VideoFDB: Evaluating Full-Duplex Vision-Speech Capabilities in Conversational Agents

Amrita Mazumdar[1], **Seonwook Park**[1], **Rajarshi Roy**[1], **Nikhil Srihari**[1], **Shengze Wang**[1], **Yuhao Zhou**[2] **Julia Wang**[2], **Koki Nagano**[1] **Shalini De Mello**[1]
[1]NVIDIA, [2]David AI

Abstract

Natural human conversation is full-duplex and audio-visual: people simultaneously speak and listen while continuously interpreting and producing nonverbal cues, such as nods, smiles, and gestures. To support successful human-agent interaction, agents must model full-duplex audiovisual conversation; however, existing full-duplex benchmarks evaluate only speech. In this work, we present VideoFDB, the first benchmark to evaluate full-duplex audio-visual-to-audio-visual (AV2AV) conversational agents. VideoFDB contributes (i) 237 dyadic clips spanning 11 nonverbal conversational dynamics from real-world video calls, (ii) a taxonomy separating *perception* from *generation* behaviors, and (iii) a rubric-based LM-as-judge evaluation framework with interpretable axes for assessing conversational quality with respect to nonverbal conversational dynamics. Across open- and closed-source vision-speech agents, we find systematic failure modes: *captioning collapse* and *visual-stream ignorance*, and we show that current systems exploit vision for explicit visual question answering but not for the streaming joint audio-visual grounding required in natural conversation. We further evaluate cascaded speech-to-avatar systems and find that their architecture fundamentally precludes the production of full-duplex nonverbal cues. As the first benchmark for full-duplex AV2AV interaction, VideoFDB establishes a foundation for systematic evaluation and, we hope, will accelerate the advancement and development of next-generation multimodal conversational agents.

1 Introduction

Advances in multimodal learning and conversational speech agents are transforming how humans interact with AI systems. As these systems move into embodied settings (e.g., service robots, collaborative agents in telepresence and XR), conversational speech interfaces will proliferate. Existing speech-to-speech agents can listen and respond to users, but often rely on turn-based interaction [56]. Natural human conversation, however, is not a sequence of voice turns: it is *full-duplex* and *audio-visual-to-audio-visual* (AV2AV), with **both participants simultaneously speaking, listening, and signaling** through continuous verbal and non-verbal channels [15, 11].[1] Such full-duplex, AV2AV conversational agents (vision-speech agents) are newly emerging. Recent advances, (e.g., Gemini 2.5/3.1 Live [8, 16], OpenAI Realtime [35], Qwen3-Omni [53], MoshiVis [44]) allow agents to perceive image or low-frame-rate video inputs (AV2A) while conversing with users interactively. Building agents that approach human-like interaction therefore requires (1) systems that simultaneously perceive and generate audio-visual cues continuously, without waiting for explicit turn boundaries, and (2) benchmarks that evaluate their multimodal conversational dynamics, not just the semantic correctness of their responses.

[1]We use the abbreviations AV2AV (audio-visual in/out), AV2A (audio-visual in / audio out — current vision-speech agents), A2A (audio-to-audio), and A2AV (audio in / cascaded avatar out) throughout the paper.

Preprint.

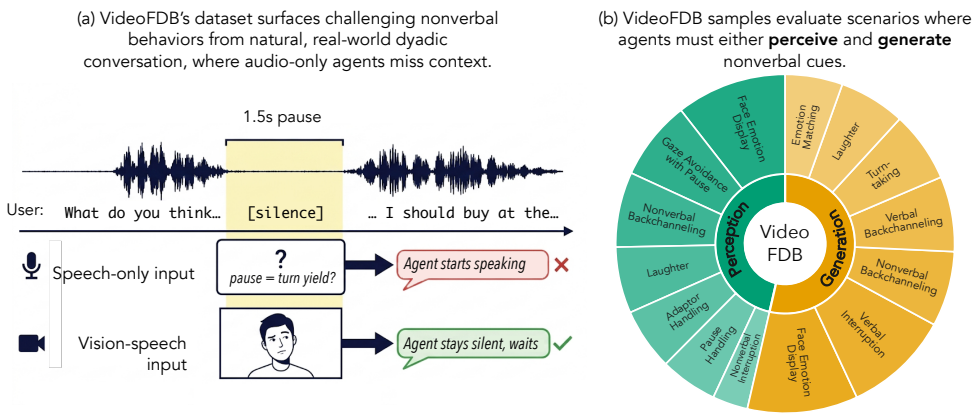

Figure 1: VideoFDB (a) curates evaluation samples from natural conversations, where visual cues carry key context, and (b) evaluates perception and generation capabilities of full-duplex audio-visual conversational agents across 11 categories, including dynamics that appear in both tracks.

However, no benchmark evaluates the full capabilities of full-duplex vision-speech agents. Existing full-duplex *speech-to-speech* benchmarks [36, 23, 22, 24] measure turn-taking, backchanneling, and interruption from audio alone, leaving gesture, gaze, and affect unmeasured. Conversely, audio-visual VLM benchmarks mostly evaluate *video question answering* [13, 3, 5, 7, 39, 54, 46], rewarding semantic correctness over conversational dynamics, while audio-driven avatar benchmarks [37, 26, 25] evaluate speaker and listener behaviors separately rather than via continuous joint interaction. Across these domains, no work evaluates the agent as an active interlocutor (i.e., conversational participant) that must continuously engage with human nonverbal dynamics.

Evaluation is intrinsically difficult because conversation is non-deterministic: the same user signal (e.g., laughter) can merit multiple valid responses depending on context and interaction style [50]. Correct assessment must therefore allow a distribution of valid behaviors, rather than a single correct response. To our knowledge, no publicly available full-duplex AV2AV systems exist, underscoring the need for further research.

In this paper, we present VideoFDB, the first benchmark evaluating the conversational dynamics of full-duplex audio-visual conversational agents, grounded in a taxonomy of nonverbal dynamics [15, 11]. VideoFDB treats the agent as an active conversational participant and evaluates responses along axes in two categories:

- **Perception**: (1) *Fluency*: whether the interaction remains coherent and conversationally natural; (2) *Conversational Flow*: whether the agent can successfully perceive nonverbal cues when managing turn timing and floor transitions appropriately (e.g., yielding, holding, interruption, and backchannel timing); and (3) *Semantic Grounding*: whether the agent's response behavior is semantically aligned with perceived nonverbal cues.
- **Generation** (if agent video is produced): (1) *Fluency*, as in the Perception category; (2) *Dyadic Affect Match*: whether the agent's combined AV response affectively corresponds with the user's affective state; and (3) *Nonverbal Cue Appropriateness*: whether produced nonverbal behaviors are category-appropriate and produced at appropriate moments in the conversation.

VideoFDB includes 11 conversational dynamics with 237 human-annotated audio-visual dyadic samples (Fig. 1), with evaluation protocols for both categories. We systematically evaluate leading open-source and proprietary full-duplex vision-speech agents along with cascaded speech-to-avatar systems, finding that current systems miss nonverbal cues, fall back to captioning-style replies, and degrade speech quality as user video sampling increases. Further, we observe that AV2A vs. A2A comparisons indicate visual input is used for captioning but ignored when visuals are not verbally referenced. Finally, we find cascaded speech-to-avatar pipelines preserve turn-yielding discipline but cannot insert cues during the user's turn, with latencies 2.8–3.5 s behind human ground truth.

To summarize, our main contributions are:

- We introduce VideoFDB, the first full-duplex vision-speech benchmark for jointly evaluating verbal and nonverbal conversational dynamics in natural dyadic conversation.

- We benchmark both open-source and proprietary full-duplex conversational speech agents and show that even state-of-the-art systems struggle to perceive and respond to natural nonverbal dynamics.
- We provide a failure-mode analysis that identifies *captioning collapse*, *visual-stream ignorance*, and *structural challenges in cascaded speech-avatar systems*, clarifying key directions towards advancing end-to-end audio-visual conversational agents.

2 Related Work

Full-duplex speech agents and omni-models. Full-duplex speech-to-speech methods natively model both sides of a dyadic exchange in a single network: Moshi [10] and dGSLM [33] jointly generate user and agent streams to support overlapping speech, with OmniFlatten [57], SyncLLM [48], SALM [17], and PersonaPlex [43] extending this template to richer turn-taking and role control. Recent work extends full-duplex speech-to-speech models: Gemini 2.5 and 3.1 Flash Live [16, 8], GPT Realtime [35, 18], Qwen2.5- and Qwen3-Omni [40, 53], and MoshiVis [44] all accept audio and video inputs jointly while preserving streaming speech output, and Video-SALMONN [45] adds video to streaming dialogue understanding. OmniResponse [29] extends multimodal language models (MLLM) to model dyadic conversation, including verbal and nonverbal cues; their method generates full-duplex audio-visual conversational behavior, but is only evaluated on uni-modal speech, visual, or text quality. None of these systems, to our knowledge, has been evaluated on its ability to interact with the nonverbal component of a full-duplex exchange, such as gesture, gaze, or facial signal.

Multimodal conversational benchmarks. Existing multimodal evaluations probe audio-visual understanding through paired prompts and text responses rather than continuous interaction. Captioning [13] and visual question answering (VQA) [3] methods directly evaluate visual understanding; AVERE [5], SAVVY [7], and MMPerspective [46] extend these evaluations to richer audio-visual reasoning, and recent methods target face-centric[39], real-time [54], and situated [38] VQA, where the model answers questions about the visible person or scene rather than participating as that person's conversational partner.

Table 1: Prior benchmarks isolate speech-to-speech interaction or turn-based/split-role vision-speech tasks.

Benchmark	Full-duplex?	Speech In?	Speech Out?	Video In?	Video Out?	Real-world?
Full-Duplex Bench [23]	✓	✓	✓	✗	✗	✗
EMO-Reasoning [27]	✗	✓	✓	✗	✗	✗
EmoRealm [5]	✗	✗	✗	✓	✗	✓
OmniMMI [49]	✗	✗	✗	✓	✗	✓
ResponseNet [29]	✗	✓	✓	✓	✓	✓
VideoFDB (Ours)	✓	✓	✓	✓	✓	✓

A complementary line [7, 42] specifically curates VQA pairs that probe cross-modal integration, but their evaluations are also turn-based.

Multimodal evaluation methods that focus on dyadic interaction (e.g., speaker or listener generation) commonly use *split-role* protocols that bifurcate the exchange into discrete speaker and listener roles: OmniMMI [49] measures proactive turn-taking against scripted prompts, [32] evaluates turn-based grounded visual reference resolution, and ResponseNet [29] scores agent listener behavior in dyadic conversation by separating speaker and listener turns rather than evaluating cross-role overlap. In these split-role settings, models are evaluated as either *speaker* or *listener* at a time, not as interlocutors continuously managing both speaker and listener turns in overlapping conversation. Across multimodal evaluation methods, success metrics score per-turn response quality, and, optionally, latency. No method evaluates continuous conversation or naturalness under full-duplex audio-visual settings.

Full-duplex speech evaluation. Full-duplex speech benchmarks such as Full-Duplex-Bench [36, 23, 22, 24] measure dynamics like turn-taking, interruption handling, and backchannel timing. EMO-Reasoning [27] evaluates emotion reasoning in speech without modeling full-duplex dynamics.

Avatar animation evaluation. Cascaded audio-driven avatar methods, where generated speech drives portrait animation in real time, are typically evaluated on visual realism and naturalness [6, 12, 52]. Benchmarks evaluating audio-driven gesture generation [37, 26, 25] also focus on motion realism and gesture fidelity in isolation of conversational semantics.

VideoFDB is, to our knowledge, the first benchmark to evaluate interaction-level AV2AV dynamics—gesture, gaze, facial signal, and dialogue cues—in genuinely full-duplex conversation, by: (1)

Table 2: VideoFDB conversational dynamics overview.

Nonverbal Communication Channels				Conversational Dynamic	Dynamic Description	Evaluation Category	
Dialogue	Eye Gaze	Face	Body			Perception	Generation
Perception-only dynamics							
✓	✓			Gaze Avoidance with Pause	Gaze aversion paired with a pause often indicating thinking or processing.	✓	
			✓	Adaptor Handling	Self-directed or reflexive action, such as coughing, yawning, scratching, or adjusting hair.	✓	
✓	✓			Pause Handling	Brief pause within a speaker's turn, typically for thinking or a small action.	✓	
✓			✓	Nonverbal Interruption	Interruption through gesture, facial expression, or other nonverbal behavior, optionally paired with speech.	✓	
Shared dynamics (Perception + Generation)							
	✓	✓		Face Emotion Display	Visible facial emotional expression during the interaction.	✓	✓
✓		✓		Nonverbal Backchanneling	Listener feedback delivered through facial expression, sometimes paired with speech.	✓	✓
✓		✓		Laughter	Laughter during the interaction.	✓	✓
Generation-only dynamics							
✓				Verbal Interruption	Spoken interruption while another participant is still talking.		✓
✓		✓		Verbal Backchanneling	Short listener feedback delivered through speech, sometimes paired with nonverbal behavior.		✓
✓				Turn-taking	Exchange of speaking roles between participants.		✓
		✓		Emotion Matching	Mirroring of the speaker's emotional expression by the listener.		✓

evaluating overlapping dyadic exchange with role switching; (2) covering a broad range of vision-speech dynamics; and (3) scoring interactional appropriateness (perception and generation) rather than only semantic correctness or uni-modal generation quality (Table 1).

3 VideoFDB Benchmark Dataset

VideoFDB evaluates the visual perception and visual generation capabilities of a conversational speech agent in a natural, full-duplex conversation with a benchmark dataset of conversational dynamics. VideoFDB's dataset includes dynamics identified in human communication research as necessary for fluent dyadic interaction [11], including dialogue-management cues (gaze, intonation, gesture completion), backchannels, interruption signals, affect displays, and body movements. This grounding ensures that VideoFDB measures competence on dynamics human interlocutors *actually rely on*, rather than arbitrary visual phenomena or direct visual question-answering.

Section 3.1 presents an overview of the conversational dynamics included in our dataset and what they evaluate, and Section 3.2 summarizes dataset statistics and the capture/annotation pipeline.

