

---

# TOWARD FINE-GRAINED SOCIO-AFFECTIVE INTELLIGENCE IN FULL-DUPLEX CONVERSATIONAL AI

**Junjie Chen**

Hefei University of Technology  
jorji.chen@gmail.com

## ABSTRACT

Full-duplex conversational models increasingly support overlapping speech, interruption handling, and synchronized audiovisual behavior. Yet responsive interaction and realistic motion generation do not by themselves establish an ability to interpret subtle human behavior in psychological and conversational context. This mini-survey connects recent work on streaming interaction models, dyadic audiovisual datasets, parametric 3D human representations, and interaction-oriented benchmarks with research on socio-affective cognition and the limits of facial emotion inference. It distinguishes the observation of fine-grained facial and bodily events from uncertain, context-sensitive interpretations and from decisions about how an agent should respond. Building on that distinction, it proposes an open research agenda: explicit motion-parameter understanding, evidence-aware annotation, asynchronous affective reasoning, real-time full-duplex response generation, longitudinal personalization, and outcome-centered evaluation. The proposed architecture is a research hypothesis rather than a validated system.

**Keywords:** full-duplex conversational AI; socio-affective computing; micro-behavior; human motion; SMPL-X; FLAME; multimodal interaction; affective reasoning.

Imagine a conversation in which someone quietly says, “*I’m fine.*” There is a short pause before the answer. Their brow briefly tightens, and one hand moves toward the opposite arm. A conventional voice assistant may hear the words and proceed as if nothing happened. A video-language model may accurately describe the frown and hand movement. Neither capability, on its own, tells us what the person feels—or what the assistant ought to do next.

The pause could reflect discomfort, uncertainty, fatigue, concentration, or simply an idiosyncratic speaking habit. A thoughtful conversational agent should be able to **observe the behavior, consider multiple explanations in context, choose an appropriate response, and revise its interpretation as the conversation unfolds**. Perhaps it gently asks whether the person would like to elaborate. Perhaps the broader context suggests that no intervention is needed.

This small example points to a larger research question:

**How can we build psychologically grounded, fine-grained, multimodal, efficient, evolving, and full-duplex conversational AI systems that better understand and respond to human emotions?**

Recent advances in speech, video, and 3D motion modeling make the question timely. Yet a key distinction remains: producing synchronized human-like behavior is not equivalent to understanding the social and affective significance of that behavior. This post reviews the emerging technical landscape and outlines a possible path toward more context-aware, adaptive human–AI interaction.

## 1 FROM INTELLIGENT REASONING TO INTELLIGENT INTERACTION

Large foundation models have become increasingly capable of analysis, planning, and long-context reasoning. But talking with a person is not simply a sequence of well-formed answers. Conversation

---

is a **continuous, reciprocal process**: speakers overlap, hesitate, signal attention, offer backchannels, change their minds, and communicate through voice, face, gaze, and body movement.

Traditional spoken assistants often follow an explicit *listen* → *transcribe* → *reason* → *synthesize* pipeline. That design is convenient but can impose unnatural turn boundaries. A truly **full-duplex** model instead listens and speaks over overlapping time intervals, managing interruptions and short acknowledgments as they occur.

Kyutai’s [Moshi](#) (Défossez et al., 2024) helped establish a speech-native, multistream formulation for full-duplex dialogue. [PersonaPlex](#) (Roy et al., 2026) built on this direction, adding stronger voice and role conditioning while maintaining natural turn-taking and interruption behavior. More recently, [Thinking Machines Lab’s Interaction Models](#) (Thinking Machines Lab, 2026) proposed a particularly useful systems perspective: keep a time-aware interaction model continuously responsive while delegating sustained reasoning and longer-horizon work to an asynchronous background model.

These systems illustrate an architectural shift. **The ability to reason deeply and the ability to interact fluently need not reside in the same model at the same time scale.** For emotionally sensitive interaction, this distinction may be even more important: an agent should respond promptly without pretending that its first impression of someone’s internal state is necessarily correct.

## 2 WHAT DOES “FINE-GRAINED SOCIO-AFFECTIVE INTELLIGENCE” MEAN?

I use *fine-grained socio-affective intelligence* as a **working research term**, not as a claim that psychology has settled on a single taxonomy.

**Fine-grained** refers to the scale of the observable signal: a brief brow movement, a subtle head retreat, a change in gaze, a self-touch, a change in voice tension, or a pause that matters only relative to the ongoing exchange. The important dimensions are not merely spatial resolution but also **temporal precision, multimodal alignment, and sensitivity to conversational context.**

**Socio-affective** refers to the interaction between emotional processes and social understanding. In psychology and social neuroscience, affect-related processes such as empathy and compassion are often distinguished from socio-cognitive processes such as perspective taking or theory of mind, while also being understood as dynamically interacting capacities (Kanske, 2018). An artificial agent cannot be presumed to *feel* human empathy merely because it produces considerate language. A more operational goal is to perceive socially relevant evidence, reason cautiously about possible human states, and respond appropriately.

