched protocol that measures video-induced response changes, donor-directed effects and native-video benefit separately, with categorical word and onset outcomes for both generalist and specialist models.
2. An evaluation of eight model snapshots with explicit control coverage, sample sizes and uncertainty, showing why a response change alone does not establish visual speech use.
3. A matched comparison of native and dubbed video, supported by exploratory probes and interventions, that identifies presentation sensitivity and limits the mechanistic conclusions available from behavioural effects.

2 Related work

Human audiovisual speech perception. Noise benefit, temporal sensitivity and responses to conflicting syllables provide complementary measures of audiovisual perception (Sumby & Pollack, 1954; Ross et al., 2007; Ma et al., 2009; van Wassenhove et al., 2007; Wallace & Stevenson, 2014; Mcgurk & Macdonald, 1976; Magnotti & Beauchamp, 2015; 2017). None is a necessary test of every form of visual speech use. McGurk susceptibility is particularly contested (Alsius et al., 2018; Getz & Toscano, 2021; Van Engen et al., 2022; Ahn et al., 2024), and small samples can exaggerate group differences (Magnotti & Beauchamp, 2018). Our main behavioural evidence therefore comes from the multi-talker GRID experiments. The McGurk materials have item-level human norms (Magnotti et al., 2020); the lipreading reference from Wang et al. (2026) uses different stimuli.

Specialist audiovisual speech recognition. AV-HuBERT (Shi et al., 2022) produces non-veridical responses at a rate comparable to humans on the recordings of Mallick et al. (2015) (59% versus 56%) when transcripts are scored by initial phoneme (Ma et al., 2026). Those recordings come from the same laboratory as ours but are different files. Noise-augmented AVSR networks also produce fusion responses more often than clean-trained networks (Grasse & Tata, 2025). We run Auto-AVSR on our own stimuli to assess the sensitivity of the proposed measurements. Its transcript scores vary substantially with the scoring rule, so we report strict option-word, nearest-option and onset outcomes separately. Recognition accuracy alone need not imply human-like cue use (Jain & Harte, 2026).

Speech-adapted language models. Llama-AVSR, Omni-AVSR and MMS-LLaMA combine audiovisual speech encoders with Llama or Qwen decoders and report low word error rates on LRS3 (Cappellazzo et al., 2025; 2026; Yeo et al., 2025). HumanOmni-Speaker is a close comparison: it combines Qwen2.5-Omni with a motion encoder and speech-oriented training, reports visual and audiovisual recognition, and ablates frame rate (Bai et al., 2026). We could not obtain its checkpoint for the experiments reported here. Wang et al. (2026) instead modify temporal attention, positional encoding and the injection of temporal features into decoder layers. Our sampling sweep leaves each encoder unchanged; it cannot distinguish these architectural explanations from differences in supervision. Domain transfer also matters: AVSRBench reports Auto-AVSR WER on GRID sentences increasing from 12.6% with audio alone to 21.5% with audio and video (Jain & Harte, 2026). Harm from video therefore does not imply an absence of audiovisual speech training.

Table 1: Video interfaces and tested sampling settings. Gemini rates are nominal; the API does not expose the selected frames. Qwen2.5-Omni, Qwen3-Omni and MiniCPM-o 2.6 also have internal analyses. Checkpoint and input-processing details are in Appendix A.

Model	Video interface
Qwen2.5-Omni-7B	Video tokens, 2–30 frames/s
Qwen3-Omni-30B-A3B	Video tokens, 2–30 frames/s
MiniCPM-o 2.6	1 or 15 frames per second of audio
MiniCPM-o 4.5	1 or 15 frames per second of audio
Gemma 3n E4B	Images, 2 or 8 frames/s
Phi-4-multimodal	Images, 2, 8 or 25 frames/s
Gemini 2.5 Flash	Video file, nominal 1 frame/s
Gemini 2.5 Pro	Video file, nominal 1 frame/s

Benchmarks and modality conflict. HumanVBench evaluates alignment detection and speech-content matching (Zhou et al., 2026); AV-SpeakerBench tests speaker-grounded content, timing and attributes (Nguyen et al., 2026); SVHalluc manipulates the consistency of narration and visible events (Zhang et al., 2026). Our distinctive comparison holds the waveform of a short speech token fixed while changing its visual donor and measures movement toward the donor's word. Benchmarks with one designated correct modality often reward selecting that modality (Nazi et al., 2026; Mukherjee et al., 2026; Zhou et al., 2026; Luo et al., 2026). We use speech-only audiovisual content, controlled noise and cross-dubbing; the McGurk task additionally permits a response matching neither stream.

Internal representations and interventions. Prior studies examine visual suppression of audio, cross-modal exchange, late changes in decoded answers and layer-dependent modality use (Selvakumar et al., 2026; Jung et al., 2026; Sudheer et al., 2026; Suharitdamrong et al., 2026; Yan et al., 2026). Gaps between decodable information and model responses have been reported for premise conflict (Quang et al., 2026), prosody (Peng & Nuchged, 2026) and text-suppressed audio (Cho et al., 2026). Steering can also change responses in audio language models (Lin et al., 2026). Our probes examine a speech-related instance of this broader problem. The additive interventions do not establish a reliable connection between the decoded visual label and the answer, so they cannot distinguish a missing readout from a suppressed one.

3 METHODS

Models and interfaces. Table 1 lists the eight snapshots; Appendix A gives checkpoint hashes and API dates. We score the six open-weight models from first-token probabilities and Gemini 2.5 Flash and Pro from generated answers. Hosted models receive padded clips and a task-specific prompt. The specialist uses its own preprocessing and generates a transcript. Open-model inference uses bf16 precision and one language-model forward pass per item, without speech generation. Each processor determines how video enters the context. Nominal frame rate does not specify the sampled frames of a short padded clip, and the hosted API does not expose those frames.

Stimuli. The main task uses the GRID command words *bin*, *lay*, *place* and *set* (Cooke et al., 2006), cut from word alignments with 80 ms of padding. We sampled 200 clips from each of seven talkers; 1,379 of the 1,400 clips have readable video. The words form two onset classes: bilabial (*bin*, *place*) and alveolar (*lay*, *set*). The syllable tests use the release of Magnotti et al. (2020): 20 auditory /ba/ tokens paired with one visual /ga/ articulation, plus a second talker's /ba/, /da/ and /ga/ recordings arranged in a 3×3 factorial. Human response rates are available for the 20 McGurk items.

Speech in noise and silent recognition. Native GRID clips are tested with and without video at -12, -6 and 0 dB SNR and in quiet. Speech-shaped noise is scaled using the whole-clip signal level: 400 items per condition from talkers 1 to 4, and 300 from talkers 7 to 9. Additional runs vary the noise seed and SNR. Silent recognition is four-way forced choice on all 1,379 valid clips. Control coverage differs by model (Table 2): most models have repeated-frame and permuted-frame

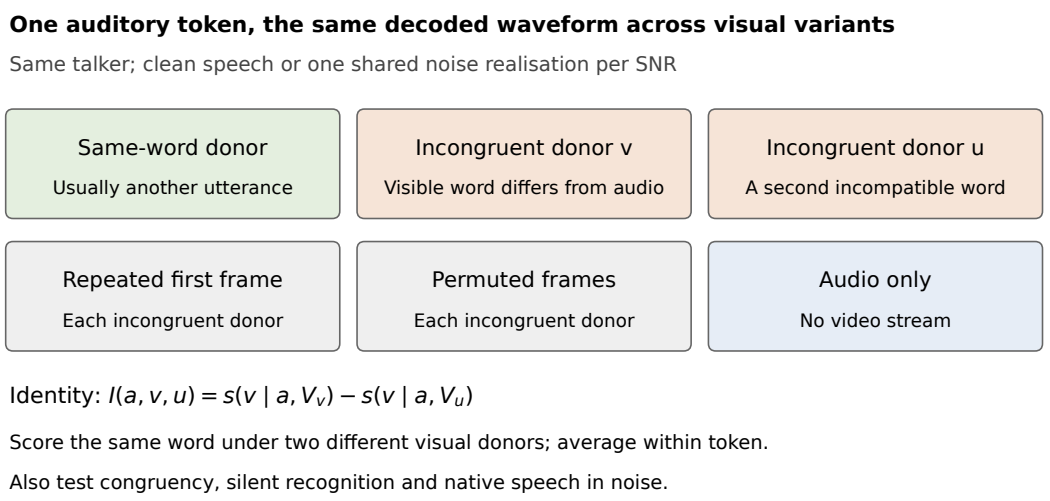


Figure 1: Evaluation protocol. Each auditory token is paired with videos of the same or different words from the same talker, repeated or permuted frames, and an audio-only reference. Decoded-audio hashes verify waveform identity across visual variants. Congruent donors are usually alternate utterances of the same word. Silent recognition and native speech in noise provide complementary tests; Auto-AVSR is evaluated on the same clips.

cells on talkers 7 to 9; Qwen2.5-Omni and MiniCPM-o 2.6 also have them on talkers 1 to 4. Pro's controls cover talker 7 only, and Gemma 3n has no matched GRID control set.

Waveform-matched cross-dubbing. For each of 329 auditory tokens, one canonical waveform is paired with each available visual word from the same talker, controls for the incongruent donors, and no video (3,566 files). Congruent donors are alternate utterances for 317 tokens, the original recording for 7, and unavailable for 5. Each token's files decode to the same audio hash. Noisy versions at -6 and 0 dB reuse one masker per token and SNR (7,132 files). Their level convention uses the standard deviation of non-silent samples; whole-clip SNR is consequently a median 0.8 dB lower than the nominal value (Appendix C).

Matched reconstruction experiment. We selected 96 tokens before inspecting outcomes: eight per word from each of talkers 7 to 9. At -6 and 0 dB, one noisy waveform per token is reused across reprocessed native video, an alternate utterance of the same word, an incompatible donor from the other onset class, repeated and permuted frames, and an audio-only reference. The unpadded native clip and a second audio-only reference use the identical prefix of that waveform. This design separates visual comparisons from the effect of extending the presentation with a masker-only tail. Noise scaling follows the native task. We refer to this set as the *bridge* experiment (Appendix D).

Visual controls and supplementary tests. The repeated-frame control repeats frame 0; the permuted-frame control uses a fixed-seed permutation. Both preserve the frame count. Token counts are matched where the processor exposes them. Neither control removes all speech information: a still mouth shape or an unordered set of shapes may remain informative, and permutation introduces unnatural motion. We therefore name contrasts by their interventions rather than treating them as isolated effects of presence, content or order. Supplementary tests vary frame rate on the McGurk set and shift audio from -400 to $+400$ ms within 500 ms silent pads. Offset tests and internal analyses cover the three Qwen-lineage models.

Scoring. The canonical syllable prompt requests exactly one of *ba*, *da* or *ga*; GRID runs substitute the four command words, with wording differences documented in Appendix A.2. For open models, we sum each option's first-token variants and renormalise over the choice set. The argmax supplies the categorical answer. Hosted models return a named option at temperature 0. Audits report choice-set mass, prompt sensitivity and option-order effects (Appendix A.4). We distinguish categorical word outcomes, option probabilities where available, and onset class. The specialist's transcript

Table 2: Silent-word accuracy and clean categorical identity effects, with 95% item and token intervals, respectively. All models use 1,379 silent clips; cross-dubs use 329 tokens except Pro (not run). Noise cohorts are A: talkers 1–4 and 7–9 ($n = 698$ per SNR); B: talkers 7–9 ($n = 298$); B7: talker 7 ($n = 81$). ‡Hosted models use padded clips and categorical answers.