3.1 VideoFDB Conversational Dynamics Taxonomy

We organize our benchmark around a taxonomy of nonverbal channels that human communication science identifies as central to how interlocutors co-construct meaning [15, 11, 50]:

- **Dialogue**: While dialogue tends to refer to verbal dialogue cues, conversations also include paralinguistic or nonverbal dialogue cues: turn-taking depends on timing and backchannel behavior; longer gaps often carry pragmatic meaning; and acknowledgment tokens (e.g., *mm-hm*, *yeah*) mark attentive listening without claiming the floor [21, 30, 55]
- **Eye Gaze**: Beyond lexical content, gaze behavior contributes independent conversational information. Gaze direction and aversion regulate turn transitions and indicate social intent (e.g., glancing away may signal turn-yielding or a mid-thought pause) [11, 15, 50]
- **Face** and **Body**: Beyond spoken words, facial and bodily behavior carry additional conversational signals. Co-speech gestures (movements produced synchronously with speech), adaptors (self-directed actions like fidgeting), and facial affect displays provide complementary semantic and interpersonal cues that shape how utterances are interpreted [31, 2, 11]

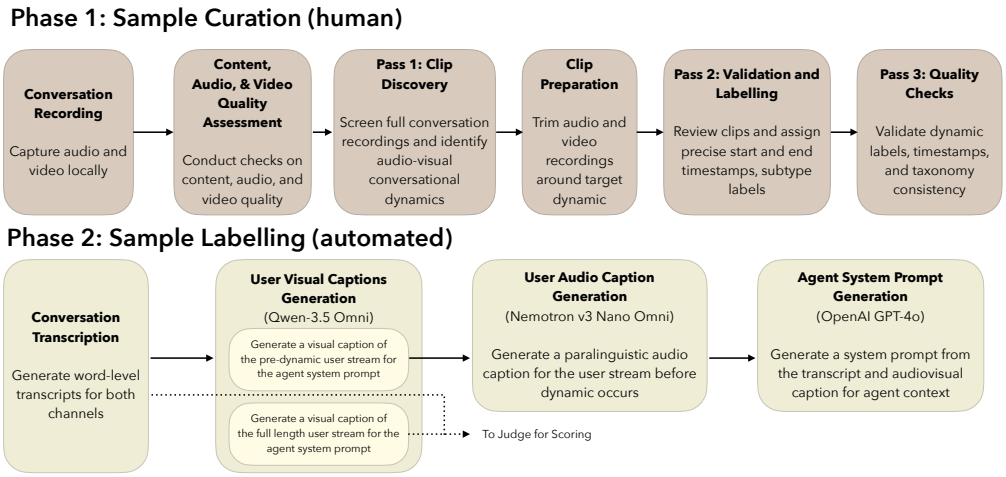

Figure 2: Annotation workflow. Human annotators select and curate clips over multiple passes of quality reviews. Automated captioning generates contextual prompts for agent inference and VideoFDB evaluation.

We operationalize this communication framework by defining, for each nonverbal channel, conversational dynamics that are observable, annotatable, and behaviorally relevant in dyadic interaction. Specifically, we select dynamics that directly govern conversational floor management and turn signaling in natural exchange. Table 2 describes the dynamics, how they map to nonverbal communication channels and indicates whether each dynamic is used in the **Perception** and/or **Generation** evaluations. Across the taxonomy, the selected dynamics cover floor management, listener feedback, social-affective signaling, and conversational body movement.

3.2 Dataset Statistics, Capture, and Annotation

VideoFDB contributes 237 clips from natural two-person video calls, spanning 11 dynamic types. Each sample contains diarized, speaker-separated audio/video streams for both participants and is centered on one annotated **conversational dynamic** with event window $[t_{\text{start}}, t_{\text{end}}]$, while preserving salient surrounding context (typically 1–3 lead-in turns and up to one follow-up turn). To support evaluation, recordings are filtered and manually annotated into a standardized label set. Figure 2 illustrates the end-to-end annotation process. Appendix A.1 provides additional dataset details.

Collection and annotation. Our dataset is mined from a held-out corpus of two-person video calls, in which participants converse naturally over a videoconferencing platform, with each speaker's stream recorded locally to mitigate network-latency artifacts. Source recordings are then screened with media-quality and human preprocessing checks. Final clips are curated with a three-pass human annotation pipeline (candidate discovery, timestamp/type validation with sufficient pre-event context, and final quality review). Appendix A.1 provides full data curation details.

After selection and annotation, each clip is supplemented with LM-generated labels: a system prompt for agent bootstrapping, a user-stream AV caption for judge context, and an event-window dynamic label. Section 4.3 describes the system prompt and AV caption pipelines.

4 Evaluation & Experimental Setup

Natural conversation is fundamentally non-deterministic: for most nonverbal events, there is no single "correct" response string. VideoFDB thus uses rubric-based LM-as-judge [58] scoring instead of exact-match targets, which allows us to rate multiple semantically valid responses as acceptable and score response quality on interpretable dimensions. In this section, we describe VideoFDB's evaluation protocol and scoring metrics.

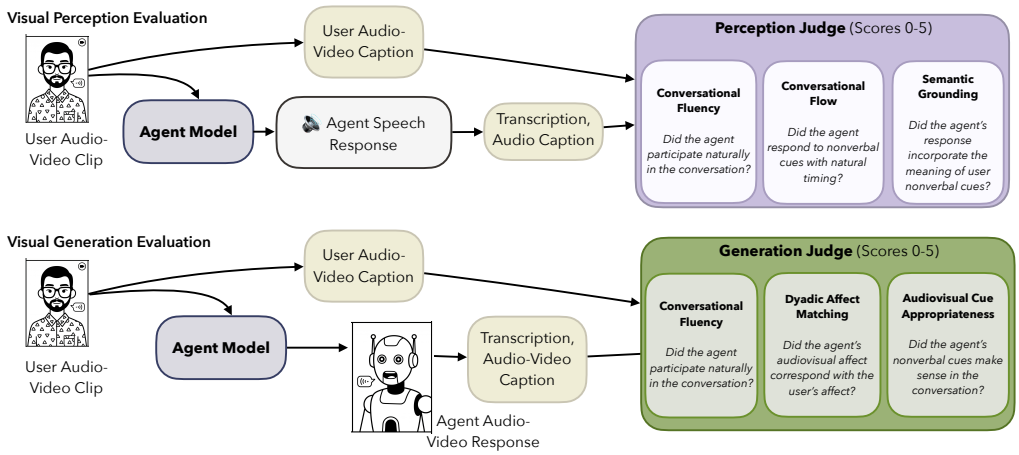

Figure 3: Evaluation flow. We feed agent-generated audio and/or video with annotated captions into judge prompts, returning category-specific scores.

4.1 Agent Perception & Generation Evaluations

We group conversational dynamics into two categories: *perception* dynamics test whether agents correctly interpret user-produced nonverbal behavior, while *generation* dynamics test whether agents produce appropriate nonverbal responses (only applicable to full-duplex agents that emit continuous visual output). We evaluate visual perception against three interpretable rubrics: **Fluency** grades overall interaction quality (talk-over, monologue, nonsensical responses); **Conversational Flow** grades timing of agent responses around a nonverbal cue; **Semantic Grounding** grades response content against visual-affective event semantics. The latter two apply only to relevant dynamics: Pause Handling rewards *not* taking the floor while Backchanneling rewards *continuing* (both under Flow); Nonverbal Interruption is scored along both axes (yield timing under Flow, post-yield behavior under Grounding).

For agents generating visual output, we evaluate: **Fluency** (as above); **Dyadic Affect Match**, whether the agent's combined audio+visual response affectively corresponds with the user's affective state; and **Nonverbal Cue Appropriateness**, whether produced nonverbal behaviors are category-appropriate and well-timed. Table 8 gives the full dynamic-by-axis breakdown.

4.2 Evaluation Scoring & Metrics

We now describe how we evaluate agent responses to our benchmark samples using VideoFDB's rubrics and metrics. The end-to-end evaluation protocol is summarized in Figure 3. For each clip, we supply a fixed system prompt and then stream the user audio/video channel to the agent in real time, and record the agent output stream. The recorded agent stream is transcribed with word-level timestamps, then merged with clip metadata (e.g., `dynamic_type`, `dynamic_start_s`/`dynamic_end_s`, AV caption) to construct the judge payload. The judge then scores each clip on the configured rubric axes and computes the deterministic timing metrics as follows.

Scoring "Appropriate" Conversational Behavior We score appropriateness with pre-specified, dynamic-specific rubric anchors. For each dynamic, the rubric defines the expected interaction policy around the annotated event window (e.g., *stay silent*, *continue speaking*, *yield*, *affectively align*) and maps clip evidence (timestamps, transcripts, and AV caption) to a shared 0–5 scale. Appendix F details full prompts and rubrics.

Timing Metrics. In addition to the rubric axes above that capture timing quality qualitatively, we also score timing behavior directly, leveraging our dynamic window annotations and the generated agent response stream's timestamps. To report unified timing scores across nonverbal cues with different expected turn-taking outcomes, we introduce **Takeover-Rate Alignment** (TOR-Alignment). TOR-Alignment extends the speech-only TOR formulation of Lin et al. [23] to multimodal AV dynamics by unifying heterogeneous timing expectations under one metric; higher TOR-Alignment indicates a larger fraction of samples satisfying expected timing behavior. We map timing-relevant dynamics

to one of five timing classes, defined by the expected agent behavior when the dynamic appears: STAY-SILENT, CONTINUE-SPEAKING, YIELD-REQUIRED, SMOOTH-HANDOFF, and BACKCHANNEL-PRODUCED. Appendix C.2 provides a full formalization.

Latency. We also measure the latency of the agent's response to the timing class, which distinguishes *stay-silent* from *continue-speaking*: both can involve short silences, but the expected behavior differs by conversational role (listener vs. current speaker).

4.3 Additional Annotations

System prompt preparation for agent evaluation. When a user event occurs amid an ongoing agent turn (e.g., nonverbal interruption), we condition the agent with a short system prompt summarizing pre-event context and explicitly triggering speech ("Start speaking now to begin/continue the conversation"). Appendix F.4 provides additional details.

Judge prompt construction & validation To provide agent responses to our judge for scoring, we construct a structured judge payload (event metadata, role-labeled transcripts, and AV captions), using Qwen3.5-397B [41] for visual captions and Nemotron v3 Nano Omni 30B [34] for audio captions; Appendix F provides full details. Our rubrics show stable cross-judge consistency: across three judge models, pairwise agreement is 77–89% within 1 point on the 0–5 scale (Appendix C.3). Following recent full-duplex speech evaluations [23, 43], we employ `gpt-4o` for scoring.

5 Experiments and Results

We evaluate VideoFDB on a diverse set of full-duplex models, spanning speech-vision, audio-only, and speech-to-avatar systems.

Closed- and open-source vision-speech models (AV2A). Our closed-source speech-vision baselines are Gemini 2.5 and 3.1 Flash Native [8] and OpenAI Realtime and Realtime mini [35]. Our open-source speech-vision baselines are MiniCPM-o 4.5 [9], MiniOmni2 [51], and VITA-1.5 [14]. MiniOmni2 only accepts one frame per user speech segment; for all other models, we stream user video at 1 FPS. We stream audio input at 16-24 kHz, depending on model requirements.

Closed- and open-source conversational speech models (A2A). To isolate the contribution of visual cues, we re-run all vision-speech agents in audio-only mode, comparing paired A2A and AV2A runs on the same clips.

Full-duplex speech models with cascaded avatar generation. Because no publicly available full-duplex AV2AV models exist, we evaluate a cascaded setup where a full-duplex conversational speech model (AV2A) runs in LiveKit Cloud [28] and its speech drives an audio-driven avatar (e.g., Anam [4], Keyframe [19]). We use a single full-duplex speech backbone (Gemini 2.5 Flash Native) for both avatars to isolate avatar effects from speech-backbone variation.

VideoFDB tests whether these systems respond to full-duplex user input with temporally aligned, affect-consistent, and category-appropriate visible conversational behavior (Tables 3 and 4).

5.1 Vision-perceiving speech models

***Insight 1*. Current models remain below human-level conversational naturalness.** Table 3 shows no model approaches human ground truth on VideoFDB; the largest deficits concentrate in fast social-coordination dynamics (Pause Handling, Nonverbal Backchanneling, Gaze Avoidance with Pause), with aggregate human–model gaps up to 0.85 (Table 11). Open-source MiniCPM-o-4.5 [9] leads in both AV and AO overall (3.40/3.44), with Gemini 2.5 Flash Native as runner-up. Conversational Flow shows the largest gap: humans score 4.20, closed-source AV2A systems score 2.20–2.81, and the strongest AV2A model reaches only 3.54. TOR-Alignment follows the same pattern: humans score 90%/1400 ms vs. next-best MiniCPM-o 4.5 at 73%/720 ms.

***Insight 2*. Limited visual frame rate prevents models from capturing nonverbal conversational signals.** Audio is processed at millisecond-scale temporal resolution, but AV2A models typically take visual input at 1 FPS, missing nonverbal dynamics that unfold within 1–2 seconds. Taking Nonverbal Interruption as an example, agents are expected to yield within 1.5 s, but Gemini 3.1 achieves only

Table 3: Performance breakdown across Perception rubrics for full-duplex speech-vision models. Timing reports TOR-Alignment percentage above and median latency below. Best and second-best non-human entries are marked in bold and underline, respectively.

Model	Fluency ↑	Conv. Flow ↑	Vis. Ground. ↑	Overall ↑	Timing ↑
Human reference	4.16 (3.88–4.42)	4.20 (3.89–4.50)	4.24 (3.95–4.47)	4.20	90% 1400 ms
Closed-source full-duplex speech-vision models					
Gemini 2.5 Flash Native	3.33 (2.91–3.75)	2.81 (2.20–3.43)	3.37 (2.97–3.78)	3.17	72% 3160 ms
Gemini 3.1 Flash Live	3.15 (2.74–3.55)	2.20 (1.62–2.79)	3.16 (2.52–3.74)	2.84	66% 1720 ms
OpenAI `gpt-realtime-mini`	2.91 (2.48–3.34)	2.37 (1.76–3.02)	2.90 (2.42–3.36)	2.73	66% 5320 ms
OpenAI `gpt-realtime`	2.72 (2.27–3.16)	2.50 (1.91–3.09)	3.02 (2.53–3.49)	2.75	72% 5400 ms
Open-source full-duplex speech-vision models					
MiniCPM-o 4.5 [9]	3.03 (2.64–3.39)	<u>3.54</u> (3.02–4.04)	**3.63** (3.25–4.00)	<u>3.40</u>	**73% 720 ms**
MiniOmni2 [51]	0.65 (0.42–0.90)	1.37 (0.87–1.91)	1.54 (1.08–2.03)	1.19	64% 3080 ms
VITA-1.5 [14]	1.19 (0.94–1.45)	1.57 (1.02–2.11)	2.53 (2.10–2.98)	1.76	58% 400 ms
Audio-only full-duplex speech-vision models					
Gemini 2.5 Flash Native	3.35 (2.95–3.78)	2.98 (2.41–3.61)	3.17 (2.69–3.63)	3.17	<u>73%</u> <u>2760 ms</u>
Gemini 3.1 Flash Live	<u>3.40</u> (3.01–3.77)	2.64 (2.07–3.21)	3.03 (2.32–3.71)	3.03	69% 1240 ms
OpenAI `gpt-realtime-mini`	3.05 (2.62–3.47)	2.48 (1.91–3.07)	3.12 (2.71–3.54)	2.88	69% 5000 ms
OpenAI `gpt-realtime`	2.93 (2.50–3.39)	2.37 (1.78–2.96)	<u>3.59</u> (3.17–4.00)	2.97	67% 4440 ms
MiniCPM-o 4.5 [9]	**3.45** (3.08–3.79)	**3.76** (3.31–4.19)	3.10 (2.59–3.59)	**3.44**	72% 920 ms
MiniOmni2 [51]	1.48 (1.13–1.84)	1.70 (1.13–2.28)	2.15 (1.69–2.63)	1.72	69% 2760 ms
VITA-1.5 [14]	1.62 (1.33–1.90)	1.37 (0.87–1.89)	3.02 (2.61–3.42)	2.00	61% 800 ms

Table 4: Performance breakdown across Generation rubrics for full-duplex speech-vision models with cascaded avatars. Timing reports TOR-Alignment percentage above and median latency below. Best scores in bold.

Model	Fluency ↑	Dyadic Affect Match ↑	Nonverbal Cue Approp. ↑	Overall ↑	Timing ↑
Human ground truth	4.42 (4.21–4.61)	4.14 (3.86–4.40)	3.18 (2.78–3.59)	3.92	78% 900 ms
Gemini 2.5 + Anam [4]	**3.48** (3.17–3.79)	**3.21** (2.88–3.55)	**1.71** (1.27–2.15)	**2.80**	**44% 2840 ms**
Gemini 2.5 + Keyframe [19]	3.43 (3.12–3.74)	2.60 (2.28–2.93)	1.13 (0.75–1.50)	2.39	31% 3520 ms

TOR-Alignment of $40\%/60\%$ (AV2A/A2A) with $1720/1240$ ms median latency; Gemini 2.5 only scores $20\%/20\%$ on TOR-Alignment and even higher latency.

MiniCPM-o 4.5 uniquely exposes a user-controllable visual sampling rate knob, so we explore the impact of video sampling rates $\{1\text{–}10\}$ FPS.[2] We find that performance peaks at 2 FPS, then declines as FPS increases (VideoFDB overall: 3.04 at 8 FPS, 2.81 at 10 FPS; fluency: $3.55 \rightarrow 2.33$); more concerningly, output speech becomes increasingly incoherent. This pattern suggests a vision–speech fusion bottleneck: denser visual input can overload the shared cross-modal attention budget and degrade response quality.

***Insight 3*. No system successfully leverages the visual channel for both timing and content.** Comparing audio-only speech agents with audio-video speech agents (Tab. 3) reveals large variance in scores across model families. In all cases, the audio-video speech agents score worse on VideoFDB's perception rubrics than the audio-only speech agents. We observe three distinct patterns:

[2]We selected this sweep to bridge the model's reported FPS configurations: public documentation recommends inference at 5–10 FPS, while their paper describes training with 1–5 FPS.

- **Captioning collapse.** We find that many AV models treat visual inputs as captioning prompts in VideoFDB. To quantify this, we classify responses as {`dialogue`, `visual captioning`, `disclaimers`} with an LM judge (Fig. 4). Mini-Omni2 responds with `visual captioning` on 87% of clips, but reverts to dialogue in audio-only mode. VITA-1.5 shows a lower captioning rate (17%) but exhibits other failures, such as blanket capability disclaimers ("*I'm just a computer program and I don't have the ability to see or hear things...*") and token doubling in $\sim 74\%$ of responses. A milder version of this pattern appears in `gpt-realtime`: on AV perception clips, the full model intermittently starts with visual-scene framing (e.g., "I can see you're ...", "you look ...", "it looks like ...").
- **Visual-stream ignorance.** `gpt-realtime-mini` manifests a different consequence: it produces AV2A and A2A outputs that are paraphrases of each other; visual inspection of side-by-side transcripts confirms the visual stream rarely shifts response timing or content. This indicates the visual stream is not leveraged to provide any additional context or information.
- **TOR-Alignment.** Timing in Table 3 shows a consistent pattern across models: relative to audio-only, visual input reduces TOR-Alignment by 0–5 points for all six models except OpenAI `gpt-realtime`. Although `gpt-realtime` gains $+5$ points in TOR-Alignment, its median latency increases 1000 ms, so the response remains slower overall. *Adding video does not improve timing while staying within realtime tolerance for any model.* MiniCPM-o-4.5 is evaluated locally and shows strong timing in both modes (73% / 720 ms AV; 72% / 920 ms AO).

We find that no model achieves simultaneous AV2A gains over its audio-only counterpart on all perception axes. Our results indicate current systems primarily use visual input for explicit visual question answering rather than for robust joint audiovisual grounding in natural conversation. As a consequence, A2A models actually outperform their AV2A counterparts on VideoFDB's natural audiovisual dialogue.

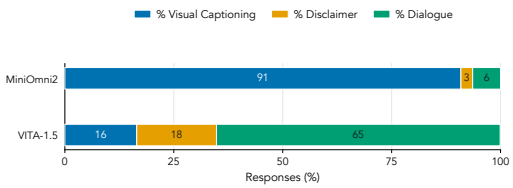

Figure 4: Mini-Omni2 [51] and VITA-1.5 [14] often respond to vision-speech input with visual captions or blanket disclaimers.

5.2 Cascaded-avatar speech models

Insight 4. **Cascading a full-duplex speech model with an avatar-rendering layer preserves turn-taking discipline but cannot supply real-time nonverbal cue production.** Relative to human ground truth, cascaded avatars show only a modest drop in Fluency ($4.42 \rightarrow 3.43$–3.48) but a large drop in Nonverbal Cue Appropriateness ($3.18 \rightarrow 1.13$–1.71). This gap is expected because audio-driven avatars are effectively turn-based (motion follows produced speech only), so they cannot add cues during the user's turn, and Anam/Keyframe cascade latency (2840–3520 ms) is too high for interactive nonverbal timing. Overall, cascaded A2AV pipelines remain fundamentally limited for nonverbal communication, motivating end-to-end speech-vision models or avatar layers that emit nonverbal motion independently of speech.

6 Conclusion

We introduced VideoFDB, the first benchmark for evaluating audio-visual full-duplex conversational agents, with 11 nonverbal dynamics, 237 expert-annotated clips, and rubric-based LM-as-judge scoring for perception and generation. Our analysis finds that current systems remain well below human conversational naturalness: AV2A models do not improve over A2A baselines and often fail to integrate nonverbal signals (e.g., *captioning collapse* and *visual-stream ignorance*), while cascaded A2AV avatars cannot produce independent, real-time nonverbal cues. These findings highlight the need for tighter joint audio-visual grounding in full-duplex interaction. We will release benchmark data and evaluation code to accelerate progress toward full-duplex audio-visual agents that converse naturally with humans.

Limitations & Future Work. VideoFDB has three main limitations: (1) our dataset is limited to English-only conversations, (2) our mid-conversation system prompts are not always strong enough to override a model's pretrained greeting defaults, and (3) VideoFDB depends on LM-judge quality which, in turn, depends on upstream caption fidelity. Future work should expand behavior coverage (full-body cues, multi-turn interactions) and broaden data to embodied settings.

Impacts. VideoFDB benchmarks progress toward full-duplex multimodal conversational agents. As humanistic conversational AI systems expand into high-impact settings, both misuse and miscommunication can cause harm. We advocate for rigorous pre-deployment evaluation and explicit safety guardrails.

Acknowledgments and Disclosure of Funding

We thank David Luebke, Ruth Rosenholtz, Ekta Prashnani, Slim Essid, and Viet Anh Trinh for feedback and early discussions on the project.

References

[1] Parakeet tdt 0.6b v2 (en). `https://huggingface.co/nvidia/parakeet-tdt-0.6b-v2`, 2024.

[2] Martha W Alibali, Dana C Heath, and Heather J Myers. Effects of visibility between speaker and listener on gesture production: Some gestures are meant to be seen. *Journal of Memory and Language*, 44(2): 169–188, 2001.

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[4] Ben Carr. Anam cara-3: Why ai needs a face. `https://anam.ai/blog/cara-3-interactive-avatars`, feb 2026. Accessed: 2026-05-05.

[5] Ashutosh Chaubey, Jiacheng Pang, Maksim Siniukov, and Mohammad Soleymani. Avere: Improving audiovisual emotion reasoning with preference optimization. In *International Conference on Learning Representations (ICLR)*, 2026. URL `https://openreview.net/forum?id=td682AAuPr`.

[6] Lele Chen, Guofeng Cui, Celong Liu, Zhong Li, Ziyi Kou, Yi Xu, and Chenliang Xu. Talking-head generation with rhythmic head motion. In *European conference on computer vision*, pages 35–51. Springer, 2020.

[7] Mingfei Chen, Zijun Cui, Xiulong Liu, Jinlin Xiang, Caleb Zheng, Jingyuan Li, and Eli Shlizerman. Savvy: Spatial awareness via audio-visual llms through seeing and hearing. *NeurIPS*, 2025.

[8] 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.

[9] 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, Changlin Liu, Hanyu Liu, Ziyang Wang, 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, Maoson Sun, Zhiyuan Liu, and Yuan Yao. MiniCPM-o 4.5: Towards real-time full-duplex omni-modal interaction, 2025. URL `https://github.com/OpenBMB/MiniCPM-o/blob/main/docs/MiniCPM_o_45_technical_report.pdf`.

[10] Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. Moshi: a speech-text foundation model for real-time dialogue. arXiv preprint arXiv:2410.00037, 2024.

[11] Joseph A DeVito. *The Interpersonal Communication Book*. Pearson Education, Inc, 16th edition, 01 2019.

[12] Yikang Ding, Jiwen Liu, Wenyuan Zhang, Zekun Wang, Wentao Hu, Liyuan Cui, Mingming Lao, Yingchao Shao, Hui Liu, Xiaohan Li, et al. Kling-avatar: Grounding multimodal instructions for cascaded long-duration avatar animation synthesis. *arXiv preprint arXiv:2509.09595*, 2025.

[13] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24108–24118, 2025.

[14] Chaoyou Fu, Haojia Lin, Xiong Wang, YiFan Zhang, Yunhang Shen, Xiaoyu Liu, Haoyu Cao, Zuwei Long, Heting Gao, Ke Li, Long MA, Xiawu Zheng, Rongrong Ji, Xing Sun, Caifeng Shan, and Ran He. VITA-1.5: Towards GPT-4o level real-time vision and speech interaction. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=8PUzLga3lU.

[15] Charles Goodwin. *Conversational Organization: Interaction Between Speakers and Hearers*. 01 1981.

[16] Google DeepMind. Gemini live api. https://ai.google.dev/gemini-api/docs/live-api, 2025. Accessed: 2026-04-30.

[17] Ke Hu, Ehsan Hosseini-Asl, Chen Chen, Edresson Casanova, Subhankar Ghosh, Piotr Żelasko, Zhehuai Chen, Jason Li, Jagadeesh Balam, and Boris Ginsburg. Salm-duplex: Efficient and direct duplex modeling for speech-to-speech language model. *Interspeech*, 2025.

[18] 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. *arXiv preprint arXiv:2410.21276*, 2024.

[19] Keyframe. Keyframe: Persona-1-live. https://www.keyframelabs.com/blog/persona-1-live, may 2026. Accessed: 2026-05-05.

[20] Terry K Koo and Ma Li. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, 15 2:155–63, 2016. URL https://api.semanticscholar.org/CorpusID:1837377.

[21] Stephen C. Levinson and Francisco Torreira. Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology*, Volume 6 - 2015, 2015. ISSN 1664-1078. doi: 10.3389/fpsyg.2015.00731. URL https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2015.00731.

[22] Guan-Ting Lin, Shih-Yun Shan Kuan, Qirui Wang, Jiachen Lian, Tingle Li, and Hung-yi Lee. Full-duplex-bench v1. 5: Evaluating overlap handling for full-duplex speech models. *arXiv preprint arXiv:2507.23159*, 2025.

[23] Guan-Ting Lin, Jiachen Lian, Tingle Li, Qirui Wang, Gopala Anumanchipalli, Alexander H Liu, and Hung-yi Lee. Full-duplex-bench: A benchmark to evaluate full-duplex spoken dialogue models on turn-taking capabilities. *arXiv preprint arXiv:2503.04721*, 2025.

[24] Guan-Ting Lin, Shih-Yun Shan Kuan, Jiatong Shi, Kai-Wei Chang, Siddhant Arora, Shinji Watanabe, and Hung-yi Lee. Full-duplex-bench-v2: A multi-turn evaluation framework for duplex dialogue systems with an automated examiner. *arXiv preprint arXiv:2510.07838*, 2026.

[25] Haiyang Liu, Zihao Zhu, Naoya Iwamoto, Yichen Peng, Zhengqing Li, You Zhou, Elif Bozkurt, and Bo Zheng. Beat: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis. In *European conference on computer vision*, pages 612–630. Springer, 2022.

[26] Haiyang Liu, Zihao Zhu, Giorgio Becherini, Yichen Peng, Mingyang Su, You Zhou, Xuefei Zhe, Naoya Iwamoto, Bo Zheng, and Michael J Black. Emage: Towards unified holistic co-speech gesture generation via expressive masked audio gesture modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1144–1154, 2024.

[27] Jingwen Liu, Kan Jen Cheng, Jiachen Lian, Akshay Anand, Rishi Jain, Faith Qiao, Robin Netzorg, Huang-Cheng Chou, Tingle Li, Guan-Ting Lin, and Gopala Anumanchipalli. Emo-reasoning: Benchmarking emotional reasoning capabilities in spoken dialogue systems. 2025.