Model	Silent accuracy	Identity effect	Noise cohort
Qwen2.5-Omni	$0.251\ [0.228, 0.274]$	$-0.002\ [-0.006, +0.003]$	A
Qwen3-Omni	$0.249\ [0.226, 0.272]$	$+0.002\ [+0.000, +0.004]$	B
MiniCPM-o 2.6	$0.251\ [0.228, 0.275]$	$+0.000\ [-0.007, +0.007]$	A
MiniCPM-o 4.5	$0.244\ [0.221, 0.266]$	$-0.001\ [-0.004, +0.003]$	B
Gemma 3n E4B	$0.249\ [0.227, 0.273]$	$-0.005\ [-0.014, +0.003]$	—
Phi-4-multimodal	$0.249\ [0.227, 0.273]$	$+0.001\ [-0.009, +0.011]$	B
Gemini 2.5 Flash‡	$0.249\ [0.227, 0.272]$	$-0.005\ [-0.012, +0.001]$	B
Gemini 2.5 Pro‡	$0.273\ [0.249, 0.297]$	—	B7

is mapped to a word or to its initial-letter class (B/P/M versus L/S/T/D/N/Z). Full-denominator onset scores count unclassifiable transcripts as non-matches; conditional scores report their reduced coverage separately.

Probes and interventions. Linear probes use mean video-position states, leave-one-talker-out evaluation, and fold-local scaling, PCA and nuisance adjustment. We also apply mean ablation, activation patching and additive steering with twenty norm-matched control directions per block and an audio-direction comparison. Appendix F.2 specifies which probes have nuisance adjustment. Layer L denotes decoder-block output L; hidden-state index $L + 1$ refers to that same output.

Statistics. Primary contrasts pair conditions within a clip or auditory token. Point estimates weight items or tokens equally; 95% percentile bootstrap intervals resample the stated unit. Intervals over talker means address a separate, equally weighted talker target (Appendix A.3). An interval containing zero is not evidence of equivalence. The main contrasts were fixed after the first experiments but before analysis of the repaired stimuli; this was an analysis plan for collected data, not a prospective preregistration. Single-talker cells use one deterministic pass. Repeatability on fixed hardware and sensitivity across GPU types are reported separately from sampling uncertainty (Appendix A.3).

4 Results

Tables 2 and 3 report coverage and matched contrasts. Figure 2 shows their uncertainty; Appendices B–D give supporting results and checks.

4.1 Silent word recognition

On 1,379 silent GRID clips, seven models have four-way accuracy between 0.244 and 0.251; their item-bootstrap intervals include chance (0.25). Pro reaches $0.273\ [0.249, 0.297]$, with its small excess concentrated in four talkers (Appendix B.1). Many models strongly favour one answer. The specialist's transcript identifies the correct onset class on $0.916\ [0.898, 0.933]$ of 973 classifiable transcripts, or 0.646 of all 1,379 clips. This coarse discrimination confirms that the clips contain usable visual information, although it is not equivalent to four-way word recognition.

4.2 Directional influence under cross-dubbing

We measure two signed contrasts. *Congruency* compares the audio-word score under congruent and incongruent video. *Identity* compares the score of visual word v when the donor says v rather than another incongruent word u, with auditory token a fixed:

$$I(a, v, u) = s(v \mid a, V_v) - s(v \mid a, V_u). \tag{1}$$

Here s is an option probability or an indicator that the categorical answer matches the scored word or onset class. We average donor pairs within a token, then tokens. Positive identity effects indi-

Table 3: Paired accuracy differences on each model's matched noise cohort (Table 2). All three contrasts in a row use the same clips. Figure 2 gives their 95% item-bootstrap intervals. Models use their densest tested setting; Gemma 3n has no matched control cohort.

Model	SNR	AV − AO	Intact − repeated	Intact − permuted
Qwen2.5-Omni	−6 dB	−0.238	−0.054	−0.016
	0 dB	−0.153	−0.047	−0.029
Qwen3-Omni	−6 dB	+0.030	+0.023	−0.023
	0 dB	+0.037	−0.003	−0.003
MiniCPM-o 2.6	−6 dB	−0.014	+0.000	+0.000
	0 dB	−0.050	+0.009	+0.010
MiniCPM-o 4.5	−6 dB	−0.114	−0.003	−0.003
	0 dB	−0.027	−0.007	+0.000
Phi-4-multimodal	−6 dB	−0.040	−0.017	−0.020
	0 dB	+0.067	−0.030	−0.057
Gemini 2.5 Flash‡	−6 dB	−0.060	+0.057	+0.124
	0 dB	+0.091	+0.168	+0.138
Gemini 2.5 Pro‡	−6 dB	+0.099	+0.037	+0.111
	0 dB	+0.111	+0.136	+0.062

cate average movement toward the visible word; they do not isolate articulation from other donor differences.

On clean intact dubs, the six open-weight models have categorical identity effects within $\pm.005$, with token intervals inside $\pm.014$ (Table 4). Probability effects are within $\pm.002$ (Table 5). Repeated-frame and permuted-frame categorical effects are within $\pm.013$; two intervals exclude zero in opposite directions. Congruency effects are within $\pm.024$, with the only interval excluding zero negative (Phi-4-multimodal). Small signed means do not imply invariant responses: changing the incongruent donor changes the argmax for 2 to 36 of 329 tokens per model, and median maximum pairwise total-variation distances range from 0.007 to 0.102 (Appendix C).

The specialist's nearest-option identity effect is $+0.006$ $[-0.002, +0.013]$. Its full-denominator onset effect is $+0.071$ $[+0.052, +0.090]$, compared with near-zero effects for repeated and permuted frames. The effect occurs between visual words from different onset classes. In noise, full-denominator onset effects rise to $+0.260$ $[+0.232, +0.290]$ at -6 dB and $+0.291$ at 0 dB; nearest-option effects are $+0.081$ and $+0.048$. Conditional onset scoring gives $+0.549$ at -6 dB but retains only 142 of 329 tokens. We therefore use the full-denominator result for the main comparison and report conditional coverage separately (Table 6).

The three open-weight models tested in noise—Qwen2.5-Omni, Qwen3-Omni and MiniCPM-o 2.6—have intact categorical identity effects within $\pm.017$ and congruency effects within $\pm.009$. The only intact identity interval excluding zero is negative: Qwen2.5-Omni at -6 dB, -0.017 $[-0.032, -0.001]$. Flash's identity effects are -0.005 $[-0.012, +0.001]$ clean, $+0.012$ $[-0.004, +0.028]$ at -6 dB and -0.005 $[-0.017, +0.008]$ at 0 dB. These estimates provide little evidence of average donor-directed movement under the tested conditions.

4.3 Native speech in noise

Video changes recognition differently across models. On the matched cohorts, Qwen2.5-Omni loses 0.238 $[0.205, 0.272]$ accuracy at -6 dB, and MiniCPM-o 4.5 loses 0.114 $[0.070, 0.158]$. Qwen3-Omni's AV-minus-AO estimates are within $\pm.04$; Phi-4-multimodal gains $+0.067$ $[-0.007, +0.141]$ at 0 dB. Gemma 3n gives essentially one answer whenever GRID frames are present, so its results lack a matched control comparison.

Among the five open-weight models with matched controls, intact-minus-control point estimates lie in $[-0.06, +0.03]$ at -6 and 0 dB. Where an interval excludes zero, intact video performs worse: for example, Qwen2.5-Omni versus a repeated frame at -6 dB (-0.054 $[-0.074, -0.034]$), and Phi-4-multimodal versus permuted frames at 0 dB (-0.057 $[-0.091, -0.027]$). Some intervals remain too wide to establish equivalence at 0.05, including Qwen3-Omni versus repeated frames at -6 dB

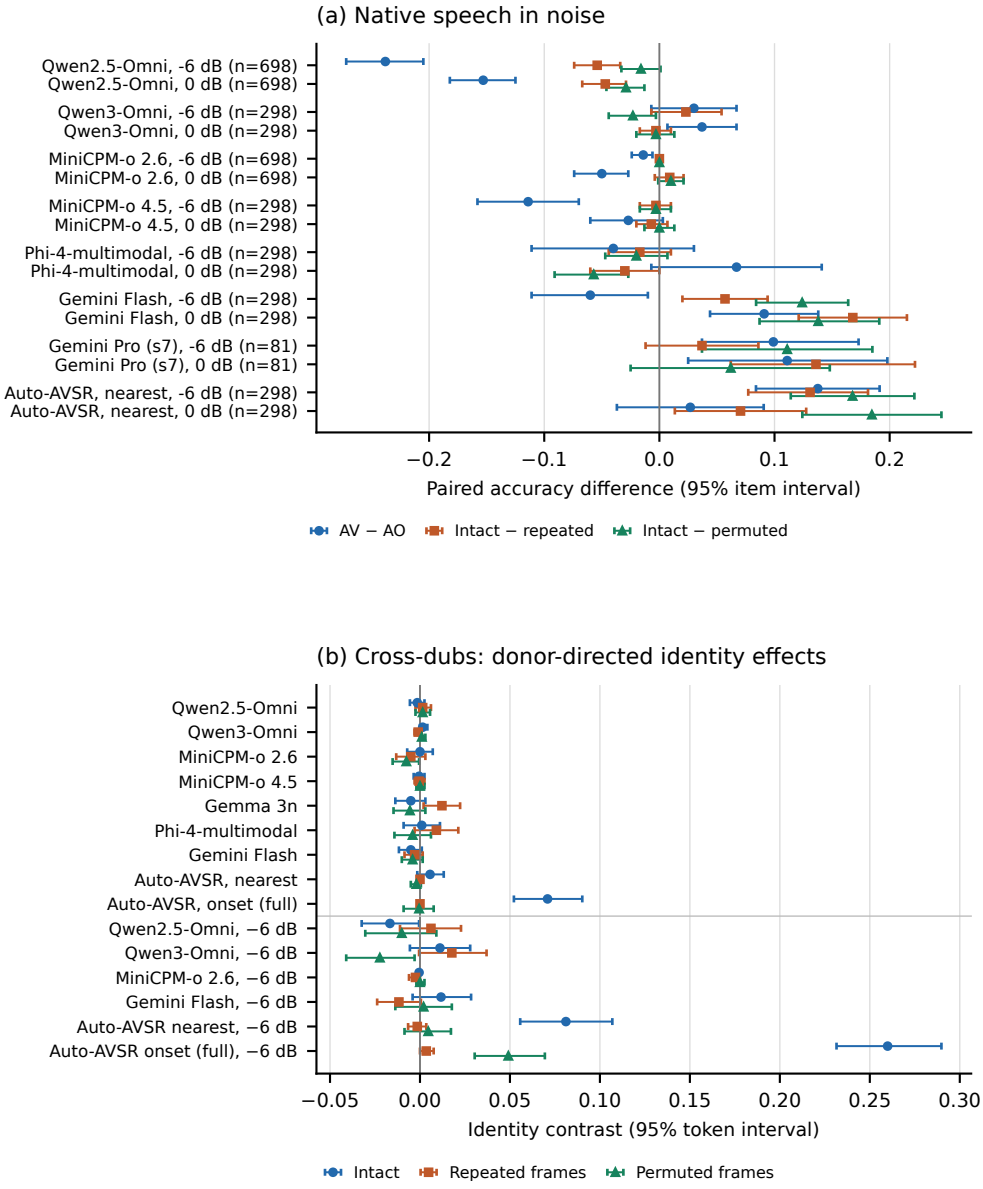


Figure 2: Primary contrasts and 95% bootstrap intervals. (a) Native speech in noise: intact video minus audio-only, repeated-frame and permuted-frame input, paired within each model's matched cohort. Labels give sample sizes. Pro's controls cover talker 7; Gemma 3n has no matched control set. Auto-AVSR uses nearest-option scoring. (b) Cross-dub identity effects on categorical word outcomes, with token-bootstrap intervals. The specialist additionally uses full-denominator onset scoring, which counts unclassifiable transcripts as non-matches.

($+0.023$ $[-0.007, +0.054]$). Quiet-speech gains in MiniCPM-o 2.6 and Phi-4-multimodal are also matched or exceeded by repeated frames (Appendix B).

Among the hosted closed-weight models, on 298 matched clips from talkers 7 to 9, Flash gains $+0.091$ $[+0.044, +0.138]$ at 0 dB; repeated and permuted frames instead reduce accuracy. At -6 dB, intact video is worse than audio-only (-0.060) but better than both controls ($+0.057$, $+0.124$). Pro gains $+0.120$ $[+0.080, +0.163]$ at -6 dB and $+0.063$ $[+0.033, +0.093]$ at 0 dB on 300 paired clips. Its four-condition comparisons use 81 talker-7 clips per SNR; two of four intact-minus-control intervals exclude zero (Appendix B.1). These native-video results and Flash's small cross-dub effects motivate the matched reconstruction experiment.

Matched reconstruction. On the 96-token bridge set, Auto-AVSR gains $+0.208$ $[+0.115, +0.302]$ from native video at -6 dB relative to audio of the same duration. The benefit remains after reprocessing and with a same-word alternate utterance. Compatible donors outperform incompatible donors by $+0.240$ $[+0.135, +0.354]$ at -6 dB and $+0.302$ $[+0.188, +0.417]$ at 0 dB (Table 7). Qwen2.5-Omni is harmed by each video variant and shows no positive compatibility effect.

Flash's native-video effects are -0.021 $[-0.104, +0.062]$ at -6 dB and $+0.021$ $[-0.062, +0.104]$ at 0 dB relative to duration-matched audio. Its compatibility effects are 0.000 $[-0.052, +0.052]$ and -0.010 $[-0.094, +0.062]$. Adding a masker-only tail to audio alone raises accuracy by $+0.104$ $[+0.021, +0.198]$ and $+0.115$ $[+0.031, +0.198]$. Thus presentation changes audio-only recognition by an amount comparable to the earlier video benefit. The bridge does not reproduce that benefit, but its differences in noise realisation, presentation and API date prevent attributing the discrepancy to one cause.

4.4 Specialist comparison

Auto-AVSR uses conformer encoders trained on LRS3, VoxCeleb2 and AVSpeech, mouth-region preprocessing, and free transcription. Its visual, audio and audiovisual checkpoints are separate trained systems; AV-minus-AO therefore compares checkpoints as well as inputs. On silent clips, strict option-word accuracy (0.033), nearest-option accuracy (0.311) and conditional onset accuracy (0.916) differ markedly. On native speech in noise, nearest-option AV-minus-AO is $+0.133$ $[+0.097, +0.173]$ at -12 dB and $+0.140$ $[+0.087, +0.197]$ at -6 dB. At -6 dB intact video also exceeds both visual controls (Appendix C.1). The specialist provides a positive comparison for these measurements, rather than a human perceptual norm.

4.5 Responses to conflicting syllables

Humans report "da" on 0.454 of trials for the 20 auditory-/ba/, visual-/ga/ items (**?**). Several models predominantly follow the audio. Qwen2.5-Omni's mean $p(\mathrm{da})$ rises from 0.123 at 2 frames/s to 0.571 at 30 ($+0.448$ $[+0.424, +0.471]$). This difference decomposes into repeated frames at 30 minus intact video at 2 ($+0.141$), permuted minus repeated frames at 30 ($+0.234$), and intact minus permuted frames at 30 ($+0.073$ $[+0.049, +0.097]$). These are intervention contrasts, not isolated causal components. The order-sensitive residual accompanies movement toward the model's silent-video default and lower scores for congruent syllables (Appendix E). Phi-4-multimodal instead moves toward "ga" as frames are added, but a repeated frame produces almost the same response (0.72 versus 0.74). Neither pattern alone establishes human-like fusion. Offset tests on the three Qwen-lineage models show no clear synchrony peak under these clean-speech conditions (Appendix E.1).

5 Exploratory representation and intervention analyses

We examine hidden states in the three Qwen-lineage models to assess whether the behavioural results coincide with a lack of decodable information (Appendix F). A linear probe, evaluated on held-out talkers, predicts the bilabial-versus-alveolar onset label from mean video-position states with accuracy 0.71 in Qwen2.5-Omni (seven talkers), 0.71 in Qwen3-Omni and 0.77 in MiniCPM-o 2.6 (four talkers each). Duration and sentence start time alone predict the label at 0.63. After fold-local linear adjustment for these variables, Qwen2.5-Omni reaches 0.61 to 0.67 from hidden-state index 8 onward (layer-selection permutation $p = 0.01$). Transfer between two word pairs reaches 0.53 to 0.66, and a four-way word probe reaches 0.39 where the model's own silent-word accuracy is 0.25. These are supervised label-prediction results within a four-word lexicon. Missing start times and nonlinear nuisance information remain limitations (Appendix F.2).

Mean ablation and activation patching show that sensitivity to modality positions varies across layers (Fig. 3a). They do not identify a unique layer at which the answer is determined. In a separate steering study, directions estimated from talkers 1 to 4 were applied to 48 noisy cross-dubs from held-out talkers. Across both signs and six gains, the visual direction did not provide consistent, direction-specific control relative to twenty norm-matched directions per block. An analogous audio direction moved responses in the expected directions at blocks 8 and 12, although its effect was not

monotonic in gain (Fig. 3c). The visual intervention's failure does not establish whether the decoded information is unused, suppressed or inaccessible to this intervention.

6 DISCUSSION

What the measurements establish. Adding video, changing the visible word and preserving native synchrony are different interventions. In the open-weight models with matched controls, repeated or permuted frames reproduce much of the change in noisy-speech accuracy. On clean cross-dubs, average categorical identity effects are small in all six open-weight models ($|I| \leq 0.005$, token intervals within $\pm.014$), despite variation in their outputs across donors. Flash and Pro benefit from native video in selected conditions, but Flash's advantage does not replicate in the matched follow-up. Auto-AVSR shows both donor-directed effects and a native-video benefit that survives reprocessing. A single audiovisual accuracy score would obscure these differences.

Unresolved explanations. Input sampling may limit access to articulation. The number and timing of sampled frames matter, and increasing frame rate through an unchanged encoder does not test the temporal architectures of **?** or **?**. Natural synchrony and domain match may also matter. The specialist tolerates cross-utterance dubbing on the follow-up set, but this need not hold for other systems; specialist AVSR itself can degrade under domain shift (**?**). Training is another possibility: speech-adapted systems succeed with related language-model backbones (Cappellazzo et al., 2025; 2026; Yeo et al., 2025), but we do not compare matched training conditions, and undisclosed speech data cannot be assumed absent. Finally, the probes and interventions leave open how decodable information contributes to the model's own responses.

Limitations. The main task uses four English words from seven talkers. Donors differ in pose, motion and timing as well as articulation, and repeated or permuted frames retain some speech information. Most congruent dubs use a different utterance of the same word (317 of 324 available tokens). The noisy cross-dubs use a different level convention from the native task; the follow-up matches that convention but covers only 96 tokens, three talkers and three systems. Flash's earlier and follow-up runs differ in noise realisation, presentation and API date, so the source of the discrepancy remains unresolved. Pro has controls for one talker and no cross-dub or follow-up results. The McGurk set contains one visual articulation. Internal analyses cover one model lineage, and the nuisance-adjusted onset probe is limited to Qwen2.5-Omni, with 579 of 1,379 start times imputed. These restrictions limit conclusions about other tasks, models and populations.

Conclusion. Our audit shows that responsiveness to video, sensitivity to the visible word and decodable speech information can come apart. In the tested omni-modal models, video changes recognition without consistently moving responses toward the visual donor, and native-video benefits depend on the tested presentation. The specialist comparison shows that the same stimuli can support donor-directed effects and benefits that survive reprocessing. Probes also recover word and onset labels from hidden states, but the interventions leave their role in recognition unresolved. Within the scope of these controlled speech tasks, the results support evaluating integration through complementary evidence: waveform-matched donor comparisons, visual controls, native-video benefit and a specialist reference. This approach makes it possible to identify what visual input contributes while keeping the remaining architectural and training explanations open.

Reproducibility statement

Appendix A documents model snapshots, input processing, stimuli, prompts, scoring and resampling. Appendices C and D specify waveform matching and reconstruction; Appendix F gives the probe and intervention procedures, including nuisance adjustment and control directions. The accompanying materials contain builder and analysis scripts, manifests, run records and environment exports. Derived videos are excluded where redistribution rights are unspecified.

Ethics statement

We collected no new human-participant data. Human response rates come from the aggregate release of Magnotti et al. (2020); we did not access individual participant records. GRID (Cooke et al., 2006) is distributed under CC BY 4.0. The OSF release does not specify redistribution rights for derived clips, so we provide construction scripts and checksums rather than those videos. Models were used under their published licences or API terms; the hosted systems may change without a version that users can pin. The small English-language stimulus set does not support conclusions about other languages, accents or speaker populations. Assistive applications should establish the visual contribution under their intended conditions. Lipreading capabilities also raise privacy and surveillance concerns, although this study does not train or release a new recognition model. The authors declare no competing interests. Institutional clusters provided compute.

AI use statement

We used generative AI tools to help with coding and editing. The tools were not used to generate synthetic datasets or conduct qualitative or thematic analysis. The authors are responsible for the paper's methods, results and final text.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. Phi-4 Technical Report, December 2024. URL http://arxiv.org/abs/2412.08905. arXiv:2412.08905 [cs].

EunSeon Ahn, Areti Majumdar, Taraz G Lee, and David Brang. Evidence for a Causal Dissociation of the McGurk Effect and Congruent Audiovisual Speech Perception via TMS to the Left pSTS. *Multisensory research*, 37(4-5):341–363, August 2024. ISSN 2213-4794. doi: 10.1163/22134808-bja10129. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC11388023/.

Agnès Alsius, Martin Paré, and Kevin G. Munhall. Forty Years After Hearing Lips and Seeing Voices: the McGurk Effect Revisited. *Multisensory Research*, 31(1-2):111–144, 2018. ISSN 2213-4794, 2213-4808. doi: 10.1163/22134808-00002565. URL https://brill.com/view/journals/msr/31/1-2/article-p111_7.xml.

Detao Bai, Zhiheng Ma, and Xihan Wei. HumanOmni-Speaker: Identifying Who said What and When, July 2026. URL https://arxiv.org/abs/2603.21664. Version Number: 3.

Umberto Cappellazzo, Minsu Kim, Honglie Chen, Pingchuan Ma, Stavros Petridis, Daniele Falavigna, Alessio Brutti, and Maja Pantic. Large Language Models are Strong Audio-Visual Speech Recognition Learners. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, April 2025. doi: 10.1109/ICASSP49660.2025.10889251. URL https://ieeexplore.ieee.org/document/10889251/. ISSN: 2379-190X.

Umberto Cappellazzo, Xubo Liu, Pingchuan Ma, Stavros Petridis, and Maja Pantic. OMNI-AVSR: Towards Unified Multimodal Speech Recognition With Large Language Models. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*,

pp. 17772–17776, May 2026. doi: 10.1109/ICASSP55912.2026.11461431. URL https://ieeexplore.ieee.org/document/11461431. ISSN: 2379-190X.

Tianle Chen, Chaitanya Chakka, Arjun Reddy Akula, Xavier Thomas, and Deepti Ghadiyaram. Some Modalities are More Equal Than Others: Decoding and Architecting Multimodal Integration in MLLMs, December 2025. URL http://arxiv.org/abs/2511.22826. arXiv:2511.22826 [cs.CV].

Hyebin Cho, Suho Yoo, Jaehyuk Jang, Changick Kim, and Joon Son Chung. Who Wins the Conflict? Mechanistic Interpretability of Text Bias in Audio LLMs, September 2026. URL http://arxiv.org/abs/2606.18924. arXiv:2606.18924 [cs.SD].

Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao. An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, 120(5 Pt 1):2421–2424, November 2006. ISSN 0001-4966. doi: 10.1121/1.2229005.