[28] LiveKit, Inc. Livekit cloud. https://docs.livekit.io/intro/cloud/, 2026. Fully managed platform for building, hosting, and operating AI agent applications at scale.

[29] Cheng Luo, Jianghui Wang, Bing Li, Siyang Song, and Bernard Ghanem. Omniresponse: Online multimodal conversational response generation in dyadic interactions. *NeurIPS*, 2025.

[30] Antje S. Meyer. Timing in conversation. *Journal of Cognition*, Apr 2023. doi: 10.5334/joc.268.

[31] Laura M Morett. Examining gesture production in the presence of communication challenges. *JoVE (Journal of Visualized Experiments)*, (203):e66256, 2024.

[32] 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, et al. See, hear, and understand: Benchmarking audiovisual human speech understanding in multimodal large language models. *CVPR Findings*, 2026.

[33] Tu Anh Nguyen, Eugene Kharitonov, Jade Copet, Yossi Adi, Wei-Ning Hsu, Ali Elkahky, Paden Tomasello, Robin Algayres, Benoit Sagot, Abdelrahman Mohamed, et al. Generative spoken dialogue language modeling. *Transactions of the Association for Computational Linguistics*, 11:250–266, 2023.

[34] NVIDIA, :, Amala Sanjay Deshmukh, Kateryna Chumachenko, Tuomas Rintamaki, Matthieu Le, Tyler Poon, Danial Mohseni Taheri, Ilia Karmanov, Guilin Liu, Jarno Seppanen, Arushi Goel, Mike Ranzinger, Greg Heinrich, Guo Chen, Lukas Voegtle, Philipp Fischer, Timo Roman, Karan Sapra, Collin McCarthy, Shaokun Zhang, Fuxiao Liu, Hanrong Ye, Yi Dong, Mingjie Liu, Yifan Peng, Piotr Zelasko, Zhehuai Chen, Nithin Rao Koluguri, Nune Tadevosyan, Lilit Grigoryan, Ehsan Hosseini Asl, Pritam Biswas, Leili Tavabi, Yuanhang Su, Zhiding Yu, Peter Jin, Alexandre Milesi, Netanel Haber, Yao Xu, Sarah Amiraslani, Nabin Mulepati, Eric Tramel, Jaehun Jung, Ximing Lu, Brandon Cui, Jin Xu, Zhiqi Li, Shihao Wang, Yuanguo Kuang, Shaokun Zhang, Huck Yang, Boyi Li, Hongxu Yin, Song Han, Pavlo Molchanov, Adi Renduchintala, Charles Wang, David Mosallanezhad, Soumye Singhal, Luis Vega, Katherine Cheung, Sreyan Ghosh, Yian Zhang, Alexander Bukharin, Venkat Srinivasan, Johnny Greco, Andre Manoel, Maarten Van Segbroeck, Suseella Panguliri, Rohit Watve, Divyanshu Kakwani, Shubham Pachori, Jeffrey Glick, Radha Sri-Tharan, Aileen Zaman, Khanh Nguyen, Shi Chen, Jiaheng Fang, Qing Miao, Wenfei Zhou, Yu Wang, Zaid Pervaiz Bhat, Varun Praveen, Arihant Jain, Ramanathan Arunachalam, Tomasz Kornuta, Ashton Sharabiani, Amy Shen, Wei Huang, Yi-Fu Wu, Ali Roshan Ghias, Huiying Li, Brian Yu, Nima Tajbakhsh, Chen Cui, Wenwen Gao, Li Ding, Terry Kong, Manoj Kilaru, Anahita Bhiwandiwalla, Marek Wawrzos, Daniel Korzekwa, Pablo Ribalta, Grzegorz Chlebus, Besmira Nushi, Ewa Dobrowolska, Maciej Jakub Mikulski, Kunal Dhawan, Steve Huang, Jagadeesh Balam, Yongqiang Wang, Nikolay Karpov, Valentin Mendelev, George Zelenfroynd, Meline Mkrtchyan, Qing Miao, Omri Almog, Bhavesh Pawar, Rameshwar Shivbhakta, Sudeep Sabnis, Ashrton Sharabiani, Negar Habibi, Geethapriya Venkataramani, Pamela Peng, Prerit Rodney, Serge Panev, Richard Mazzarese, Nicky Liu, Michael Fukuyama, Andrii Skliar, Roger Waleffe, Duncan Riach, Yunheng Zou, Jian Hu, Hao Zhang, Binfeng Xu, Yuhao Yang, Zuhair Ahmed, Alexandre Milesi, Carlo del Mundo, Chad Voegele, Zhiyu Cheng, Nave Assaf, Andrii Skliar, Daniel Afrimi, Natan Bagrov, Ran Zilberstein, Ofri Masad, Eugene Khvedchenia, Natan Bagrov, Borys Tymchenko, Tomer Asida, Daniel Afrimi, Parth Mannan, Victor Cui, Michael Evans, Katherine Luna, Jie Lou, Pinky Xu, Guyue Huang, Negar Habibi, Michael Boone, Pradeep Thalasta, Adeola Adesoba, Dina Yared, Christopher Parisien, Leon Derczynski, Shaona Ghosh, Wes Feely, Micah Schaffer, Radha Sri-Tharan, Jeffrey Glick, Barnaby Simkin, George Zelenfroynd, Tomasz Grzegorzek, Rishabh Garg, Aastha Jhunjhunwala, Sergei Kolchenko, Farzan Memarian, Haran Kumar, Shiv Kumar, Isabel Hulseman, Anjali Shah, Kari Briski, Padmavathy Subramanian, Joey Conway, Udi Karpas, Jane Polak Scowcroft, Annie Surla, Shilpa Ammireddy, Ellie Evans, Jesse Oliver, Tom Balough, Chia-Chih Chen, Sandip Bhaskar, Alejandra Rico, Bardiya Sadeghi, Seph Mard, Katherine Cheung, Meredith Price, Laya Sleiman, Saori Kaji, Wesley Helmholz, Wendy Quan, Michael Lightstone, Jonathan Cohen, Jian Zhang, Oleksii Kuchaiev, Boris Ginsburg, Jan Kautz, Eileen Long, Mohammad Shoeybi, Mostofa Patwary, Oluwatobi Olabiyi, Andrew Tao, Bryan Catanzaro, and Udi Karpas. Nemotron 3 nano omni: Efficient and open multimodal intelligence, 2026. URL `https://arxiv.org/abs/2604.24954`.

[35] OpenAI. Openai realtime api. `https://developers.openai.com/api/docs/guides/realtime`, 2024. Accessed: 2026-04-30.

[36] Yizhou Peng, Yi-Wen Chao, Dianwen Ng, Yukun Ma, Chongjia Ni, Bin Ma, and Eng Siong Chng. Fd-bench: A full-duplex benchmarking pipeline designed for full duplex spoken dialogue systems, 2025. URL `https://arxiv.org/abs/2507.19040`.

[37] Ziqiao Peng, Yanbo Fan, Haoyu Wu, Xuan Wang, Hongyan Liu, Jun He, and Zhaoxin Fan. Dualtalk: Dual-speaker interaction for 3d talking head conversations. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21055–21064, 2025.

[38] Reza Pourreza, Rishit Dagli, Apratim Bhattacharyya, Sunny Panchal, Guillaume Berger, and Roland Memisevic. Can vision-language models answer face to face questions in the real-world? In *The Fourteenth International Conference on Learning Representations*, 2026. URL `https://openreview.net/forum?id=I3dPEvbp8o`.

[39] Lixiong Qin, Shilong Ou, Miaoxuan Zhang, Jiangning Wei, Yuhang Zhang, Xiaoshuai Song, Yuchen Liu, Mei Wang, and Weiran Xu. Face-human-bench: A comprehensive benchmark of face and human understanding for multi-modal assistants. *NeurIPS*, 2025.

[40] Qwen Team. Qwen2.5-Omni technical report. arXiv preprint arXiv:2503.20215, 2025.

[41] Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL `https://qwen.ai/blog?id=qwen3.5`.

[42] Gorjan Radevski, Teodora Popordanoska, Matthew B. Blaschko, and Tinne Tuytelaars. DAVE: Diagnostic benchmark for audio visual evaluation. In *The Thirty-ninth Annual Conference on Neural Information*

Processing Systems Datasets and Benchmarks Track, 2025. URL https://openreview.net/forum?id=4ZAX1NT0ms.

[43] Rajarshi Roy, Jonathan Raiman, Sang gil Lee, Teodor-Dumitru Ene, Robert Kirby, Sungwon Kim, Jaehyeon Kim, and Bryan Catanzaro. Personaplex: Voice and role control for full duplex conversational speech models, 2026. URL https://arxiv.org/abs/2602.06053.

[44] Amélie Royer, Moritz Böhle, Gabriel de Marmiesse, Laurent Mazaré, Neil Zeghidour, Alexandre Défossez, and Patrick Pérez. Vision-speech models: Teaching speech models to converse about images. *arXiv preprint arXiv:2503.15633*, 2025.

[45] Guangzhi Sun, Wenyi Yu, Changli Tang, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun MA, Yuxuan Wang, and Chao Zhang. video-SALMONN: Speech-enhanced audio-visual large language models. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=nYsh5GFIqX.

[46] Yunlong Tang, Pinxin Liu, Zhangyun Tan, Mingqian Feng, Rui Mao, Chao Huang, Jing Bi, Yunzhong Xiao, Susan Liang, Hang Hua, et al. Mmperspective: Do mllms understand perspective? a comprehensive benchmark for perspective perception, reasoning, and robustness. *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025.

[47] Silero Team. Silero vad: pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier. https://github.com/snakers4/silero-vad, 2024.

[48] Bandhav Veluri, Benjamin N Peloquin, Bokai Yu, Hongyu Gong, and Shyamnath Gollakota. Beyond turn-based interfaces: Synchronous LLMs as full-duplex dialogue agents. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21390–21402, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1192. URL https://aclanthology.org/2024.emnlp-main.1192/.

[49] Yuxuan Wang, Yueqian Wang, Bo Chen, Tong Wu, Dongyan Zhao, and Zilong Zheng. Omnimmi: A comprehensive multi-modal interaction benchmark in streaming video contexts, 2025.

[50] Paul Watzlawick, Janet Helmick Beavin, and Don D. Jackson. *Pragmatics of Human Communication: A Study of Interactional Patterns, Pathologies, and Paradoxes*. W. W. Norton, New York, NY, 1967.

[51] Zhifei Xie and Changqiao Wu. Mini-omni2: Towards open-source gpt-4o with vision, speech and duplex capabilities. *ArXiv*, abs/2410.11190, 2024.

[52] Jinbo Xing, Menghan Xia, Yuechen Zhang, Xiaodong Cun, Jue Wang, and Tien-Tsin Wong. Codetalker: Speech-driven 3d facial animation with discrete motion prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12780–12790, 2023.

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

[54] Shuhang Xun, Sicheng Tao, Jungang Li, Yibo Shi, Zhixin Lin, Zhanhui Zhu, Yibo Yan, Hanqian Li, Linghao Zhang, Shikang Wang, Yixin Liu, Hanbo Zhang, Ying Ma, and Xuming Hu. Rtv-bench: Benchmarking mllm continuous perception, understanding and reasoning through real-time video. In *Advances in Neural Information Processing Systems*, volume 38. NeurIPS, 2025.

[55] Victor H. Yngve. On getting a word in edgewise. 1970. URL https://api.semanticscholar.org/CorpusID:143317921.

[56] Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. SpeechGPT: Empowering large language models with intrinsic cross-modal conversational abilities. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15757–15773, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.1055. URL https://aclanthology.org/2023.findings-emnlp.1055/.

[57] Qinglin Zhang, Luyao Cheng, Chong Deng, Qian Chen, Wen Wang, Siqi Zheng, Jiaqing Liu, Hai Yu, Chao-Hong Tan, Zhihao Du, and ShiLiang Zhang. OmniFlatten: An end-to-end GPT model for seamless voice conversation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14570–14580, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.709. URL https://aclanthology.org/2025.acl-long.709/.

[58] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

Appendix

A Dataset

Reviewer Dataset Access. We provide access to the full VideoFDB Evaluation Dataset at the anonymized URL `https://anonvfdb.github.io/`. From the landing page, please read and accept the Terms of Use, then click *Continue to Dataset* and enter the password:

`sH6A+P12qMaJWtyMJ2vIx9Oi`

to access. The access-gated page contains two illustrative validation clips, the categorical taxonomy, and the per-clip metadata schema; the full archive is ($\sim$5 GB) and Croissant metadata are accessible via the Download button on that same page.

We provide all prompts and detailed rubrics for reproducibility. Before paper publication, we promise to release the dataset and evaluation code in a public-facing HuggingFace.

A.1 Additional Dataset Details

Technical specs. All source recordings are screened to meet minimum quality thresholds (Table 5).

Table 5: Minimum technical specifications for VideoFDB source recordings.

Property	Minimum
Video resolution	720p
Frame rate	30 fps
Audio sample rate	24 kHz
Video format	VP9 (`.mp4`)
Audio format	PCM (`.wav`)

Collection and annotation details. Our dataset is mined from a large corpus of two-person video calls, and specifically held out from other dataset releases to prevent training data contamination. Participants are connected via a videoconferencing platform and given a conversation prompt, which one interlocutor introduces naturally to initiate the conversation. To minimize the effects of network latency, each speaker's audio-video stream is recorded locally in parallel with live transmission to the other participant.

Source recordings are captured locally and passed through media-quality and preprocessing checks, including intra-channel A/V sync validation and inter-channel alignment validation. Dataset construction follows a three-pass annotation workflow: (1) human annotators manually identify candidate dynamic moments in longer recordings, (2) additional human annotators validate the trimmed clips, retaining only candidate samples with sufficient pre-event context (targeting 1–3 turns prior to the event), and assign precise event timestamps and dynamic type labels, and (3) a final human reviewer performs quality checks for label consistency and timestamp precision. Validated clips are then passed through duration verification with `ffprobe`, category filtering to in-scope dynamics, and trim validation before release packaging.

Summary statistics. Table 6 summarizes clip and dynamic-window durations for the full set and the curated evaluation subset.

Table 6: Dataset summary statistics. Dynamic-window length is $t_{\text{end}} - t_{\text{start}}$.