Junbo Cui, Bokai Xu, Chongyi Wang, Tianyu Yu, Weiyue Sun, Yingjing Xu, Tianran Wang, Zhihui He, Wenshuo Ma, Tianchi Cai, Jiancheng Gui, Luoyuan Zhang, Xian Sun, Fuwei Huang, Moye Chen, Zhuo Lin, Hanyu Liu, Qingxin Gui, Qingzhe Han, Yuyang Wen, Huiping Liu, Rongkang Wang, Yaqi Zhang, Hongliang Wei, Chi Chen, You Li, Kechen Fang, Jie Zhou, Yuxuan Li, Guoyang Zeng, Chaojun Xiao, Yankai Lin, Xu Han, Maosong Sun, Zhiyuan Liu, and Yuan Yao. MiniCPM-o 4.5: Towards Real-Time Full-Duplex Omni-Modal Interaction, April 2026. URL http://arxiv.org/abs/2604.27393. arXiv:2604.27393 [cs.CL].

Gemma Team. Gemma 3n model overview. https://ai.google.dev/gemma/docs/gemma-3n, 2025. Google DeepMind; model google/gemma-3n-E4B-it.

Laura M. Getz and Joseph C. Toscano. Rethinking the McGurk effect as a perceptual illusion. *Attention, Perception & Psychophysics*, 83(6):2583–2598, August 2021. ISSN 1943-393X. doi: 10.3758/s13414-021-02265-6.

Lukas Grasse and Matthew S. Tata. Artificial Neural Networks Trained on Noisy Speech Exhibit the McGurk Effect, October 2025. URL http://arxiv.org/abs/2411.05715. arXiv:2411.05715 [cs.SD].

Rishabh Jain and Naomi Harte. AVSRBench: A Multi-Condition AVSR Benchmark, September 2026. URL http://arxiv.org/abs/2609.10366. arXiv:2609.10366 [eess.AS].

Jihoo Jung, Chaeyoung Jung, Ji-Hoon Kim, and Joon Son Chung. Probing Cross-modal Information Hubs in Audio-Visual LLMs, May 2026. URL http://arxiv.org/abs/2605.10815. arXiv:2605.10815 [cs.AI].

Tsung-En Lin, Kuan-Yi Lee, and Hung-Yi Lee. Silence is Golden: Mitigating Hallucinations in Large Audio-Language Models via Layer-Weighted Vector Steering, August 2026. URL http://arxiv.org/abs/2510.12851. arXiv:2510.12851 [cs.SD].

Ziyang Luo, Nian Liu, and Junwei Han. Chain of Modality: From Static Fusion to Dynamic Orchestration in Omni-MLLMs, April 2026. URL http://arxiv.org/abs/2604.14520. arXiv:2604.14520 [cs.CV].

Haotian Ma, Zhengjia Wang, Xiang Zhang, John F. Magnotti, and Michael S. Beauchamp. A deep neural network model of audiovisual speech recognition reports the McGurk effect. *Psychonomic Bulletin & Review*, 33(2):84, February 2026. ISSN 1531-5320. doi: 10.3758/s13423-025-02846-8.

Pingchuan Ma, Alexandros Haliassos, Adriana Fernandez-Lopez, Honglie Chen, Stavros Petridis, and Maja Pantic. Auto-AVSR: Audio-Visual Speech Recognition with Automatic Labels. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, June 2023. doi: 10.1109/ICASSP49357.2023.10096889. URL https://ieeexplore.ieee.org/document/10096889/. ISSN: 2379-190X.

Wei Ji Ma, Xiang Zhou, Lars A. Ross, John J. Foxe, and Lucas C. Parra. Lip-Reading Aids Word Recognition Most in Moderate Noise: A Bayesian Explanation Using High-Dimensional Feature Space. *PLOS ONE*, 4(3):e4638, March 2009. ISSN 1932-6203. doi: 10.1371/journal.pone.0004638. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0004638.

John F. Magnotti and Michael S. Beauchamp. The noisy encoding of disparity model of the McGurk effect. *Psychonomic Bulletin & Review*, 22(3):701–709, June 2015. ISSN 1531-5320. doi: 10.3758/s13423-014-0722-2.

John F. Magnotti and Michael S. Beauchamp. A Causal Inference Model Explains Perception of the McGurk Effect and Other Incongruent Audiovisual Speech. *PLOS Computational Biology*, 13(2):e1005229, February 2017. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1005229. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005229.

John F. Magnotti and Michael S. Beauchamp. Published estimates of group differences in multisensory integration are inflated. *PLoS ONE*, 13(9):e0202908, September 2018. ISSN 1932-6203. doi: 10.1371/journal.pone.0202908. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC6145544/.

John F. Magnotti, Kristen B. Dzeda, Kira Wegner-Clemens, Johannes Rennig, and Michael S. Beauchamp. Weak observer–level correlation and strong stimulus-level correlation between the McGurk effect and audiovisual speech-in-noise: A causal inference explanation. *Cortex*, 133:371–383, December 2020. ISSN 0010-9452. doi: 10.1016/j.cortex.2020.10.002. URL https://www.sciencedirect.com/science/article/pii/S0010945220303749.

Debshila Basu Mallick, John F. Magnotti, and Michael S. Beauchamp. Variability and stability in the McGurk effect: contributions of participants, stimuli, time, and response type. *Psychonomic Bulletin & Review*, 22(5):1299–1307, October 2015. ISSN 1531-5320. doi: 10.3758/s13423-015-0817-4.

Harry Mcgurk and John Macdonald. Hearing lips and seeing voices. *Nature*, 264(5588):746–748, December 1976. ISSN 1476-4687. doi: 10.1038/264746a0. URL https://www.nature.com/articles/264746a0. Number: 5588.

Swapnanil Mukherjee, Agyeya Negi, Tanuja Ganu, and Ponnurangam Kumaraguru. C$ˆ3$PO: Evaluating Cross-Modal Composition and Counterfactual Performance in Omnimodal Models, August 2026. URL http://arxiv.org/abs/2608.05381. arXiv:2608.05381 [cs.AI].

Zabir Al Nazi, Shubhashis Roy Dipta, and Md Rizwan Parvez. Omni-Modal Dissonance Benchmark: Systematically Breaking Modality Consensus to Probe Robustness and Calibrated Abstention, March 2026. URL http://arxiv.org/abs/2603.27187. arXiv:2603.27187 [cs.LG].

Le Thien Phuc Nguyen, Zhuoran Yu, Samuel Low Yu Hang, Subin An, Jeongik Lee, Yohan Ban, SeungEun Chung, Thanh-Huy Nguyen, JuWan Maeng, Soochahn Lee, and Yong Jae Lee. See, Hear, and Understand: Benchmarking Audiovisual Human Speech Understanding in Multimodal Large Language Models, April 2026. URL http://arxiv.org/abs/2512.02231. arXiv:2512.02231 [cs.CV].

Linkai Peng and Baorian Nuchged. Represented but Ignored: A Causal Account of Prosodic Underuse in Audio-Language Models, June 2026. URL http://arxiv.org/abs/2608.19211. arXiv:2608.19211 [cs.CL].

Trung Nguyen Quang, Yiming Gao, Fanyi Pu, Kaichen Zhang, Shuo Sun, and Ziwei Liu. Senses Wide Shut: A Representation-Action Gap in Omnimodal LLMs, May 2026. URL http://arxiv.org/abs/2605.13737. arXiv:2605.13737 [cs.AI].

Lars A. Ross, Dave Saint-Amour, Victoria M. Leavitt, Daniel C. Javitt, and John J. Foxe. Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. *Cerebral Cortex*, 17(5):1147–1153, May 2007. ISSN 1047-3211. doi: 10.1093/cercor/bhl024.

Ramaneswaran Selvakumar, Kaousheik Jayakumar, S. Sakshi, Sreyan Ghosh, Ruohan Gao, and Dinesh Manocha. Do Audio-Visual Large Language Models Really See and Hear?, April 2026. URL http://arxiv.org/abs/2604.02605. arXiv:2604.02605 [cs.AI].

Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. Learning Audio-Visual Speech Representation by Masked Multimodal Cluster Prediction, March 2022. URL http://arxiv.org/abs/2201.02184. arXiv:2201.02184 [eess.AS].

Adarsh Sudheer, David Li, Omar Elbanna, Ishaan Kodarapu, Arjun Bahuguna, and Vasu Sharma. Compositional Failure in Audio-Visual LLMs: Late-Layer Prior Dominance Under Cross-modal Conflict, September 2026. URL http://arxiv.org/abs/2608.27785. arXiv:2608.27785 [cs.CL].

Wish Suharitdamrong, Muhammad Awais, Xiatian Zhu, and Sara Atito. From Senses to Decisions: The Information Flow of Auditory and Visual Perception in Multimodal LLMs, June 2026. URL http://arxiv.org/abs/2606.10147. arXiv:2606.10147 [cs.AI].

W. H. Sumby and Irwin Pollack. Visual Contribution to Speech Intelligibility in Noise. *The Journal of the Acoustical Society of America*, 26(2):212–215, March 1954. ISSN 0001-4966, 1520-8524. doi: 10.1121/1.1907309. URL https://pubs.aip.org/jasa/article/26/2/212/598972/Visual-Contribution-to-Speech-Intelligibility-in.

Gemini Team. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities, December 2025. URL http://arxiv.org/abs/2507.06261. arXiv:2507.06261 [cs].

Kristin J. Van Engen, Zilong Xie, and Bharath Chandrasekaran. Audiovisual sentence recognition not predicted by susceptibility to the McGurk effect. *Attention, Perception & Psychophysics*, 79 (2):396–403, February 2017. ISSN 1943-393X. doi: 10.3758/s13414-016-1238-9.

Kristin J. Van Engen, Avanti Dey, Mitchell S. Sommers, and Jonathan E. Peelle. Audiovisual speech perception: Moving beyond McGurk. *The Journal of the Acoustical Society of America*, 152(6): 3216, December 2022. ISSN 1520-8524. doi: 10.1121/10.0015262.

Virginie van Wassenhove, Ken W. Grant, and David Poeppel. Temporal window of integration in auditory-visual speech perception. *Neuropsychologia*, 45(3):598–607, January 2007. ISSN 00283932. doi: 10.1016/j.neuropsychologia.2006.01.001. URL https://linkinghub.elsevier.com/retrieve/pii/S002839320600011X.

Mark T. Wallace and Ryan A. Stevenson. The construct of the multisensory temporal binding window and its dysregulation in developmental disabilities. *Neuropsychologia*, 64:105–123, November 2014. ISSN 0028-3932. doi: 10.1016/j.neuropsychologia.2014.08.005. URL https://www.sciencedirect.com/science/article/pii/S0028393214002681.

Liyong Wang, Junliang Xing, Tianyu Hu, Jianfei Jiang, Jihuai Zhao, and Huimin Ma. Bridging the Temporal Gap in Multimodal LLMs: Deeply Stacking Temporal Tokens for Audio-Visual Speech Recognition. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 27748–27759, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.1381. URL https://aclanthology.org/2026.findings-acl.1381/.

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, March 2025a. URL http://arxiv.org/abs/2503.20215. arXiv:2503.20215 [cs.CL].

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, September 2025b. URL http://arxiv.org/abs/2509.17765. arXiv:2509.17765 [cs.CL].

Xinru Yan, Boxi Cao, Yaojie Lu, Hongyu Lin, Weixiang Zhou, Le Sun, and Xianpei Han. Beyond Text-Dominance: Understanding Modality Preference of Omni-modal Large Language Models, April 2026. URL http://arxiv.org/abs/2604.16902. arXiv:2604.16902 [cs.AI].

Jeong Hun Yeo, Hyeongseop Rha, Se Jin Park, and Yong Man Ro. MMS-LLaMA: Efficient LLM-based Audio-Visual Speech Recognition with Minimal Multimodal Speech Tokens. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 20724–20735, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1065. URL https://aclanthology.org/2025.findings-acl.1065/.

Chenshuang Zhang, Kyeong Seon Kim, Chengxin Liu, and Tae-Hyun Oh. SVHalluc: Benchmarking Speech-Vision Hallucination in Audio-Visual Large Language Models, May 2026. URL http://arxiv.org/abs/2606.02642. arXiv:2606.02642 [eess.AS].