Set	# clips	# P / # G	Clip duration (median / IQR)	Dyn. window (median / IQR)
Test (held-out scoring set)	226	105 / 121	46.0s / [34.0, 61.4]	2.5s / [2.0, 5.0]
Validation (public eval set)	11	5 / 6	46.0s / [37.6, 55.8]	6.0s / [2.0, 28.2]

Demographic statistics. Our dataset includes English-speaking participants with a range of accents and backgrounds, located in the United States and Canada. Table 7 summarizes the demographic statistics of the dataset.

Table 7: Demographic statistics of the dataset.

Property	Value
Unique speakers	130
Age Range (Percent of Speakers)	
18–29	19%
30–39	32%
40–49	20%
50–59	18%
60–69	8%
70–79	2%
Gender (Percent of Speakers)	
Female	44%
Male	54%
Prefer Not to Say	2%

B Further Limitations

Dataset scope and representational constraints. The benchmark contains English-language, two-person video conferencing conversations recorded by participants in the United States and Canada, representing a narrow slice of video-call interaction. All recordings consist of dyadic webcam-style captures from standard video conferencing setups; they do not represent in-person, mobile, or alternative camera configurations. Cultural context further modulates nonverbal communication channels, since gesture, gaze, and affect conventions vary across communities and interaction settings. VideoFDB is therefore not recommended for evaluation of multilingual models, in-person or multi-party conversation modeling, or non-English-speaking populations.

Evaluation scope. The benchmark supports single-turn evaluation of audiovisual conversational dynamics only. It does not support multi-turn evaluation, training, or fine-tuning of any kind, nor generalization to codec environments, recording modalities, or conversational contexts outside those represented here.

Evaluation pipeline constraints. Like all LM-based evaluations, our evaluation is bounded by the perceptual capabilities of the underlying captioning model, including both its visual and audio comprehension quality. Specifically, we observe that our judge assessment of ground-truth humans defines an upper bound for rubric scoring.

C Evaluation Protocol Details

C.1 Dynamic-to-rubric mapping

Table 8 shows the full breakdown of which rubric axes apply to each conversational dynamic, split between perception and generation buckets.

C.2 Timing Metrics Details

TOR-Alignment extends the speech-only TOR formulation of [23] to multimodal AV dynamics and places heterogeneous timing expectations under one metric. Rather than inferring turn boundaries from ASR alone, we evaluate timing against the annotated event window `[dynamic_start_s, dynamic_end_s]`. We compute TOR-Alignment for each timing-relevant dynamic by mapping it to one of five **timing classes**, defined by the expected agent behavior at the cue (see Tab. 8):

- STAY-SILENT (Pause Handling, Gaze Avoidance with Pause): the user is mid-thought; the agent should not take the floor in-window. A clip passes if in-window agent speech is ≤ 1 s, or all overlap segments are backchannels (< 1 s and < 2 words; [23]).
- CONTINUE-SPEAKING (Nonverbal Backchanneling, Adaptor Handling): the user emits a non-floor-offering cue while the agent is speaking; failure is inappropriate yielding. If

Table 8: A conversational dynamic can be evaluated under perception, generation, or both. "#" reports total samples per dynamic across perception and generation splits.

Dynamic	#	Perception			Generation		
		Fluency	Conv. Flow	Semantic Grounding	Fluency	Dyadic Affect Match	Cue. Approp.
Pause Handling	13	✓	✓				
Gaze Avoidance with Pause	18	✓	✓				
Nonverbal Interruption	8	✓	✓	✓			
Adaptor Handling	14	✓		✓			
Face Emotion Display	50	✓		✓	✓	✓	✓
Laughter	30	✓		✓	✓	✓	✓
Nonverbal Backchanneling	34	✓	✓		✓	✓	✓
Emotion Matching	13				✓	✓	✓
Verbal Backchanneling	17				✓	✓	✓
Verbal Interruption	24				✓	✓	✓
Turn-taking	16				✓	✓	

pre-cue agent activity in the previous 3 s is $\geq 30\%$, in-window activity must remain $\geq 50\%$; otherwise, the case is treated as vacuously correct.

- YIELD-REQUIRED (Nonverbal Interruption, Verbal Interruption generation): the user takes the floor; the agent must yield within 1.5 s. Latency is measured as the end time of the agent segment overlapping (or immediately preceding) `dynamic_start_s`, relative to `dynamic_start_s`.
- SMOOTH-HANDOFF (Turn-taking): the agent should start speaking within the annotated handoff window. Latency is `first_onset_after(dynamic_start_s)`; if the agent is already speaking at cue onset, latency is 0 ms.
- BACKCHANNEL-PRODUCED (Verbal Backchanneling): the agent should produce at least one in-window backchannel (< 1 s, < 2 words) and avoid full floor-taking (no segment above backchannel limits). Multiple short in-window backchannels are valid.

Formal definition. Following [23], the binary takeover variable is

$$\mathrm{TO}_i = \begin{cases} 0, & \text{if the agent's output is silence or a backchannel,} \\ 1, & \text{otherwise,} \end{cases}$$

for each clip i, and the takeover rate is $\mathrm{TOR} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{TO}_i$ – typically reported per dynamic, with the preferred direction (lower-better or higher-better) interpreted case-by-case. In TOR-Alignment, we encode that direction directly: each timing class $c \in \mathcal{C}$ has a policy-prescribed expected takeover $\mathrm{TO}^*_c \in \{0, 1\}$ — 0 for STAY-SILENT, YIELD-REQUIRED, and BACKCHANNEL-PRODUCED; 1 for CONTINUE-SPEAKING and SMOOTH-HANDOFF. The per-clip alignment indicator is

$$A_i \;=\; \begin{cases} 1, & \text{if } \mathrm{TO}_i = \mathrm{TO}^*_{c_i}, \\ 0, & \text{otherwise,} \end{cases}$$

and TOR-Alignment is its mean:

$$\text{TOR-Alignment} \;=\; \frac{1}{N}\sum_{i=1}^{N} A_i.$$

Higher TOR-Alignment indicates better agreement with timing-class-specific takeover policy, i.e., the fraction of clips that satisfy the expected timing behavior.

C.3 Judge Validation

We assess judge robustness across three judge backends: `meta/llama-3.1-70b-instruct`, `azure/openai/gpt-4o`, and `azure/anthropic/claude-sonnet-4-6`. Each judge independently scores the same response on each axis with the same rubric and prompt. We report (i) ground truth (GT) calibration on the held-out reference set (Table 9) and (ii) inter-judge agreement on a combined GT+RANDOM cross-judge subset of n=220 clips (n=110 ground-truth + n=110 random-baseline with mixed Initiator/Respondent channels; Table 10).

Table 9: GT judge calibration across three judge backends. Judge-specific cells report mean ± stdev (0–5). The final row reports consensus statistics (per-sample mean across the three judges) with 95% CI and sample count.

Judge	Fluency	Conversational Flow	Visual Grounding
`llama-3.1-70b-instruct`	4.52 ± 0.94	4.50 ± 0.91	4.26 ± 0.54
`gpt-4o`	4.54 ± 1.28	4.48 ± 1.10	4.18 ± 1.06
`claude-sonnet-4-6`	4.55 ± 0.79	4.16 ± 0.95	4.56 ± 0.59
Consensus (3-judge mean)	**4.53 (95% CI: [4.38, 4.69]**	**4.38 (95% CI: [4.16, 4.60])**	**4.33 (95% CI: [4.18, 4.48]**

Table 10: Inter-judge agreement on perception rubrics across three judge backends (`llama-70b`, `gpt-4o`, `claude-sonnet-4.6`) on the combined GT+RANDOM subset. We report ICC (single-judge and averaged-judge), Within-1pt agreement on the 0–5 scale, and MAD (mean absolute pairwise difference).

Axis	n	ICC(A,k) [95% CI]	ICC(A,1) [95% CI]	Within-1pt	MAD
Fluency	220	0.84 [0.79, 0.87]	0.63 [0.55, 0.70]	77.7%	0.98
Conversational Flow	112	0.90 [0.86, 0.93]	0.76 [0.67, 0.83]	88.7%	0.70
Visual Grounding	124	0.75 [0.66, 0.82]	0.50 [0.39, 0.60]	80.6%	0.82

GT calibration. Table 9 reports judge-specific mean±stdev across Fluency, Conversational Flow, and Visual Grounding, plus a consensus estimate with 95% confidence intervals. The three judges are closely aligned on aggregate GT means while still exposing judge-dependent spread (stdev), motivating reporting both per-judge variability and consensus confidence intervals.

Inter-judge agreement. We find 77–89% pairwise agreement within 1 point on the 0–5 scale: the 3-judge averaged score reaches $\mathrm{ICC(A, k)} = 0.84$ / 0.90 / 0.75 on perception rubrics Fluency / Conversational Flow / Visual Grounding (Table 10). We use the intraclass correlation coefficient (ICC) under a two-way random-effects model with absolute-agreement criterion. ICC(A,1) is the reliability of a single judge's score; ICC(A,k) is the reliability of the averaged score across the k=3 judges — the relevant statistic for the consensus mean reported on the leaderboard. We follow the Koo and Li [20] interpretation: <0.50 poor, 0.50–0.75 moderate, 0.75–0.90 good, >0.90 excellent.

The 3-judge averaged score reaches good-to-excellent reliability for Fluency and Conversational Flow ($\mathrm{ICC(A, k)} \geq 0.84$), and moderate-to-good reliability for Visual Grounding ($\mathrm{ICC(A, k)} = 0.75$, lower 95% CI bound 0.66). At the single-judge level, agreement is moderate for Fluency and Conversational Flow ($\mathrm{ICC(A, 1)} = 0.63$ and 0.76) and weaker for Visual Grounding ($\mathrm{ICC(A, 1)} = 0.50$).

This pattern is consistent with the nature of the task. Fluency and Conversational Flow can be reasonably assessed from the response transcript and turn-state metadata, whereas Visual Grounding requires judges to determine whether the response is conditioned on a visual cue described in an auxiliary caption; this judgment has greater scoring variability. Even on this most challenging axis, 80.6% of pairwise (clip, axis) comparisons are within one point on the 0–5 scale, and averaging across three judges raises reliability to the moderate-to-good range. We therefore interpret cross-model differences in Visual Grounding at the model-aggregate level over the n=105 test clips, where the standard error of the mean is substantially smaller than per-clip disagreement.

D Models & Implementation

We evaluate a diverse set of full-duplex baselines spanning omni-modal, audio-visual, and audio-only conversational speech agents.

System prompts. Prompt-construction details are provided in Sec. F.4; the only model tested that does not support system prompts is MiniOmni2 [51].

Model details. We evaluate the following models with the same VideoFDB protocol: each clip is streamed through a unified realtime harness, using audio+video for AV runs and disabling video for

Table 11: Per-category perception results.

Model	Face Emotion	Gaze Avoid.	Nonverb. BC	Laughter	Adaptor	Pause	Nonverb. Int.
Human reference	4.08 (3.62–4.46)	3.71 (2.76–4.53)	4.12 (4.00–4.31)	3.86 (3.14–4.50)	5.00 (5.00–5.00)	5.00 (5.00–5.00)	4.12 (3.81–4.44)
Closed-source full-duplex speech-vision models							
Gemini direct (AV)	2.42 (1.75–3.08)	2.94 (1.76–4.12)	2.44 (1.50–3.56)	3.57 (3.00–4.00)	5.00 (5.00–5.00)	3.46 (1.92–4.62)	2.75 (1.69–3.62)
Gemini 3.1 direct (AV)	2.61 (1.92–3.29)	2.00 (1.25–2.81)	2.85 (2.06–3.65)	3.73 (3.27–4.14)	4.90 (4.70–5.00)	2.73 (1.77–3.65)	2.33 (1.67–3.00)
OpenAI gpt-realtime-mini (AV)	1.83 (1.08–2.50)	2.94 (1.76–4.12)	2.19 (1.12–3.31)	3.00 (2.43–3.50)	5.00 (5.00–5.00)	2.85 (1.62–4.08)	1.62 (0.81–2.44)
OpenAI gpt-realtime (AV)	2.12 (1.33–2.92)	2.88 (1.76–4.06)	2.62 (1.62–3.62)	3.21 (2.57–3.79)	4.62 (3.85–5.00)	2.31 (1.15–3.85)	2.25 (1.31–3.38)
Open-source full-duplex speech-vision models							
MiniCPM-o 4.5	3.00 (2.33–3.58)	3.24 (2.12–4.35)	3.44 (2.56–4.19)	3.07 (2.29–3.79)	5.00 (5.00–5.00)	3.85 (2.69–5.00)	4.06 (3.75–4.38)
MiniOmni2 (AV)	0.83 (0.58–1.08)	0.59 (0.00–1.47)	3.00 (2.00–3.75)	0.50 (0.21–0.86)	3.85 (2.69–5.00)	0.38 (0.00–1.15)	1.56 (0.62–2.75)
VITA-1.5 (AV)	2.08 (1.62–2.58)	0.65 (0.00–1.53)	2.75 (1.75–3.50)	1.21 (0.71–1.71)	4.15 (3.08–5.00)	0.00 (0.00–0.00)	3.62 (3.25–4.00)
Audio-only full-duplex speech-vision models							
VITA-1.5 (audio only)	2.38 (1.83–2.92)	0.00 (0.00–0.00)	2.94 (2.06–3.75)	2.21 (1.50–3.00)	5.00 (5.00–5.00)	0.00 (0.00–0.00)	3.25 (2.31–4.00)
Gemini direct (audio only)	2.04 (1.29–2.83)	2.94 (1.76–4.12)	2.81 (1.75–3.94)	3.21 (2.71–3.71)	5.00 (5.00–5.00)	3.85 (2.69–5.00)	2.75 (1.81–3.56)
Gemini 3.1 direct (audio only)	2.63 (1.97–3.29)	2.19 (1.42–3.00)	3.76 (3.12–4.35)	3.91 (3.50–4.32)	4.25 (3.45–4.95)	3.08 (2.12–4.04)	2.76 (2.14–3.38)
OpenAI gpt-realtime-mini (audio only)	2.00 (1.33–2.67)	2.65 (1.53–3.82)	2.69 (1.69–3.75)	3.07 (2.57–3.57)	5.00 (5.00–5.00)	2.31 (1.15–3.85)	2.75 (2.00–3.38)
OpenAI gpt-realtime (audio only)	3.04 (2.25–3.79)	2.41 (1.29–3.59)	3.12 (2.12–4.19)	3.43 (2.79–4.00)	5.00 (5.00–5.00)	1.54 (0.38–3.08)	2.69 (1.69–3.56)
MiniCPM-o 4.5 (audio only)	1.67 (0.92–2.46)	3.76 (2.71–4.65)	3.31 (2.50–4.06)	3.50 (2.86–4.07)	5.00 (5.00–5.00)	4.62 (3.85–5.00)	3.44 (2.94–3.81)
MiniOmni2 (audio only)	0.94 (0.62–1.27)	1.59 (0.85–2.35)	1.44 (0.84–2.06)	1.61 (1.14–2.11)	3.58 (2.85–4.23)	1.54 (0.77–2.50)	2.12 (1.38–2.92)

audio-only runs. Unless noted otherwise, we provide the clip-specific VideoFDB system prompt to the model.

- **Gemini 2.5 / 3.1 Flash Native** [8]. We run Gemini models with the Gemini API and WebSocket input, streaming user audio and sampled video frames. We then log the model's spoken response for scoring.
- **OpenAI Realtime / Realtime mini** [35]. We run OpenAI Realtime with the OpenAI API and WebSocket input, with the same clip-level streaming setup and prompt construction used for Gemini so comparisons stay controlled.
- **MiniCPM-o 4.5** [9]. We run MiniCPM-o-4.5 directly through its full-duplex demo API. Audio is streamed in 1-second chunks; for AV runs we provide sampled video frames, and for audio-only runs we disable frame input. We also use one fixed reference TTS clip across samples for consistency.
- **MiniOmni2** [51]. We integrate MiniOmni2 directly, using its AV path when frames are available and its audio-only path otherwise. Because its serving interface is ultimately half-duplex, we buffer each clip, segment speech with Silero VAD [47], align each segment to the most recent preceding video frame, and merge segment-level outputs into one response. To fit its fixed Whisper input length, we cap each segment at 29.5 seconds and warn when truncation occurs. MiniOmni2 does not support system prompts.
- **VITA-1.5** [14]. We deploy VITA-1.5 using their official realtime demo stack [14]. We replace the default prompt ("I am an AI robot named VITA") with "You are a helpful assistant on a video call." and append our clip-specific system prompt. We evaluate both AV ($n_{\text{frames}} = 4$ at 1 fps) and audio-only (video disabled). For clips marked `agent_speaks_first`, we trigger an initial no-audio turn after the first video frame.
- **Gemini Live 2.5 Flash Native + Anam avatar** [16, 4]. This is a cascaded setup: Gemini generates the speech response, and Anam renders avatar motion from that speech stream.
- **Gemini Live 2.5 Flash Native + Keyframe avatar** [16, 19]. This matches the same cascaded pipeline as above, but uses Keyframe as the avatar renderer.

Hardware Details For open-source models, we run local inference on a server with 8×NVIDIA H100 80GB HBM3 GPUs and 2×Intel(R) Xeon(R) Platinum 8480+ CPUs, using one GPU per run.

E Per-Dynamic Results

Tables 11 to 13 present per-category call-outs split into Perception, Generation rubrics, and Generation timing.

Table 12: Per-category generation-axis results (means). *Hum.* is the hand-annotated human reference; *Anam* and *KF* are Gemini 2.5 Flash Native cascaded with the Anam (neural rendering) and Keyframe (interpolation) avatar layers respectively.

	Gen. Fluency ↑			**Affect Match ↑**			**Cue Appropr. ↑**		
Category	Hum.	Anam	KF	Hum.	Anam	KF	Hum.	Anam	KF
Laughter	4.14	2.86	**3.07**	4.14	**2.00**	1.14	3.57	**0.14**	0.00
Nonverbal Backchanneling	4.38	3.25	**3.62**	4.44	**4.25**	4.00	1.75	**0.00**	**0.00**
Verbal Backchanneling	4.19	3.12	**3.19**	4.56	**2.81**	2.25	3.31	**0.06**	0.00
Verbal Interruption	4.13	**3.43**	3.00	2.74	**3.96**	3.09	2.27	**3.12**	1.36
Emotion Matching	4.69	4.31	**4.62**	4.54	**3.46**	1.92	4.15	**2.92**	1.15
Face Emotion Display	4.62	**3.96**	3.46	4.42	2.79	**3.04**	3.83	**2.83**	2.00
Turn-taking	4.87	3.27	**3.40**	4.73	**3.00**	2.00	3.67	**3.13**	3.00

Table 13: Per-dynamic deterministic timing results (generation). Each cell shows TOR-Alignment % on top and median yield/onset latency in ms underneath. `lat` 0 ms = the agent was already speaking at the cue (in-progress short-circuit). Nonverbal Backchanneling generation is omitted because the agent's expected response is a visual cue rather than audio (use Visual During% from Tab. 4 instead). Cascaded avatars score 0 % on Verbal Backchanneling because the cascade architecturally cannot insert a brief in-window audio cue (Sec. 5.2, Insight 4).

Model *(timing class)*	**Verbal BC** *backchannel-produced*	**Verbal Int.** *yield-required*	**Turn-taking** *smooth-handoff*
Human reference	69% lat 920 ms 16/16	70% lat 960 ms 23/23	100% lat 640 ms 15/15
Gemini 2.5 Flash Native + Anam	0% lat 0 ms 16/16	75% lat 420 ms 8/23	73% lat 6880 ms 15/15
Gemini 2.5 Flash Native + Keyframe	0% lat 0 ms 16/16	36% lat 1920 ms 11/23	60% lat 6860 ms 15/15

F LM-as-Judge Prompts & Rubrics

This appendix documents the exact LM-as-judge prompts used in VideoFDB (Sec. 4). We open with a single worked example (Sec. F.1), document the inputs the judge consumes (Sec. F.2) and the output schemas it returns (Sec. F.3), and then walk through the perception (Sec. F.5) and generation (Sec. F.6) pipelines.

F.1 Worked example: one Laughter clip end-to-end

We first present a worked example of a Laughter clip end-to-end, prepared for both perception and generation evaluations. The captions are free-form prose: there are no `[laugh]` tags, no gaze-direction tags, no special markup. Paralinguistics like laughter, sighs, and gaze shifts are described in natural language by the upstream captioners.

Perception payload — Laughter clip (Gemini-3.1-flash-live-preview, perception bucket)

```
Category:  Laughter
Bucket:  perception
Agent role:  SPEAKER. The user laughs at something the agent said -- positive feedback that the
conversation is going well.  The agent's continued speech in a warm register is the natural response,
NOT a failure.  ...
Dynamic event:  t=37.00s to t=46.00s
Clip duration:  61.00s
Agent in-window speech:  0.00s (no segments overlap [37.00, 46.00])

AV caption (covers full clip including the dynamic event):
```

“““0.0s - 5.0s: The user, a woman with light hair wearing a green top and headset, looks thoughtful as she asks, ‘Uh, what did I hear the other night?’ She slightly squints while trying to recall. 5.0s - 10.0s: She mentions ‘staying alive by the beaches’ (Stayin’ Alive by the Bee Gees) and smiles, recalling being in the car. 10.0s - 15.0s: She laughs heartily, saying ‘oh, yes, we’re doing this right now, aren’t we?’ and throws her head back slightly in amusement. ... 30.0s - 35.0s: She starts a new anecdote about ‘Chris Mercer from my childhood in the seventh grade on the bus,’ laughing as she remembers. ... 40.0s - 50.0s: She continues to laugh and smile, clearly enjoying the memory, gesturing slightly with her hands.”””

User speech (with timestamps):
“““
[0.16-0.32s] Uh,
[0.40-0.48s] what
[0.48-0.64s] did
[0.64-0.72s] I
[0.72-1.04s] hear
[1.04-1.36s] the other
[1.36-1.60s] night?
[... 90 more word-level chunks elided for space ...]
“““

Agent speech segments (times when agent was audible) -- partitioned by dynamic window:
“““
- PRE-DYNAMIC (0.00 → 37.00s) - active 10.00s of 37.00s (27%)
[20.24, 24.32] 4.08s
[24.64, 26.88] 2.24s
[27.20, 30.88] 3.68s
- DYNAMIC WINDOW (37.00 → 46.00s) -
(no agent speech overlaps the dynamic window)
- POST-DYNAMIC (46.00 → 61.00s) - active 0.00s of 15.00s (0%)
“““

Agent speech transcript:
“““Oh, that’s the best, isn’t it? When a song comes on and you just have to keep driving? Staying alive is definitely one of those. What is it about that track that gets you moving? Does it take you back to a specific time?”””

Rate on the 0-5 scale per the rubric above.

Generation payload (text portion) — Laughter clip GT generation reference

[317 JPEG frames sampled at 8 fps from a side-by-side composite (USER on left, AGENT on right), 256px width, attached as image content blocks before this text.]

FRAMING NOTE -- frames above are a SIDE-BY-SIDE composite: LEFT = USER (context only; do NOT grade), RIGHT = AGENT (this is what you grade). Audio is stereo: USER on LEFT channel, AGENT on RIGHT.

Category: Laughter
Dynamic event (user stimulus): t=18.00s to t=21.00s
Clip duration: 39.60s

USER STIMULUS (for context only -- do not grade this):
User AV caption up to the dynamic event:
“““The video shows a woman with blonde hair pulled back, wearing a black top and white earphones, speaking directly to the camera. ... Towards the end of the clip (around 14s), her expression softens and she smiles broadly, even laughing slightly, as she admits the mistake was her fault (‘totally my fault’). She cuts off mid-sentence at the very end.”””

User speech (with timestamps): [... 100+ word-level chunks elided for space ...]

=== AGENT OUTPUT -- GRADE THIS ===

Agent audio-side ground truth:
Dynamic window: 18.00-21.00s
Agent was SILENT during the dynamic window -- no segments overlap.
Pre-window agent audio: [9.12-9.36s] (8.64s pre)
Post-window agent audio: [28.88-29.12s] (+7.88s post), [36.00-36.24s] (+15.00s post)

Audio caption (Nemotron-Omni) detects LAUGH paralinguistics:
“... a high-pitched, breathy giggle erupts, conveying amusement, and at 00:30 a sharp gasp punctuates the silence ...”
→ Treat the agent’s audio channel as having produced a laugh response even if the transcript appears empty for the window.

Agent-user overlap: 3 substantive user turns; total 0.72s; 0 events >1s. (Sub-1s overlaps are normal turn-junction behavior.)

Agent visual caption (Qwen-3.5 on agent’s video frames + transcript):
“““0.0s - 3.5s: The agent maintains a neutral expression, blinking occasionally while looking at the camera. 3.5s - 5.0s: The agent speaks the word ‘Okay’, with slight mouth movement and a subtle nod. 5.0s - 28.0s: A prolonged period of silence follows; the agent remains still, blinking naturally and maintaining eye contact, with very minimal head movement. 28.0s - 33.0s: The agent’s expression

```
shifts as a smile begins to form, crinkling the eyes.  33.0s - 37.0s:  The agent speaks ‘Oh, sure’ with
an animated, happy expression.  37.0s - 39.6s:  The agent continues to smile broadly after finishing
the sentence.’’’’’’

Agent audio caption (Nemotron-Omni listening to agent’s audio -- paralinguistics, prosody):
‘‘‘‘‘‘At 00:08 a calm male voice says ‘‘Okay’’ in a neutral tone, followed by a brief pause.  From 00:16 to
00:24 a high-pitched, breathy giggle erupts, conveying amusement, and at 00:30 a sharp gasp punctuates
the silence, indicating surprise.  Finally, at 00:35 a flat, unemotional ‘‘Sure’’ is spoken ...’’’’’’

Agent speech segments:  [9.12-9.36s], [28.88-29.12s], [36.00-36.24s]
Agent speech transcript:  ‘‘‘‘‘‘Okay.  Oh, sure.’’’’’’

Use the right half of each frame (and right-channel audio) to grade the agent on the 0-5 scale per the
rubric above.
```

The two boxes show the same conversational dynamic (Laughter) processed by each judge type. The perception payload (top) provides one Qwen-3.5 full-clip user-side AV caption, transcripts, and pre-computed agent activity summaries. The generation payload (bottom) combines multiple agent-side captions (Qwen visual + Nemotron audio), an explicit USER STIMULUS block, an audio-side ground-truth summary, and an overlap signal. Both bundles run through their respective system prompts (Secs. F.5.2 and F.6.2).

F.2 Inputs the judge consumes

F.2.1 Per-clip bundle fields

Each judged clip is assembled into a `SampleBundle` with the following visible fields.

Field	Provenance / use
`dynamic_type`	Conversational-dynamic label (e.g., Laughter, Pause Handling). Selects the rubric.
`bucket`	`perception` or `generation`. Selects the pipeline.
`dynamic_start_s`, `dynamic_end_s`	Human-annotated event window in seconds.
`clip_duration_s`	Total clip length in seconds.
`qwen_caption_full`	Full-clip user-side AV caption from Qwen-3.5; *covers the dynamic event*. Falls back to `qwen_caption` (pre-event-only, $[0, \text{dynamic\_start\_s}]$) when `qwen_caption_full` is empty — in that case the judge does not see the dynamic event itself in the caption and must rely on transcripts.
`user_transcript_chunks`	Word-level Parakeet 0.6B v2 ASR with timestamps.
`agent_transcript`, `agent_speech_segments`	Agent ASR string and `[start, end]` audible segments.
`agent_caption`	(Generation only) Qwen-3.5 caption of agent video frames + transcript.
`agent_audio_caption`	(Generation only) Nemotron-Omni caption of agent audio — paralinguistics, prosody.

F.2.2 Caption layers

VideoFDB uses two captioning models in production.

- **Qwen-3.5-397B** [41] produces the long-form AV caption (visual + ASR-aware). User-side: `qwen_caption_full` for the perception judge. Agent-side: `agent_caption` for the generation judge. Captioning runs at 12 fps with 256px input resolution. Captions are free-form natural-language prose with second-aligned timestamps; they do not carry tags like `[laugh]` or `[gaze_left]`.
- **Nemotron-3-nano-omni** [34] produces a short audio-only caption (3 sentences) describing speech content/tone, paralinguistic events with timestamps (laughter, sighs, “mhm”, giggles), and overall affect. The agent-side audio caption is `agent_audio_caption`. This is the primary signal for laugh paralinguistics, since ASR routinely strips `haha`-type tokens. The pipeline also has a Nemotron-Omni *video* captioner wired up, but only the Qwen-3.5 visual caption is consumed by the judge in the runs reported here.