Ting Zhou, Daoyuan Chen, Qirui Jiao, Bolin Ding, Yaliang Li, and Ying Shen. HumanVBench: Probing Human-Centric Video Understanding in MLLMs with Automatically Synthesized Benchmarks, April 2026. URL http://arxiv.org/abs/2412.17574. arXiv:2412.17574 [cs.CV].

A EVALUATION DETAILS

Experiments used NVIDIA L40S, H100 (MIG partitions) and H200 GPUs. Open-model inference used bf16 precision and the supplied input processors. Hosted models were accessed through their APIs; their names do not identify immutable checkpoints.

A.1 STIMULI AND VISUAL CONTROLS

GRID words are cut from released alignments with 80 ms padding. Talkers s5, s6 and s10 were excluded because of alignment or filename mismatches. The silent-word set contains 1,379 readable clips from s1–s4 and s7–s9; 21 s8 clips lack readable video. Native speech-in-noise cells use 400 items from s1–s4 and 300 from s7–s9. Their maskers are white noise filtered to the clip's long-term spectrum and scaled to its whole-clip signal level. Cross-dubs and the bridge use the constructions in Appendices C and D.

Repeated-frame controls repeat frame 0; permuted controls use a fixed permutation of all decoded frames. Both preserve frame count and source frame rate, with libx264 encoding at CRF 16. Syllable controls preserve the audio track; cross-dubs use one canonical waveform per token. For native Qwen2.5-Omni clips, video-token counts match in all 1,200 checked intact/control triples. Matching counts does not remove all articulation information: a still mouth shape or unordered mouth configurations may remain informative.

Gemini's GRID inputs add 0.75 s of silence and cloned endpoint frames on each side, with task-specific prompts, because unpadded word recognition was poor. Syllable cells use the released unpadded clips. This API padding is distinct from the masker-only tail introduced by bridge reconstruction.

The McGurk release (Magnotti et al., 2020) contains 20 auditory /ba/ tokens paired with one visual /ga/ articulation. Three congruent /ba/, /da/ and /ga/ recordings come from a second talker. Offset tests add 500 ms silent pads and shift the speech within them; endpoint video frames are cloned. Every offset retains the same speech samples and duration. Positive offsets mean video leads audio.

A.2 PROMPTS AND SCORING

The canonical prompt is "You will perceive a person saying a single syllable. Answer with exactly one of: ba, da, ga. Answer:". GRID runs replace the choices with *bin, lay, place, set*; MiniCPM-o also replaces "syllable" with "word". Gemini receives "This short audio clip contains one English word spoken by a person. Which word is it?" with the options, or the corresponding "short video clip" wording. Open-model scores sum first-token forms with and without a leading space by log-sum-exp, then renormalise over the choice set. Their argmax and Gemini's generated answer define categorical outcomes. Appendix A.4 reports format sensitivity and choice-set checks.

Auto-AVSR transcripts are scored by the first exact option word, nearest option by letter edit distance (ties become OTHER), or initial-letter class: B/P/M versus L/S/T/D/N/Z. Full-denominator onset scores treat unclassifiable transcripts as non-matches. Conditional onset scores exclude them and have outcome-dependent coverage (Appendix C); the two measures must not be interchanged.

A.3 UNCERTAINTY AND NUMERICAL REPEATABILITY

Primary differences are paired within clips or auditory tokens. Percentile bootstrap intervals weight the named unit equally. McGurk contrasts use 10,000 resamples of the 20 auditory tokens. GRID item/token intervals and intervals from resampling talker means target different populations, especially with unequal talker sample sizes. Single-talker syllable cells do not support population-level equivalence claims. Probe permutation procedures are specified in Appendix F.2.

Repeated inference on fixed hardware gave identical probabilities and answers for Qwen2.5-Omni, Qwen3-Omni and MiniCPM-o 2.6. In a matched cross-hardware comparison, changes in $p(\mathrm{da})$ had median 0.0065 and maximum 0.032, with one of 67 answers changing; a separate baseline differed by 0.045. These checks describe numerical sensitivity in the tested cells, not a universal effect-size threshold.

A.4 Prompt and scoring sensitivity

For Qwen2.5-Omni, the McGurk 2-to-30 frames/s rise is $+0.448$ with the canonical prompt, $+0.545$ with reversed options, $+0.155$ with a fourth option (*ta*), and $+0.246$ with "What syllable did you perceive? Choose one of: ba, da, ga. Answer:". The foil absorbs 0.29–0.60 of the choice-set mass. The rise retains its sign but depends strongly on response format; a full Latin square of orders was not run. On 400 GRID words, changing "syllable" to "word" leaves the -6 dB AV-minus-AO loss at -0.280 versus -0.275, but changes the quiet loss from -0.018 to -0.150.

In the 207-cell core reruns, the first-token choice set contains median mass 0.992–0.998 for Qwen2.5-Omni and 0.997–1.000 for Qwen3-Omni; its argmax always matches the unrestricted top token. Logged word-task medians are 0.997, 0.999 and 0.947 for Qwen2.5-Omni, Qwen3-Omni and MiniCPM-o 2.6, respectively; MiniCPM-o's minimum is 0.529. MiniCPM-o 4.5's median is 0.999 (minimum 0.946). These audits do not replace an unrestricted-transcription evaluation of the omni models. Re-running syllable audio-only cells from bit-exact 48 kHz demuxed tracks changes Qwen2.5-Omni's McGurk $p(\mathrm{ba})$ by 0.009 on average (maximum 0.023), compared with the standard 16 kHz audio-only path.

B Silent words and native speech in noise

The common-cohort accuracy intervals and matched noisy-speech contrasts are in Tables 2 and 3. Here we give the response biases and coverage details needed to interpret them.

Silent-word biases. Silent-word responses are strongly biased: Qwen2.5-Omni answers *set* on 66% and *lay* on 34% of clips; MiniCPM-o 2.6 always answers *place*, and Flash always answers *set*. Pro answers *set*, *place* and *lay* on 84%, 11% and 5% of clips. Its 0.273 accuracy has item interval $[0.249, 0.297]$, with the excess concentrated in talkers 2, 3, 4 and 7. The talker-mean interval $[0.256, 0.291]$ answers a different sampling question.

Quiet-speech controls. On 400 words from talkers 1–4, MiniCPM-o 2.6 at $k = 1$ gains $+0.180$ over audio-only with intact video and $+0.185$ with either repeated or permuted frames. On talkers 7–9 the corresponding gains are $+0.147$, $+0.150$ and $+0.144$. Phi-4-multimodal's quiet accuracies are 0.603 audio-only, 0.740 intact, 0.807 repeated and 0.813 permuted. These controls reproduce or exceed the intact-video gains.

B.1 Hosted-model coverage and repeat check

Flash's native noisy-word comparison uses 300 paired clips and 298 with complete visual controls; six of 4,200 requests were rejected as corrupted video. Its 0 dB AV-minus-AO estimate is positive for each of talkers 7–9.

An interleaved repeat returned the original answer on 98.4% of 623 answered requests. On 155 clips complete in both runs, intact-minus-repeated was $+0.129$ $[+0.071, +0.194]$ and intact-minus-permuted $+0.123$ $[+0.052, +0.194]$, versus $+0.135$ for each originally. AV-minus-AO was $+0.052$ $[-0.013, +0.116]$. The visual-control contrasts persisted, but the audio-only comparison remained uncertain on this smaller cohort.

Pro's AO/AV accuracies on 300 clips are 0.520/0.640 at -6 dB and 0.757/0.820 at 0 dB. Its four-condition intersection contains only 81 talker-7 clips per SNR. Intact-minus-repeated is $+0.037$ $[-0.012, +0.098]$ and $+0.136$ $[+0.062, +0.222]$; intact-minus-permuted is $+0.111$ $[+0.037, +0.185]$ and $+0.062$ $[-0.037, +0.148]$. Pro's remaining controls, cross-dubs and bridge were not completed under the request cap.

C Waveform-matched cross-dubbing

The analysis uses 329 auditory tokens and same-talker donors. Each token's audio is padded to a common timeline; donor videos repeat their last frame to match. Repeated and permuted controls accompany every incongruent donor, with an audio-only reference (3,566 clean files). Noisy versions

reuse one masker per token and SNR across all visual conditions (7,132 files). Decoding verifies one audio hash per token, or per token and SNR, with no violations. An earlier build truncated audio to donor duration and is excluded from all reported cross-dub estimates.

Tables 4–6 report categorical and probability outcomes with token-bootstrap intervals. As a separate sensitivity check, intervals from seven talker means lie within $\pm.006$ for intact probability identity effects and $\pm.012$ for congruency effects in the open models. The specialist's conditional onset identity interval is $[+0.040, +0.094]$ on 322 classifiable tokens.

Weighting and coverage. Contrasts average donor pairs within a token and then weight tokens equally. Congruent donors are alternate utterances for 317 tokens, the original recording for 7, and unavailable for 5. Specialist conditional onset scoring excludes pairs with either transcript outside the onset classes, retaining 322/329 tokens clean, 142 at -6 dB and 254 at 0 dB; at -6 dB its repeated and permuted conditions retain only 131 and 94. Full-denominator scoring treats unclassifiable transcripts as non-matches and retains all 329 tokens. Thus control comparisons share a denominator only under the latter convention. Within-onset donor pairs cannot produce an onset identity effect; on between-onset pairs, the specialist's full-denominator effect is $+0.106$ clean and $+0.387$ at -6 dB.

Donor sensitivity. Across incongruent intact donors, the categorical answer changes on 9, 2, 29, 3, 32 and 36 of 329 tokens for Qwen2.5-Omni, Qwen3-Omni, MiniCPM-o 2.6, MiniCPM-o 4.5, Gemma 3n and Phi-4-multimodal. Median maximum pairwise total-variation distances are 0.029, 0.007, 0.035, 0.012, 0.102 and 0.054. Small signed effects therefore do not imply invariant output. Donors also differ in pose, timing and non-phonetic appearance.

Noise level convention. The noisy remix scales the masker to the standard deviation of the token's samples with magnitude above 10^{-4}, whereas the native speech-in-noise sets scale to the whole-clip RMS; over the 329 canonical waveforms the nominal v2 SNR is a median 0.76 dB lower (range 0.2 to 3.4 dB) than the whole-clip convention would give. Within-token matching is unaffected; comparisons between the native and the dubbed tasks at the same nominal SNR are approximate.

C.1 Specialist comparison

Auto-AVSR (Ma et al., 2023) uses conformer encoders trained on LRS3, VoxCeleb2 and AVSpeech, mouth-region preprocessing and free transcription. On 1,379 silent clips, strict option-word accuracy is 0.033, nearest-option accuracy 0.311 $[0.287, 0.336]$, and conditional onset accuracy 0.916 $[0.898, 0.933]$ on 973 classifiable transcripts. The native audiovisual comparisons below use 300 clips from talkers 7–9 and item-bootstrap intervals.

SNR	AV nearest	AV − AO (nearest)	AV − AO (onset class)	intact − static (nearest)	intact − permuted (nearest)
−12 dB	0.133	+0.133 [+0.097, +0.173]	+0.207 [+0.163, +0.253]	—	—
−6 dB	0.223	+0.140 [+0.087, +0.197]	+0.207 [+0.147, +0.267]	+0.133 [+0.080, +0.183]	+0.168 [+0.114, +0.218]
0 dB	0.557	+0.030 [−0.033, +0.093]	+0.113 [+0.047, +0.180]	+0.073 [+0.017, +0.130]	+0.185 [+0.128, +0.245]
quiet	0.850	−0.017 [−0.050, +0.017]	−0.003 [−0.027, +0.020]	+0.097 [+0.050, +0.143]	+0.117 [+0.064, +0.171]

On noisy cross-dubs, intact video produces larger full-denominator onset identity effects than either control (Table 6). Permutation leaves a residual effect, consistent with information retained in unordered mouth configurations. Conditional scoring gives larger estimates on fewer tokens.