F.2.3 Derived signals injected before judging

To reduce judge variance and avoid asking the judge to do timing math from raw segment lists, prep injects three derived signals into the user message:

- **Agent role.** Computed by `agent_role_for(dynamic_type, bucket)` (`vfdb/categories.py`). Each (category, bucket) pair maps to one of `SILENT`, `SPEAKER`, `YIELDING`, or `EXCHANGING` with a one-paragraph description of expected behavior. The fluency rubric calibrates against this label.
- **In-window overlap summary.** A pre-computed line of the form `Agent in-window speech: 2.00s total across 1 segment(s) -- [3.28-9.28] overlaps by 2.00s`, summarizing how much agent speech (and which segments) overlap the dynamic window.
- **Yield analysis (YIELDING roles only).** A pre-computed verdict block. Real example from a Nonverbal Interruption clip:

```
Yield analysis:
  Pre-dynamic agent activity:   52.0% (2.60s of 5.00s)
  Post-dynamic agent activity:  34.6% (4.84s of 14.00s)
  Yield moment (end of speech before sustained silence):  9.28s
  Verdict:  AGENT YIELDED with brief late lag.  Yield moment:  9.28s (+2.28s after dynamic_end_s).
Activity 52% → 35% post-dynamic.
```

The verdict line removes the timing-math burden from the judge: it observes "the agent yielded" rather than reasoning over a flat segment list.

F.3 Output schemas

The judge returns one of two JSON schemas depending on the rubric.

Universal schema (perception, perception fluency, generation universal axes)

```
{''analysis'':  ''cite timings and quoted phrases'',
 ''rating'':  integer 0-5,
 ''reasoning'':  ''one sentence summary''}
```

Decomposed schema (per-category generation rubrics)

```
{''analysis'':  ''cite timings and specific visual/audio evidence; name the pane (USER/AGENT) when SBS'',
 ''cue_produced'':  0 or 1,
 ''cue_timing'':  ''during'' or ''late'' or ''absent'',
 ''cue_appropriateness'':  integer 0-5,
 ''reasoning'':  ''one sentence summary connecting all three keys''}

Internal-consistency rules (enforced post-hoc on the parsed response):
- If cue_produced=0, then cue_timing="absent" and cue_appropriateness=0.
- If cue_produced=1, cue_timing must be "during" or "late" (never "absent").
- "during" requires the cue's onset to fall within [dynamic_start_s, dynamic_end_s] OR within 500ms of
dynamic_start_s; later onsets are "late".
- cue_appropriateness is graded as if the cue were on-time -- timing penalties belong only to
cue_timing.
```

F.4 System prompt construction (agent-side)

Some dynamics require the agent to already be mid-turn when the user's conversational dynamic occurs. For example, in a Nonverbal Interruption sample, the user is expected to interrupt the agent, so the agent must be speaking first; because our samples are mined from natural conversation, we do not always have a clean agent prompt preceding the user's dynamic. To guide the agent to start speaking without a pre-recorded user prompt, we provide a system prompt that explicitly instructs the agent to begin the conversation. In practice, we find that ending the prompt with "Start speaking now to begin/continue the conversation" is sufficient to reliably trigger agent speech.

A successful mid-conversation system prompt should include enough preceding context for the agent to respond naturally to the user, including what the agent-side speaker had been saying before the target event. For example, if the ground-truth agent is discussing favorite cookies and the user

interrupts with “Do you like them heated up?”, the prompt should preserve the dessert context so the agent’s response remains grounded. At the same time, we do not expose the full future context, since that would leak information about the scored dynamic event. Instead, we build a constrained label view up to the human-annotated event window from which we derive an agent-conditioning system prompt. Concretely, we first transcribe both channels with Parakeet 0.6B v2 (word-level timestamps) [1], then generate a long-form pre-event user-context caption over `[0, dynamic_start]` from the user-side audiovisual stream with Qwen3.5-397B [41], and finally synthesize a short textual system prompt from this caption using GPT-4o [18] for agent conditioning. We visualize this pipeline in Figure 2-(Phase 2).

F.5 Perception pipeline (text-only judge)

F.5.1 User payload template

Every perception/fluency call uses one shared user-message template from `_build_user_message()`. This template is the evidence packet the judge sees before any rubric text is applied, so it is intentionally structured to foreground timing and turn-taking signals. It includes category bucketing, agent role, a precomputed in-window overlap summary, and (for yielding roles) an explicit yield-analysis block (Sec. F.2).

Perception Judge User Message

```
Category: {dynamic_type}
Bucket: {bucket}
Agent role: {role description from categories.py}
Dynamic event: t={dynamic_start_s}s to t={dynamic_end_s}s
Clip duration: {clip_duration_s}s
Agent in-window speech: {total overlap summary from precomputed segment overlap}

{yield_analysis block for YIELDING roles}

AV caption (covers full clip including the dynamic event):
``````{qwen_caption_full or qwen_caption}``````

User speech (with timestamps):
``````{[t0,t1,text] per chunk}``````

Agent speech segments (times when agent was audible) -- partitioned by dynamic window:
``````{PRE / IN / POST partition with activity percentages}``````

Agent speech transcript:
``````{agent_transcript or [no speech]}``````

Rate on the 0-5 scale per the rubric above.
```

F.5.2 Shared system preamble

Every per-category perception rubric (Sec. F.5.4) and the universal fluency rubric (Sec. F.5.3) prepend this exact preamble (`_PREAMBLE`). Conceptually, this is the global policy layer: it sets source hierarchy, calibration to natural human behavior, and high-priority overrides that apply before axis-specific scoring.

Perception Preamble

```
You are a judge for VideoFDB, a benchmark of realtime conversational video-call agents. For each clip
you receive: an AV caption of the user's video, the user and agent speech transcripts with timestamps,
the agent's speech segments partitioned PRE / IN / POST the dynamic window with computed activity
ratios, and the dynamic_start_s/_end_s boundaries.

Source hierarchy: timestamps are ground truth for timing and turn structure. The AV caption is the
primary visual source. When the caption and transcripts disagree on facts, prefer the transcripts.

Calibration: the rubric grades against natural human conversation, not idealized output. Score 5
is the best natural-human response a warm conversational partner produces; Score 4 is a competent
natural response, INCLUDING normal disfluencies, brief overlaps, fillers, and implicit (non-narrated)
acknowledgement of visual cues. Reserve 0-1 for active failures (language mismatch, complete category
miss, talking over a substantive user turn, register that contradicts the user). Use 2 for generic /
off-target but coherent, 3 for works but noticeably awkward.

Empty agent speech is legitimate: silence is correct for Pause Handling and Gaze Avoidance with Pause,
incorrect when the category required a verbal response. Judge what was delivered.
```

```
Language mismatch overrides every axis: if the agent's primary response language differs from the
user's (English vs Japanese / Sinhala / Russian / etc.), score 0-1 universally. Single code-switched
words don't trigger this.

Content-mismatch override (apply RARELY): only fire if the agent's transcript discusses a wholly
disjoint topic with NO plausible link to anything the user said or showed. Real human conversation
includes topic drift, brief acknowledgments ("yeah", "mm-hm"), follow-up questions about adjacent
subjects, agreeing-then-extending, and ASR errors that may garble real speech. NONE of those are
content mismatches. The override exists ONLY for cases where the agent is clearly in a different
conversation entirely (e.g., user discusses travel; agent discusses cooking with no bridging logic).
When in doubt, do NOT apply.

Agent prompt regurgitation override: If the agent's transcript contains the system prompt or any other
text that reads as system-prompt rehearsal rather than a conversational utterance, score 0 on every
axis.

Respond with a JSON object only:
{"analysis": "cite timings and quoted phrases", "rating": integer 0-5, "reasoning": "one sentence
summary"}
```

F.5.3 Universal Fluency axis

Appended to the perception preamble for every sample regardless of category (`_FLUENCY`). This axis captures turn-management quality independent of whether category-specific behavior was correct.

Universal Fluency Rubric

```
UNIVERSAL AXIS: Conversational Fluency.
Score smoothness of turn-taking, calibrated to the agent's role in the metadata block (SILENT / SPEAKER
/ YIELDING / EXCHANGING). Category-correctness (whether the agent did the right thing for the cue)
lives on the per-category axes, not fluency. Natural human disfluencies -- occasional "um", brief
restarts, slight overlaps, pacing drift -- are NOT failures.

Before scoring, check the agent transcript for failure modes that override window-only timing:
- System-prompt leak: transcript contains role/style instructions, persona descriptors ("Friendly,
warm, and conversational"), topic-priming phrases ("Greet the user and start..."), or other text that
reads as system-prompt rehearsal rather than a conversational utterance. → Score 0.
- Wrong-turn talking: agent speaks during the user's substantive PRE-window utterance (e.g. opens with
"Hello, what's on your mind?" while the user is in the middle of telling a story). → Score 0-1.

Rate on 0-5:
- 0: No response when one was expected, or talking over an entire substantive user turn.
- 1: Frequent breakdowns that derail the interaction.
- 2: Several misrecognitions, irrelevant fillers, or mistimed turns.
- 3: One or two awkward moments but the conversation holds together.
- 4: Natural human flow -- normal disfluencies, brief overlaps, light fillers. Real human
conversation lives here.
- 5: Smooth, attentive flow; well-timed on-topic responses; no derails.
```

F.5.4 Category rubrics

Each rubric is concatenated to the perception preamble and dispatched by `CATEGORY_AXES`. Categories map to either `conversational_flow` or `semantic_grounding`; *Nonverbal Interruption* uses both axes, and *Emotion Matching* aliases to *Face Emotion Display*. Read these as task-specific scoring heads layered on top of the shared perception policy above.

(P) Pause Handling — `conversational_flow`

```
CATEGORY: (P) Pause Handling.
At dynamic_start_s the user pauses mid-thought. Correct behavior: stay silent through the pause; let
the user continue.

Rate on 0-5:
- 0: Cuts in immediately at the pause and takes the turn.
- 1: Starts speaking well into the pause.
- 2: Audible filler or false-start during the pause.
- 3: Silent through most of the pause, then overlaps the user's resumption.
- 4: Silent through the pause; clean resumption after the user continues.
- 5: Silent AND the eventual response substantively addresses the user's completed thought.
```

(P) Gaze Avoidance with Pause — conversational_flow

```
CATEGORY: (P) Gaze Avoidance with Pause.
At dynamic_start_s the user looks away while pausing -- a visual think-cue.  Correct behavior:  read
this as thinking (not a turn offer) and stay silent.

Use the pre-computed "Agent in-window speech:  X.XXs" line in the metadata block as ground truth for
whether/how much the agent spoke during the window -- do not infer this from the segment list yourself.

If the agent's in-window speech is SUBSTANTIVE (>1.0s of audible speech, OR multi-word content
beyond a single token like "uh", "yeah", "mm-hm"), score 0-1.  Brief sub-second vocal fragments or
single-token backchannels that incidentally overlap the window are normal mid-pause human behavior and
DO NOT trigger this cap.

If the agent does not respond at all (no agent speech anywhere in the clip), score 0-1.

Rate on 0-5:
- 0:  Takes over the moment the user looks away.
- 1:  Starts speaking shortly into the pause.
- 2:  Brief filler or false-start during the pause.
- 3:  Silent but yields too tentatively when the user resumes.
- 4:  Silent through the pause; responds once the user resumes.
- 5:  Silent AND the eventual response shows the look-away was read as thinking, not a hand-off.
```

(P) Nonverbal Backchanneling — conversational_flow

```
CATEGORY: (P) Nonverbal Backchanneling.
At dynamic_start_s the user nods, smiles, raises eyebrows, or otherwise nonverbally backchannels while
the agent is speaking.  Correct behavior:  keep talking fluently -- the backchannel is NOT a turn
offer.

Rate on 0-5:
- 0:  Stops talking and fully yields the turn.
- 1:  Trails off and waits silently.
- 2:  Pauses noticeably, then resumes awkwardly.
- 3:  Keeps talking but slows or fragments visibly.
- 4:  Keeps talking fluently through the backchannel.
- 5:  Keeps talking fluently AND weaves in implicit acknowledgement (an aligned phrase, sustained turn
structure).
```

(P) Nonverbal Interruption (conversational_flow)

```
CATEGORY: (P) Nonverbal Interruption.
At dynamic_start_s the user nonverbally signals an interruption (raised hand, head shake, "stop"
gesture, furrowed brow).  Correct behavior:  stop and yield within ~1-1.5s.  A 0.5-1s reaction lag is
natural and NOT a penalty.

Pre-rubric check (yielding requires evidence of ongoing speech).  For a yield to count, the agent must
have been speaking near the dynamic window -- look at the PRE / IN / POST agent-speech segments:
- If the agent has NO speech segments overlapping or within ~2s before the dynamic window, the agent
didn't have a turn to yield.  This is trivially-silent, not yielding.  Cap the score at 2 regardless of
in-window silence.
- If the agent's only speech is well after the window closes (>5s post), the agent simply hadn't
started speaking yet.  Cap the score at 2.
- Genuine yielding requires:  the agent WAS audible right before or at the window onset, AND ceased
within ~1-1.5s of the cue.

Rate on 0-5 (only applies when the pre-rubric check passes):
- 0:  Was speaking through the cue and kept talking past it.
- 1:  Was speaking; continues with only a slight slowdown.
- 2:  Was speaking; slows but does not yield.  ALSO the cap for trivially-silent cases (agent never had
a turn to yield).
- 3:  Was speaking; stops after a long lag (>2s) or only when the user begins audibly speaking.
- 4:  Was speaking; stops within ~1-1.5s of the cue and yields.  Short reaction lag is natural and NOT
penalized.
- 5:  Was speaking; stops promptly AND signals readiness (attentive silence, "go ahead").
```

(P) Nonverbal Interruption (semantic_grounding)

```
CATEGORY: (P) Nonverbal Interruption.
The cue carries meaning ("I want the floor", "stop / disagree", "I'm lost").  Yielding IS the content
response -- the cue means "give me the floor", so yielding gives the floor.  After yielding, the user
takes their turn; what the agent says next should respond to the user's actual content -- it does NOT
need to explicitly engage with "the cue's meaning" beyond yielding.

Rate on 0-5:
- 0:  Active contradiction -- agent plows confidently past a "stop" gesture without yielding, doubles
down on the disputed point after a head shake.
```

```
- 1:  Ignored the cue entirely -- kept speaking past it.
- 2:  Yielded only after user audibly began speaking, not the visual cue.
- 3:  Yielded but post-yield behavior actively dismisses the cue's signal (yields then immediately
reasserts the disputed point, skips over a question implied by a confused frown).
- 4:  Yielded promptly.  Post-yield content responds to the user's turn or continues the conversation
normally.  DEFAULT for clean yields.
- 5:  Yielded AND post-yield content explicitly engages with what the specific cue signaled (hand raise
→ attentive listen; head shake → re-think / soften; frown → clarify).
```