D Native-versus-dubbed bridge

The bridge contains 96 tokens selected before inspecting outcomes: eight per word from talkers 7–9, each with an alternate-utterance congruent donor and a donor from the other onset class. At -6 and 0 dB, one masker is scaled using the unpadded speech's whole-clip RMS and reused across padded conditions. These are audio-only (AO), reprocessed native video, its repeated and permuted controls, a same-word alternate utterance, and an incompatible donor. Unpadded native video and its audio-only reference (AO_n) use the identical waveform prefix. Donors are onset-aligned through the word cuts' 80 ms pre-onset padding. All 1,536 files pass decoded-audio hash checks; hosted

Table 4: Clean waveform-matched cross-dubs. Panel A gives congruency and identity effects; bracketed values are 95% token-bootstrap intervals. Panel B gives contributing tokens/pairs for intact identity and changes relative to audio-only. Word outcomes are categorical; Auto-AVSR also has full-denominator onset scores (unclassifiable transcripts count as non-matches) and conditional scores (unclassifiable pairs excluded). Conditional control scores can use different subsets. There are 329 auditory tokens and 1,910 available ordered pairs; 324 tokens have a congruent donor. ‡Hosted categorical answers. §Specialist transcript scoring. See Appendix C for construction and coverage.

A. Congruency and identity

Model / outcome	Congruency	Identity: intact	Repeated	Permuted
Qwen2.5-Omni	-0.007 $[-0.018, +0.003]$	-0.002 $[-0.006, +0.003]$	$+0.002$	$+0.002$
Qwen3-Omni	-0.003 $[-0.008, +0.000]$	$+0.002$ $[+0.000, +0.004]$	-0.001	$+0.001$
MiniCPM-o 2.6	$+0.018$ $[-0.003, +0.038]$	$+0.000$ $[-0.007, +0.007]$	-0.005	-0.008
MiniCPM-o 4.5	$+0.001$ $[-0.005, +0.007]$	-0.001 $[-0.004, +0.003]$	-0.001	$+0.000$
Gemma 3n E4B	-0.007 $[-0.029, +0.014]$	-0.005 $[-0.014, +0.003]$	$+0.012$	-0.006
Phi-4-multimodal	-0.024 $[-0.043, -0.006]$	$+0.001$ $[-0.009, +0.011]$	$+0.009$	-0.004
Gemini 2.5 Flash‡	$+0.012$ $[-0.005, +0.028]$	-0.005 $[-0.012, +0.001]$	-0.003	-0.004
Auto-AVSR, nearest option§	$+0.038$ $[+0.011, +0.066]$	$+0.006$ $[-0.002, +0.013]$	$+0.000$	-0.002
Auto-AVSR, onset full§	$+0.082$ $[+0.059, +0.106]$	$+0.071$ $[+0.052, +0.090]$	$+0.000$	-0.001
Auto-AVSR, onset conditional§	$+0.074$ $[+0.053, +0.099]$	$+0.067$ $[+0.048, +0.088]$	$+0.000$	$+0.000$

B. Coverage and video-minus-audio effects

Model / outcome	Tokens / pairs	Congruent / incongruent	Repeated / permuted
Qwen2.5-Omni	329 / 1,910	-0.009 / -0.003	-0.006 / -0.009
Qwen3-Omni	329 / 1,910	-0.003 / $+0.000$	-0.002 / $+0.002$
MiniCPM-o 2.6	329 / 1,910	$+0.015$ / -0.000	-0.041 / -0.009
MiniCPM-o 4.5	329 / 1,910	$+0.022$ / $+0.020$	$+0.020$ / $+0.021$
Gemma 3n E4B	329 / 1,910	-0.593 / -0.585	-0.567 / -0.561
Phi-4-multimodal	329 / 1,910	$+0.046$ / $+0.072$	$+0.062$ / $+0.062$
Gemini 2.5 Flash‡	329 / 1,910	$+0.003$ / -0.007	-0.017 / -0.017
Auto-AVSR, nearest option§	329 / 1,910	$+0.040$ / $+0.002$	$+0.021$ / $+0.024$
Auto-AVSR, onset full§	329 / 1,910	$+0.028$ / -0.056	$+0.015$ / $+0.001$
Auto-AVSR, onset conditional§	322 / 1,820	$+0.013$ / -0.059	$+0.003$ / -0.003

Table 5: Cross-dub contrasts using renormalised option probabilities in the open-weight models. The probability statistic was fixed before the repaired stimuli were analysed. Panel A reports congruency and identity, with 95% token-bootstrap intervals; panel B gives intact-identity coverage and video-minus-audio effects. See Table 4 for categorical outcomes.

A. Congruency and identity

Model / SNR	Congruency	Identity: intact	Repeated	Permuted
Qwen2.5-Omni	$+0.000$ $[-0.004, +0.005]$	$+0.001$ $[-0.001, +0.002]$	$+0.000$	$+0.000$
Qwen3-Omni	-0.000 $[-0.002, +0.002]$	-0.000 $[-0.001, +0.000]$	$+0.000$	$+0.000$
MiniCPM-o 2.6	$+0.002$ $[-0.001, +0.004]$	-0.001 $[-0.002, -0.000]$	$+0.002$	-0.001
MiniCPM-o 4.5	-0.002 $[-0.006, +0.001]$	-0.000 $[-0.001, +0.001]$	-0.002	-0.000
Gemma 3n E4B	$+0.002$ $[-0.006, +0.010]$	-0.001 $[-0.004, +0.003]$	$+0.011$	-0.003
Phi-4-multimodal	-0.001 $[-0.005, +0.004]$	$+0.002$ $[-0.001, +0.004]$	$+0.006$	$+0.001$
Qwen2.5-Omni, -6 dB	-0.002 $[-0.008, +0.004]$	-0.001 $[-0.004, +0.002]$	-0.004	-0.002
Qwen2.5-Omni, 0 dB	$+0.000$ $[-0.005, +0.005]$	-0.001 $[-0.003, +0.001]$	-0.001	$+0.000$
Qwen3-Omni, -6 dB	$+0.004$ $[-0.003, +0.009]$	$+0.001$ $[-0.002, +0.004]$	$+0.005$	-0.000
Qwen3-Omni, 0 dB	$+0.006$ $[-0.000, +0.011]$	-0.000 $[-0.002, +0.001]$	-0.001	$+0.001$
MiniCPM-o 2.6, -6 dB	-0.001 $[-0.003, +0.001]$	$+0.000$ $[-0.001, +0.001]$	$+0.004$	-0.000
MiniCPM-o 2.6, 0 dB	$+0.000$ $[-0.002, +0.003]$	-0.001 $[-0.002, +0.001]$	$+0.004$	-0.001

B. Coverage and video-minus-audio effects

Model / SNR	Tokens / pairs	Congruent / incongruent	Repeated / permuted
Qwen2.5-Omni	329 / 1,910	$+0.036$ / $+0.034$	$+0.037$ / $+0.037$
Qwen3-Omni	329 / 1,910	$+0.005$ / $+0.006$	$+0.007$ / $+0.006$
MiniCPM-o 2.6	329 / 1,910	$+0.023$ / $+0.021$	$+0.013$ / $+0.021$
MiniCPM-o 4.5	329 / 1,910	$+0.080$ / $+0.081$	$+0.070$ / $+0.083$
Gemma 3n E4B	329 / 1,910	-0.454 / -0.453	-0.449 / -0.438
Phi-4-multimodal	329 / 1,910	-0.018 / -0.016	-0.020 / -0.017
Qwen2.5-Omni, -6 dB	329 / 1,910	-0.193 / -0.193	-0.182 / -0.182
Qwen2.5-Omni, 0 dB	329 / 1,910	-0.124 / -0.124	-0.108 / -0.113
Qwen3-Omni, -6 dB	329 / 1,910	-0.110 / -0.115	-0.105 / -0.106
Qwen3-Omni, 0 dB	329 / 1,910	-0.024 / -0.029	-0.015 / -0.025
MiniCPM-o 2.6, -6 dB	329 / 1,910	-0.014 / -0.013	-0.015 / -0.013
MiniCPM-o 2.6, 0 dB	329 / 1,910	-0.020 / -0.020	-0.026 / -0.020

inputs also receive API padding. Table 7 reports paired token-bootstrap comparisons, separating donor compatibility, reprocessing and presentation duration.

Comparison with the earlier experiment. Only 45 of the 96 bridge tokens overlap Flash's earlier word set. Table 7 therefore does not provide a paired replication on that overlap. Noise realisation, trailing masker duration and separate API runs leave the source of the discrepancy unresolved.

Within the bridge, the padded audio-only reference improves Flash's accuracy over its unpadded prefix by $+0.104$ $[+0.021, +0.198]$ and $+0.115$ $[+0.031, +0.198]$ at -6 and 0 dB. This isolates presentation sensitivity within the follow-up; it does not explain the earlier video benefit by itself.

D.1 Masker sensitivity

Token-specific speech-shaped noise could retain cues to the source word through its spectral envelope. We compared it with noise shaped by the average spectrum of 300 clean words, using 120 words from talkers 7 to 9 at -6 and 0 dB. Under the corpus-average masker, Qwen2.5-Omni's audio-only accuracy is 0.967 and 0.975, compared with 0.750 and 0.983 under the token-specific masker. Its AV-minus-AO remains negative: -0.150 $[-0.217, -0.092]$ and -0.058 $[-0.100, -0.017]$, compared with -0.250 and -0.167.

Flash's audio-only accuracy is 0.800 and 0.875 with the corpus-average masker, versus 0.483 and 0.633 with the token-specific masker. AV-minus-AO is $+0.017$ $[-0.058, +0.092]$ and 0.000 $[-0.067, +0.067]$ with the corpus masker, versus -0.075 $[-0.158, 0.000]$ and $+0.117$ $[+0.050, +0.192]$. The maskers therefore produce different acoustic difficulty at the same nominal SNR. These results establish masker sensitivity but do not exclude spectral cue leakage or determine whether Flash's gain would reappear with the corpus masker at a lower SNR.

Table 6: Noisy waveform-matched cross-dubs. Panel A gives congruency and identity effects; bracketed values are 95% token-bootstrap intervals. Panel B gives contributing tokens/pairs for intact identity and changes relative to audio-only. Word outcomes are categorical; Auto-AVSR also has full-denominator onset scores (unclassifiable transcripts count as non-matches) and conditional scores (unclassifiable pairs excluded). Conditional control scores can use different subsets. There are 329 auditory tokens and 1,910 available ordered pairs; 324 tokens have a congruent donor. ‡Hosted categorical answers. §Specialist transcript scoring. See Appendix C for construction and coverage.

A. Congruency and identity

Model / outcome	Congruency	Identity: intact	Repeated	Permuted
Qwen2.5-Omni, −6 dB	+0.004 [−0.018, +0.026]	−0.017 [−0.032, −0.001]	+0.006	−0.010
Qwen2.5-Omni, 0 dB	−0.003 [−0.021, +0.015]	−0.006 [−0.018, +0.005]	−0.003	−0.011
Qwen3-Omni, −6 dB	−0.007 [−0.033, +0.018]	+0.011 [−0.006, +0.028]	+0.018	−0.022
Qwen3-Omni, 0 dB	+0.002 [−0.015, +0.019]	−0.003 [−0.013, +0.006]	−0.003	+0.002
MiniCPM-o 2.6, −6 dB	−0.002 [−0.009, +0.003]	−0.001 [−0.002, +0.000]	−0.003	+0.000
MiniCPM-o 2.6, 0 dB	−0.009 [−0.025, +0.008]	−0.001 [−0.008, +0.005]	+0.004	−0.003
Gemini 2.5 Flash, −6 dB‡	−0.022 [−0.049, +0.004]	+0.012 [−0.004, +0.028]	−0.012	+0.002
Gemini 2.5 Flash, 0 dB‡	−0.019 [−0.044, +0.004]	−0.005 [−0.017, +0.008]	+0.001	+0.001
Auto-AVSR nearest, −6 dB§	+0.182 [+0.134, +0.232]	+0.081 [+0.056, +0.107]	−0.002	+0.005
Auto-AVSR onset full, −6 dB§	+0.306 [+0.254, +0.358]	+0.260 [+0.232, +0.290]	+0.004	+0.049
Auto-AVSR onset conditional, −6 dB§	+0.487 [+0.417, +0.554]	+0.549 [+0.484, +0.615]	+0.008	+0.121
Auto-AVSR nearest, 0 dB§	+0.170 [+0.119, +0.219]	+0.048 [+0.027, +0.068]	+0.002	−0.001
Auto-AVSR onset full, 0 dB§	+0.387 [+0.343, +0.429]	+0.291 [+0.260, +0.322]	+0.001	+0.042
Auto-AVSR onset conditional, 0 dB§	+0.370 [+0.328, +0.412]	+0.402 [+0.356, +0.448]	+0.001	+0.052

B. Coverage and video-minus-audio effects

Model / outcome	Tokens / pairs	Congruent / incongruent	Repeated / permuted
Qwen2.5-Omni, −6 dB	329 / 1,910	−0.219 / −0.228	−0.198 / −0.209
Qwen2.5-Omni, 0 dB	329 / 1,910	−0.114 / −0.116	−0.075 / −0.089
Qwen3-Omni, −6 dB	329 / 1,910	−0.083 / −0.078	−0.072 / −0.070
Qwen3-Omni, 0 dB	329 / 1,910	+0.015 / +0.015	+0.009 / +0.001
MiniCPM-o 2.6, −6 dB	329 / 1,910	−0.028 / −0.025	−0.029 / −0.027
MiniCPM-o 2.6, 0 dB	329 / 1,910	−0.099 / −0.092	−0.121 / −0.102
Gemini 2.5 Flash, −6 dB‡	329 / 1,910	−0.056 / −0.033	−0.038 / −0.019
Gemini 2.5 Flash, 0 dB‡	329 / 1,910	−0.040 / −0.020	−0.027 / −0.026
Auto-AVSR nearest, −6 dB§	329 / 1,910	+0.210 / +0.029	+0.076 / +0.054
Auto-AVSR onset full, −6 dB§	329 / 1,910	+0.262 / −0.049	+0.067 / −0.031
Auto-AVSR onset conditional, −6 dB§	142 / 536	+0.225 / −0.291	+0.036 / −0.045
Auto-AVSR nearest, 0 dB§	329 / 1,910	+0.068 / −0.104	+0.077 / +0.044
Auto-AVSR onset full, 0 dB§	329 / 1,910	+0.130 / −0.260	−0.011 / −0.106
Auto-AVSR onset conditional, 0 dB§	254 / 1,188	+0.118 / −0.253	+0.020 / −0.041

Table 7: Matched native-versus-dubbed experiment: 96 auditory tokens from talkers 7–9, eight per talker and word, at two SNRs. Values are paired differences; brackets give 95% token-bootstrap intervals. Native and AO_n use the unpadded waveform prefix; repro and AO use the padded timeline. Alt is another utterance of the same word; incompat is a donor from the other onset class. Repeated and permuted controls derive from repro. Onset scores retain the full denominator. ‡Hosted categorical answers with API padding.

A. Native and reprocessed benefit

System / outcome	SNR	Native − AO_n	Repro − AO
Auto-AVSR, nearest option	−6 dB	$+0.208$ $[+0.115, +0.302]$	$+0.281$ $[+0.188, +0.375]$
Auto-AVSR, nearest option	0 dB	$+0.167$ $[+0.031, +0.292]$	$+0.198$ $[+0.104, +0.302]$
Auto-AVSR, onset full	−6 dB	$+0.292$ $[+0.177, +0.396]$	$+0.240$ $[+0.125, +0.354]$
Auto-AVSR, onset full	0 dB	$+0.156$ $[+0.042, +0.271]$	$+0.156$ $[+0.052, +0.260]$
Qwen2.5-Omni	−6 dB	-0.156 $[-0.260, -0.052]$	-0.198 $[-0.302, -0.094]$
Qwen2.5-Omni	0 dB	-0.073 $[-0.135, -0.021]$	-0.073 $[-0.135, -0.021]$
Gemini 2.5 Flash‡	−6 dB	-0.021 $[-0.104, +0.062]$	-0.021 $[-0.094, +0.052]$
Gemini 2.5 Flash‡	0 dB	$+0.021$ $[-0.062, +0.104]$	$+0.021$ $[-0.031, +0.083]$

B. Alternate-donor comparisons

System / outcome	SNR	Alt − AO	Incompat − AO	Alt − incompat
Auto-AVSR, nearest option	−6 dB	$+0.260$	$+0.021$	$+0.240$ $[+0.135, +0.354]$
Auto-AVSR, nearest option	0 dB	$+0.229$	-0.073	$+0.302$ $[+0.188, +0.417]$
Auto-AVSR, onset full	−6 dB	$+0.177$	-0.208	$+0.385$ $[+0.281, +0.490]$
Auto-AVSR, onset full	0 dB	$+0.135$	-0.396	$+0.531$ $[+0.417, +0.646]$
Qwen2.5-Omni	−6 dB	-0.198	-0.146	-0.052 $[-0.104, +0.000]$
Qwen2.5-Omni	0 dB	-0.073	-0.062	-0.010 $[-0.062, +0.031]$
Gemini 2.5 Flash‡	−6 dB	-0.021	-0.021	$+0.000$ $[-0.052, +0.052]$
Gemini 2.5 Flash‡	0 dB	-0.031	-0.021	-0.010 $[-0.094, +0.062]$

C. Presentation and visual controls

System / outcome	SNR	Native − repro	Repro − repeated / permuted	AO − AO_n
Auto-AVSR, nearest option	−6 dB	-0.062	$+0.229$ / $+0.260$	-0.010
Auto-AVSR, nearest option	0 dB	$+0.010$	$+0.062$ / $+0.073$	-0.042
Auto-AVSR, onset full	−6 dB	-0.042	$+0.333$ / $+0.365$	$+0.094$
Auto-AVSR, onset full	0 dB	-0.052	$+0.156$ / $+0.229$	$+0.052$
Qwen2.5-Omni	−6 dB	$+0.083$	-0.042 / -0.031	-0.042
Qwen2.5-Omni	0 dB	$+0.000$	-0.042 / -0.021	$+0.000$
Gemini 2.5 Flash‡	−6 dB	-0.104	$+0.031$ / $+0.010$	$+0.104$
Gemini 2.5 Flash‡	0 dB	-0.115	$+0.010$ / $+0.042$	$+0.115$

E SYLLABLE RESPONSE CONTROLS

The McGurk set varies 20 auditory tokens while keeping one visual articulation. Increasing Qwen2.5-Omni's frame rate moves its response toward "da", also its silent-video default. Table 8 reports paired contrasts. The $+0.448$ intact rise decomposes into repeated frames at 30 minus intact at 2 ($+0.141$), permuted minus repeated at 30 ($+0.234$), and intact minus permuted at 30 ($+0.073$). These are intervention contrasts; they do not isolate causal contributions of token count, content and order.

At the same time, Qwen2.5-Omni's congruent /da/ and /ga/ scores fall from 0.969 to 0.809 and 0.806 to 0.242. MiniCPM-o 2.6's increase in $p(\mathrm{da})$ from $k = 1$ to $k = 15$ is larger for repeated frames ($+0.295$) than intact video ($+0.248$). Phi-4-multimodal instead shifts toward "ga" (0.11 at 2 frames/s, 0.74 at 25), with repeated frames yielding 0.72 and permuted frames 0.65. These shifts need not indicate human-like fusion. Flash answers "ba" on 18 of 20 McGurk items and "da" on two; Pro answers "ba" on all 20. Audio-only and visual-control responses are similar.

Table 8: McGurk contrasts with 95% intervals from 10,000 paired resamples of the 20 auditory tokens. Qwen rises compare 2 and 30 frames/s; MiniCPM-o contrasts use the stated k.

Model	Contrast	Estimate [95% CI]
Qwen2.5-Omni	intact rise	$+0.448$ $[+0.424, +0.471]$
	frozen rise	$+0.134$ $[+0.114, +0.153]$
	shuffled rise	$+0.360$ $[+0.335, +0.387]$
	intact − frozen at 30 frames/s	$+0.307$ $[+0.288, +0.326]$
	intact − shuffled at 30 frames/s	$+0.073$ $[+0.049, +0.097]$
	shuffled − frozen at 30 frames/s	$+0.234$ $[+0.215, +0.255]$
	frozen at 30 frames/s − intact at 2	$+0.141$ $[+0.119, +0.162]$
Qwen3-Omni	intact rise	$+0.001$ $[-0.049, +0.032]$
	frozen rise	$+0.024$ $[-0.020, +0.066]$
	shuffled rise	$+0.019$ $[-0.029, +0.055]$
	intact − frozen at 30 frames/s	-0.021 $[-0.048, +0.000]$
	intact − shuffled at 30 frames/s	-0.026 $[-0.047, -0.009]$
MiniCPM-o 2.6	intact − frozen, $k = 1$	-0.015 $[-0.041, +0.004]$
	intact − shuffled, $k = 1$	$+0.004$ $[-0.009, +0.017]$
	intact − frozen, $k = 15$	-0.061 $[-0.083, -0.041]$
	intact − shuffled, $k = 15$	-0.022 $[-0.039, -0.004]$
	rise $k = 1$ to 15: intact / frozen / shuffled	$+0.248$ / $+0.295$ / $+0.275$ (no CI)

E.1 TEMPORAL OFFSETS

Human temporal sensitivity depends on the stimulus, task and offset direction (van Wassenhove et al., 2007; Wallace & Stevenson, 2014). Our test preserves the speech samples while shifting them within silent pads (Table 9). None of the tested curves shows a clear synchrony peak. Because the synchronous cells do not first establish articulation-specific benefit, this finding gives limited evidence about temporal integration outside these inputs. Permuting frames is a different intervention: it preserves the frame set while changing order and motion.

F EXPLORATORY PROBES AND INTERVENTIONS

The analyses below support Section 5. Probe hidden-state index k denotes decoder-block output $k-1$; index 0 is the projector output. Interventions use zero-indexed block outputs. Only the corrected steering study, which matches estimation and injection indices, supports our steering conclusions.

F.1 ABLATION AND PATCHING

Mean ablation replaces all positions of one modality with their pooled mean at a block. The evaluation uses eight syllable items at 30 frames/s. Qwen2.5-Omni audio ablation changes $p(\text{audio syllable})$ by about -0.25 at blocks 0 and 13, -0.09 at 16 and -0.02 at 24. Early video ablation raises it: $+0.32$ $[+0.21, +0.39]$ at block 0 and $+0.35$ $[+0.25, +0.44]$ at block 13. Qwen3-Omni's early audio effects are larger and video effects smaller. These curves locate sensitivity to this perturbation, not a unique decision layer: information may already have propagated to other positions, and mean ablation preserves the pooled mean.