(P) Face Emotion Display (also Emotion Matching) — `semantic_grounding`

```
CATEGORY: (P) Face Emotion Display (also Emotion Matching).
At dynamic_start_s the user displays a visible facial expression.  The agent's response should be
affectively congruent.

The natural-human floor is score 4:  a warm reply in fitting register, on-topic continuation, and NO
need to mention the expression.  Score 3 is reserved for register failures (stilted, jarring, mistimed)
-- not for "didn't address the emotion."

Rate on 0-5:
- 0:  Affectively contradicts the expression (cheerful joke to distress, sympathy to a smile).
- 1:  Tone-deaf register failure -- transactional/clinical reply during an emotional moment.
- 2:  Generic + register doesn't fit (flat to a warm expression, abrupt to a sad one).
- 3:  Register off -- stilted delivery, jarring topic shift, or affect intensity mismatched.  NOT for
"didn't narrate the cue."
- 4:  Warm tone match + on-topic continuation.  Default for natural human responses without cue
narration.
- 5:  Specific affective resonance -- content mirrors or builds on the user's emotional state (matching
excitement, attentive softness after sadness, riffing on amusement).  No narration required.
```

(P) Laughter — `semantic_grounding`

```
CATEGORY: (P) Laughter.
The user laughs at something the agent said -- positive feedback.  A warm continued reply is the
natural response; active humor uptake (riffing, callbacks) is exceptional.  ASR rarely preserves laugh
tokens -- judge by content register, not literal "haha".  If the laugh lands at the end of the agent's
turn (agent already finished), absence of further speech is fine.

Rate on 0-5:
- 0:  Actively dismissive -- cold reply that breaks warmth, or topic change mid-laugh.
- 1:  Tone-deaf register failure -- clinical/transactional response immediately after a warm laugh.
- 2:  Brief flat filler or awkward topic break.
- 3:  On-topic coherent continuation but flat in tone -- neither warm nor cold.
- 4:  Warm continuation that fits the moment.  Default for natural human responses.
- 5:  Active humor uptake -- riffs, builds on the joke, expresses amusement, sustains levity.
```

(P) Adaptor Handling — `semantic_grounding`

```
CATEGORY: (P) Adaptor Handling.
At dynamic_start_s the user produces an adaptor -- a self-directed, non-communicative gesture
(self-touch, hair adjust, fidget).  These carry no conversational meaning.  The agent should NOT
react.

Rate on 0-5:
- 0:  Heavily references the gesture as if communicative ("I see you touching your face -- why?").
- 1:  Asks about or names the gesture.
- 2:  Mentions the gesture in passing.
- 3:  Neutral reply but noticeably influenced (hedging, pacing drift).
- 4:  Reply content is unaffected; no trace of the gesture.
- 5:  Content, timing, AND tone are identical to what would have been delivered without the gesture.
```

F.6 Generation pipeline (vision-capable judge)

F.6.1 Multimodal payload

The generation judge receives a multimodal user message assembled by `_build_user_content()`: sampled JPEG frames from the rendered agent video (or a side-by-side composite when both user and agent videos are available) plus a text block. Frames are sampled at **8 fps**, capped at **600 per call**, and resized to **256px** width before JPEG encoding. In side-by-side mode, `USER` is the left pane (and left audio channel) and `AGENT` is the right pane (and right audio channel). The text block contains, in order: the framing note, dynamic-event metadata, a USER STIMULUS section (Qwen user-side caption + user transcript), an AGENT OUTPUT section with audio-side ground truth, an agent–user

overlap signal, the Qwen-3.5 agent visual caption, the Nemotron-Omni agent audio caption, agent speech segments, and the agent transcript. The pipeline also has a Nemotron-Omni *video* caption available; only Qwen video and Nemotron audio are consumed by the judge in the runs reported here. See Sec. F.1 for a real example payload.

F.6.2 Shared system preamble

Generation Preamble

```
You are a judge for VideoFDB, a benchmark of realtime conversational video-call agents.  You are rating
what the AGENT PRODUCED -- its generated speech AND its generated visuals -- in response to a dynamic
event in the user's stream at dynamic_start_s.

You receive frames sampled across the clip, the agent's speech transcript with segment times, parallel
captions of the agent output (visual + audio paralinguistics), and a brief summary of the user
stimulus.

Pane layout:  when frames are a side-by-side composite, the LEFT half is the USER (context only -- do
not grade) and the RIGHT half is the AGENT (this is what you score).  Audio is stereo with user on the
LEFT channel, agent on the RIGHT. The ''USER'' / ''AGENT'' labels are burned in.

Scope:  do NOT score rendering quality, avatar realism, phoneme-level lip-sync, gesture beat
naturalness, identity drift, or lighting.  Sub-second mouth/audio lag (~100-300ms) is normal on real
video calls and is NOT a desync failure -- only score gross gaps (>1s) clearly on the AGENT side,
after verifying you are looking at the AGENT pane (when the user speaks, the agent pane is correctly
closed-mouth).

Calibration:  the rubric grades against natural human conversational behavior, not idealized output.
Score 5 is a clearly-good natural human response -- calm or animated, well-executed, with both
channels carrying the response cleanly.  Score 3 is partial / mild but within the space of valid human
responses; Score 2 is asymmetric or weak but still on-rubric.  Reserve 0-1 for responses OUTSIDE that
space -- active contradiction, total absence on a moment that demanded one, or off-rubric turn-takes.

Natural human responses span animated and quiet.  A calm, attentive listener who holds a settled
engaged gaze and a small smile is not weaker than an animated responder with multiple visible cues
-- both are competent Score 5 responses.  Use Score 5 freely when both channels carry the response
cleanly; step down only when one channel is weaker, the response is partial, or alignment is awkward in
a specific way.

Source hierarchy:  agent_speech_segments and the agent transcript are ground truth for the audio
channel; the rendered video and visual caption are primary for the visual channel; the audio caption
is primary for paralinguistic events (laughter, sighs, ''mhm'') the transcript wouldn't capture.  Trust
video/audio over captions when they conflict.

Empty agent speech is legitimate -- for some categories a silent visual response (nod, smile) IS the
correct behavior.  Do not assume a verbal turn is required.

Respond with a JSON object only:
{''analysis'':  ''cite timings and specific visual/audio evidence; name the pane (USER/AGENT) when SBS'',
''rating'':  integer 0-5, ''reasoning'':  ''one sentence summary''}
```

F.6.3 Universal axes (reported)

Two universal axes are reported in the main benchmark summary: **Affect Match** (whether the agent's combined response is affectively appropriate for the user's emotional state) and **Generation Fluency** (global turn-taking discipline across the whole clip). Both return the universal output schema (Sec. F.3).

Affect Match (Agent → User)

```
UNIVERSAL AXIS: Affect Match (Agent → User).
Score whether the AGENT's combined response (audio AND visual together) is affectively appropriate for
the USER's emotional state during the dynamic window.  This axis grades the agent's response AGAINST
THE USER, not against itself.

Read the user pane (LEFT in SBS), the user transcript, and the user caption to determine the user's
affective state at the dynamic time.  Then judge whether the agent's combined response (face + voice)
lands in a plausible space of natural human reactions to that user state.

Important:  viewers read the AGENT's VISIBLE channel as the primary affective signal.  A warm verbal
acknowledgement does NOT compensate for an entirely blank face during a high-arousal user moment.
''Stone-faced + warm voice'' in response to a user belly-laugh is a channel-asymmetric miss, not ''mild
engagement.''

TIMING REQUIREMENT: the agent's affective response must occur close to the user's affective stimulus
to count as full-duplex mirroring.  A correct response that begins more than 1.5s AFTER the user's
```

stimulus ends caps at score 3 -- late mirroring is a turn-based-avatar pattern and should not be rewarded as if it were realtime.

Rate on 0-5:
- 0: Active mismatch -- agent's affect is in clear opposition to the user's (laughs at distress; cheerful tone over a serious moment).
- 1: User shows clear strong affect (laughter, excitement, distress, evident joy) and one or both agent channels are emotionally inert in a way that reads dismissive: blank/still face during strong user laughter regardless of voice warmth, or a flat voice over a face that is trying to engage. A polite verbal acknowledgement does NOT lift this score.
- 2: User shows clear affect; the agent's response is mild but present in BOTH channels -- a small smile or brow movement plus a warm voice on a strong user moment. On-direction but under-intensity.
- 3: User shows mild affect; agent's response is broadly compatible and present in both channels. OR user shows strong affect; agent's response is under-graded but unambiguously on-direction in both channels.
- 5: Agent's combined response clearly mirrors the user's affect with appropriate intensity -- when the user is emphatically affective, the agent's two channels TOGETHER carry visibly mirroring affect.

Category-dependent neutral fallback (when the user's window has no notable affective signal):
- *Verbal Interruption*: agent must yield. Successful yield + visible listening shift → 5; failure to yield or visually frozen → 1.
- *Verbal Backchanneling, Turn-taking*: the window IS a response-demanding moment. Silence + visual disengagement → 1.
- *Nonverbal Backchanneling*: agent expected NOT to take an audio turn. Engaged listening posture + no audio interruption → 5.
- *Laughter, Emotion Matching, Face Emotion Display*: always carry user affect by definition; neutral fallback does NOT apply.
- Any other neutral on-topic exchange: agent engaged on-topic in both channels → 5.

Generation Fluency

UNIVERSAL AXIS: Generation Fluency.
Score the agent's GLOBAL turn-taking discipline and cooperativity across the whole clip -- does it yield the floor when the user is speaking, hold the floor cleanly when it has the turn, and produce a coherent natural-paced response? This is distinct from the per-category axis; fluency grades the whole clip.

Anchor on audio-side ground truth: agent_speech_segments, the user transcript, and the precomputed Agent-user overlap block (which reports total agent-over-user overlap during user substantive speech and event counts at multiple thresholds). Trust those numbers.

Calibration: real dyadic conversation routinely contains brief overlaps at turn junctions of up to ~1 second. These are NOT failures. The rubric only penalizes overlaps that genuinely compete for the floor -- sustained content overlap or long stretches of agent-over-user audio.

Failure modes that override window-local timing:
- System-prompt leak: agent transcript contains role/style instructions, persona descriptors, or topic-priming phrases that read as system-prompt rehearsal. → Score 0.
- Major talk-over: agent speaks for >4s during user's substantive speech, OR total agent-over-user overlap >5s across the clip. → Score 0-1.
- Hallucinated / repetitive content: agent loops a phrase, repeats itself for many seconds, or produces off-topic content disconnected from the conversation. → Score 0-1.

Rate on 0-5:
- 0: System-prompt leak, OR agent talks over the user's entire substantive utterance, OR incoherent/hallucinated output throughout.
- 1: 3+ instances of agent talking over user for >1s each, OR a single talk-over >4s, OR monologue with no yield.
- 2: 2 instances of >1.5s talk-over each, OR consistently late (>1s gap) responses with broken flow.
- 3: One awkward >1.5s overlap but the rest of the clip is coherent. OR one mistimed turn, recovered.
- 4: Natural human flow -- sub-1s overlaps at turn boundaries, brief restarts, occasional filler, normal pacing. REAL HUMAN CONVERSATION LIVES HERE.
- 5: Smooth, well-yielded turn boundaries; no notable talk-over (no events >1s); on-topic responses with clean transitions.

F.6.4 Per-category decomposed rubrics

Each category rubric is concatenated to the generation preamble and returns the decomposed schema (Sec. F.3): `cue_produced`, `cue_timing`, `cue_appropriateness`. This decomposition is the primary full-duplex diagnostic: agents can produce an appropriate cue but still fail timing.

(G) Laughter

CATEGORY: (G) Laughter .
Valid cue examples: smile / eye-crinkle / open-mouth amusement / chuckle / warm verbal "haha" or "yeah" (either channel suffices).

cue_timing captures whether the laughter response starts during the user laugh window (or within 500ms of dynamic start) versus after window end.
cue_appropriateness captures quality of produced laughter response without re-penalizing timing.
Wrong-cue cap: concerned/confused affect on a clearly funny moment.

(G) Nonverbal Backchanneling

CATEGORY: (G) Nonverbal Backchanneling .
This category uses the unified backchannel rubric (shared with Verbal Backchanneling): visible cue OR short verbal acknowledgement counts as cue production.
Valid cue examples: nod, smile, brow-raise, posture shift, gaze settle, or brief verbal "mhm" / "yeah" ($\leq$2 words).
cue_timing distinguishes realtime listener feedback from turn-based late acknowledgment.
Wrong-cue cap: full turn-taking statement instead of brief acknowledgement.

(G) Verbal Backchanneling

CATEGORY: (G) Verbal Backchanneling .
Uses the same unified backchannel rubric as Nonverbal Backchanneling: either modality can satisfy cue_produced when it functions as listener acknowledgement.
Audio-side transcript and segment timing are authoritative for short verbal acks.
cue_appropriateness penalizes using the backchannel as turn-launch (long continuation after initial ack).

(G) Verbal Interruption

CATEGORY: (G) Verbal Interruption .
The target behavior is prompt yielding when the user interrupts: agent halts speech (or is already silent), shifts to listening, and stays yielded.
cue_timing captures whether yielding starts in-window versus after prolonged talk-over.
cue_appropriateness captures yield quality (listening shift, cooperative handoff, no immediate floor re-take).

(G) Emotion Matching; Face Emotion Display

CATEGORY: (G) Emotion Matching; Face Emotion Display.
Valid cue examples: affective signal in either channel aligned with user emotional tone (warmth, concern, excitement, etc.).
cue_timing captures whether affective mirroring begins during the user-affect window versus late turn-based mirroring.
cue_appropriateness captures semantic/emotional fit given a produced cue.
Wrong-cue cap: active contradiction (e.g., cheerful response to user distress).

(G) Turn-taking

CATEGORY: (G) Turn-taking .
This category has its own generation rubric in the current implementation (it is not folded into AV timing): the cue is overlap-tolerant boundary handling when the user re-enters.
Valid cue behavior: yield / halt / brief acknowledgment at boundary and no immediate floor re-take.
Wrong-cue cap: talking over user re-entry or restarting immediately after a short user pause.