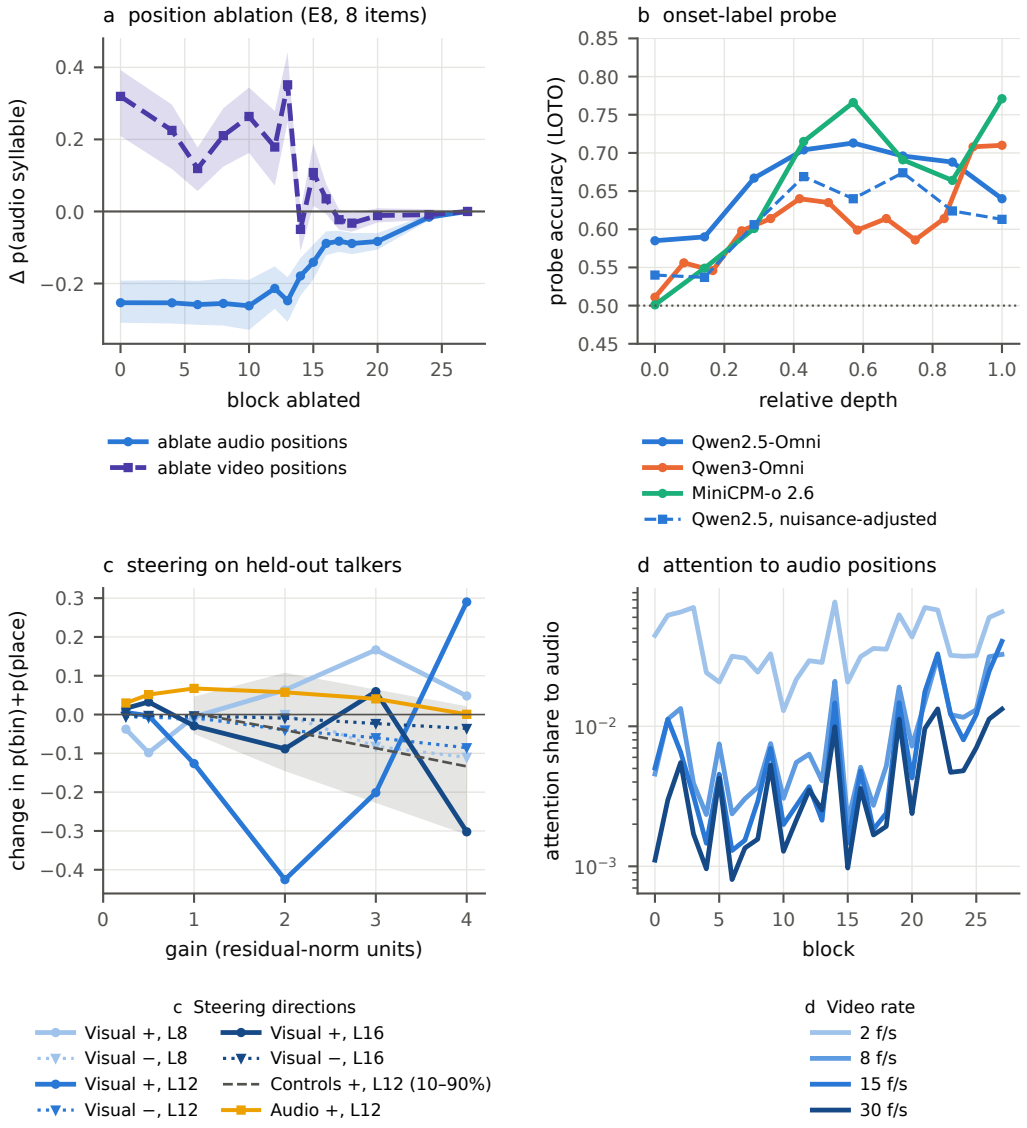


Figure 3: Exploratory internal analyses, Qwen2.5-Omni unless indicated. (a) Mean ablation on eight syllable items with 95% item-bootstrap intervals. (b) Held-out-talker onset probes: raw curves for three models and nuisance-adjusted Qwen2.5-Omni. (c) Steering on 48 noisy cross-dubs, with both signs, twenty norm-matched controls at block 12, and an audio-direction comparison. The band spans the 10th–90th percentile of direction-specific item means at matched gains. (d) Answer-position attention to audio positions across frame rates. Decodability, perturbation sensitivity and attention do not by themselves identify the model's natural computation.

Table 9: Asynchrony v2: audio padded with 500 ms of silence at both ends and shifted by the offset (positive: video leads), so every cell has identical audio content and duration. Congruent $p(\text{correct})$ (three talker-3 syllables) and McGurk $p(\text{da})$ (20 auditory tokens on one articulation). "Sync − ends" is the synchronous cell minus the mean of ± 400 ms with the 95% bootstrap across the 20 auditory tokens; a binding window predicts a positive value for McGurk $p(\text{da})$. Absolute Qwen2.5-Omni accuracy is lower than in the unpadded set (0.81) because the pads add cloned frames and silence.

Model, setting	Measure	-400	-200	-100	0	$+100$	$+200$	$+400$	Sync − ends
Qwen2.5-Omni, 2	congruent	0.636	0.617	0.566	0.580	0.474	0.358	0.729	-0.103
	McGurk $p(\text{da})$	0.334	0.324	0.346	0.388	0.446	0.427	0.565	-0.062 $[-0.098, -0.024]$
Qwen2.5-Omni, 30	congruent	0.476	0.435	0.427	0.436	0.368	0.397	0.564	-0.084
	McGurk $p(\text{da})$	0.695	0.668	0.702	0.748	0.760	0.756	0.819	-0.009 $[-0.026, +0.009]$
Qwen3-Omni, 2	congruent	0.982	0.992	0.981	0.983	0.988	0.974	0.988	-0.002
	McGurk $p(\text{da})$	0.050	0.057	0.060	0.064	0.067	0.063	0.120	-0.021 $[-0.042, +0.004]$
Qwen3-Omni, 30	congruent	0.972	0.989	0.981	0.979	0.899	0.936	0.966	$+0.010$
	McGurk $p(\text{da})$	0.029	0.031	0.037	0.037	0.047	0.034	0.143	-0.049 $[-0.060, -0.037]$
MiniCPM-o 2.6, $k = 1$	congruent	0.821	0.840	0.916	0.832	0.906	0.784	0.817	$+0.013$
	McGurk $p(\text{da})$	0.151	0.179	0.297	0.190	0.213	0.238	0.199	$+0.015$ $[-0.036, +0.070]$
MiniCPM-o 2.6, $k = 15$	congruent	0.711	0.700	0.692	0.635	0.669	0.650	0.634	-0.037
	McGurk $p(\text{da})$	0.379	0.353	0.447	0.391	0.408	0.449	0.413	-0.006 $[-0.053, +0.039]$

In the Qwen2.5-Omni patching check (25 frames/s, one pair per direction), audio /ba/ states inserted into an audio /ga/ input raise $p(\text{ba})$ from 0.062 to 0.968 at block 0, 0.890 at block 12 and 0.227 at block 16. Patching visual /ba/ into visual /ga/ changes a probability by at most 0.028; the reverse changes it by at most 0.008. The visual-only source and target baselines already differ by just 0.033, giving the visual test much less baseline contrast than the audio test. Neither intervention identifies token count as the cause of visual interference.

Panel (d) is descriptive: as video input grows from 782 to 14,076 tokens, block-0 attention to audio falls from 0.045 to 0.001 and attention to video rises from 0.071 to 0.213. These aggregate shares do not identify the information used.

F.2 Onset probes and nuisance adjustment

Raw probes use mean-pooled video-position states, StandardScaler, PCA with 128 components and logistic regression ($C = 0.1$), fitted within each leave-one-talker-out fold. Raw onset accuracies peak at 0.713 in Qwen2.5-Omni (1,379 clips, seven talkers), 0.710 in Qwen3-Omni and 0.771 in MiniCPM-o 2.6 (800 clips, four talkers each). Only Qwen2.5-Omni has the nuisance-adjusted onset audit below; the other two curves remain raw fits.

The Qwen2.5-Omni audit uses all 1,379 silent clips. Nuisance variables: clip duration measured per clip from the stream metadata (mean 0.481 s, SD 0.067) and the clip's start time within its sentence (the cut offset; available for 800 clips). Within each leave-one-talker-out fold, nuisance values are imputed with training means, a linear nuisance regression is fitted on the training features and applied to both partitions, and the scaler, PCA(128) and logistic classifier are fitted on the training partition; cross-word transfer fits every stage on the training word pair. Nuisance-only baselines (balanced accuracy): duration 0.592, start time 0.598, both 0.627 for the onset label; 0.312 for the word.

Hidden index	Onset raw	Duration adj.	Duration + start adj.	Word raw	Word adj.
0	0.586	0.573	0.540	0.344	0.290
8	0.667	0.609	0.606	0.362	0.324
12	0.704	0.674	0.669	0.389	0.331
16	0.713	0.650	0.640	0.368	0.340
20	0.696	0.661	0.674	0.415	0.388
24	0.688	0.624	0.624	0.442	0.384
28	0.640	0.629	0.613	0.415	0.391

Cross-word transfer trains on *bin/lay* and tests on *place/set*, or reverses the pairs, fitting every preprocessing stage on the training pair. Duration-and-start-adjusted accuracies range from 0.526 to 0.638 in the first direction and 0.542 to 0.662 in the reverse across the listed indices.

Observed and permuted labels use the same procedure. Labels are permuted within talker, and the test statistic is the maximum duration-and-start-adjusted onset accuracy over the full layer grid. With 99 permutations, the observed maximum is 0.674, the null mean 0.510 and the null 95th percentile 0.521; $p = (b + 1)/(B + 1) = 0.01$.

The adjustment has several limits. Start time is available for only 800 of 1,379 clips; imputing the other 579 does not remove their unmeasured start-time effects. Linear residualisation does not establish independence from nonlinear nuisance information. The 0.674 versus 0.627 comparison is not a test of an increment over the nuisance-only predictor. The transfer test uses only two fixed word pairs, so it cannot establish generalisation across a broad phonetic vocabulary. The onset permutation test adjusts for the specified layer search, not selection of the best word-probe layer. On the same clips, the model's silent-word accuracy is 0.251 and the adjusted word probe reaches about 0.39 at hidden index 28.

F.3 Steering and matched controls

Direction: class-mean contrast of mean video-position states (bilabial minus alveolar onset) from talkers 1 to 4 at feature index $L{+}1$, injected at the output of block L (matching estimation and injection indices). Items: 48 cross-dubbed GRID words at -6 dB from held-out talkers 7 to 9, one per talker, audio word and visual word; target quantity $p(\mathrm{bin}){+}p(\mathrm{place})$, baseline 0.607. Both signs are tested at gains 0.25, 0.5, 1, 2, 3 and 4, in units of the residual-state norm. Twenty norm-matched controls per block: five random, five orthogonal to the direction, five label-shuffled class means, five donor differences. Audio calibration: the analogous direction from audio positions (700 clean words) injected at audio positions.

Controls and choice-set mass at gain 4

Block	Control shifts: random, orthogonal, shuffled labels, donor (+/−)	Choice-set mass (+/−)
4	$-0.11/-0.04$, $-0.26/-0.01$, $-0.13/-0.12$, $-0.04/-0.17$	0.000 / 0.977
8	$-0.03/-0.24$, $-0.14/-0.24$, $-0.02/-0.14$, $-0.25/-0.18$	0.852 / 0.992
12	$-0.14/0.00$, $-0.16/-0.17$, $-0.10/-0.12$, $-0.13/-0.17$	0.996 / 0.995
16	$+0.10/+0.14$, $+0.06/-0.04$, $+0.05/+0.15$, $-0.08/+0.15$	0.146 / 0.995
20	$+0.03/+0.06$, $-0.07/+0.01$, $-0.04/+0.09$, $-0.10/-0.24$	0.001 / 0.996
24	$+0.20/+0.09$, $-0.04/+0.18$, $+0.10/+0.18$, $+0.17/+0.12$	0.968 / 0.996

Visual steering is non-monotonic: at block 12, positive gains 2 and 4 change bilabial probability by -0.43 and $+0.29$. Large gains can also push most probability outside the choice set (table above). The analogous audio direction moves probability in the expected directions at blocks 8 and 12, reaching $+0.16/-0.08$ and $+0.07/-0.15$, respectively, across the tested gains; its curves are also non-monotonic.

For the visual direction, the positive shift exceeds all twenty controls on 0 to 33% of held-out items at most tested high-gain settings. One setting reaches 71% (block 12, gain 4), but the same block's curve reverses direction at lower gains. The negative direction is not consistently more effective than controls. Natural per-token projections have means $+0.10$ to $+0.29$ residual norms and standard deviations 0.07 to 0.25.