This is also why **micro-expression, micro-action, and micro-gesture** should not be presented as three mutually exclusive ontological categories. Micro-expression usually refers to a brief facial expression event studied within a particular experimental tradition; micro-action is a broader description of subtle movement; micro-gesture emphasizes movement serving, or potentially serving, a communicative function. The same small movement may fit more than one description depending on the annotation scheme and conversational context. What matters scientifically is defining observable events and their timing before assigning interpretations to them.

One psychological caution is non-negotiable: **a facial movement is not a direct readout of an emotion.** The influential review *Emotional Expressions Reconsidered* (Barrett et al., 2019) documents the limitations of inferring specific emotional states from facial configurations alone. An AI system that confidently converts a fleeting frown into “the user is anxious” may look intelligent while making an unjustified inference.

An improved system should distinguish at least three layers:

1. **Observation:** What happened, where, and when?
2. **Interpretation:** Which hypotheses are compatible with the behavior and context, and how uncertain are they?
3. **Interaction:** Would acting on those hypotheses improve the conversation, and can later evidence confirm or disconfirm them?

The third layer is the real destination. Fine-grained behavior matters to conversational AI only when it contributes useful information beyond speech and conversational history.

---

### 3 WHAT RECENT MODELS HAVE ALREADY ACHIEVED

The most relevant work spans several partially connected research streams.

- **Full-duplex speech** ([Moshi](#), [PersonaPlex](#)). Continuous listening and speaking; turn-taking, interruptions, and backchannels **Open question:** Are subtle visual and affective cues used appropriately, rather than merely detected?
- **Interactive systems architecture** ([Interaction Models](#)). Real-time interaction alongside asynchronous deeper reasoning **Open question:** How should background emotional hypotheses safely influence online behavior?
- **Dyadic audiovisual data and motion** ([Meta Seamless Interaction](#)). Large-scale two-person conversational footage and synchronized behavioral motion modeling **Open question:** Can naturally occurring micro-behaviors be annotated and interpreted in context?
- **Full-duplex speech–motion generation** ([NVIDIA DyaPlex](#)). Joint streaming perception and generation of speech and body motion **Open question:** Does synchronization translate into explainable, psychologically grounded affective understanding?
- **Audiovisual interaction evaluation** ([NVIDIA VideoFDB](#)). Behavioral evaluation grounded in dyadic conversational dynamics **Open question:** How much *incremental interaction benefit* comes from fine-grained cues?

Meta’s [Seamless Interaction](#) ([Agrawal et al., 2025](#)) is especially important for data. Its associated dataset includes **more than 4,000 hours of face-to-face interactions involving more than 4,000 participants**, and the research family explores dyadic audiovisual behavior, including gestures and facial expressions that align with speech. It is not, however, automatically a fully labeled dataset of psychologically verified micro-behaviors or internal emotional states. Those are different annotation problems.

NVIDIA’s [DyaPlex](#) ([Nagano et al., 2026](#)) makes a further substantial step. It couples a frozen full-duplex speech model with a trainable motion pathway, motion tokenization, and temporal alignment machinery. Its contribution is not simply “adding 3D to a voice model”: it explicitly models reciprocal speech–motion dynamics in a causal streaming setting, including partner motion and the agent’s motion. That is a strong starting point for the field.

Accordingly, a future research project should **not** claim novelty merely from introducing SMPL-X or FLAME parameters, adding a motion tokenizer, or producing speech-aligned gestures. The more interesting frontier is whether the model can use fine-grained motion **as contextual evidence**, reason about ambiguous interpretations, and improve its conversational decisions. That last capability is a research hypothesis, not something that existing architecture diagrams alone establish.

### 4 EXPLICIT BEHAVIORAL REPRESENTATIONS: WHAT CAN 3D CONTRIBUTE?

There are at least two broad approaches to modeling subtle behavior.

**Implicit representations** encode video or motion through learned latent features, such as video embeddings, motion tokens, or autoregressive discrete codes. They may be compact and effective for downstream prediction and generation, but the semantic relationship between an internal token and a particular observed motion is not necessarily transparent.

**Explicit representations** describe interpretable geometric or kinematic quantities: facial action units, gaze direction, joint angles, hand trajectories, body orientation, and parametric 3D human models. [SMPL-X](#) ([Pavlakos et al., 2019](#)) unifies body pose, articulated hands, and an expressive face; [FLAME](#) ([Li et al., 2017](#)) provides a parametric representation of facial shape, expression, and pose. Both make it possible to express subtle changes as temporally ordered parameter sequences rather than only as pixels.

Conceptually, a behavioral state could be written as

$$\mathbf{x}_t = [\theta_t^{\text{body}}, \theta_t^{\text{hand}}, \psi_t^{\text{face}}, \mathbf{g}_t, \mathbf{p}_t],$$

where the terms denote model-dependent pose, facial expression, gaze, and positional quantities. A small movement is often more visible in **change over time** than in a single state:

$$\Delta \mathbf{x}_t = \mathbf{x}_t - \mathbf{x}_{t-1}, \quad \Delta^2 \mathbf{x}_t = \Delta \mathbf{x}_t - \Delta \mathbf{x}_{t-1}.$$

Here the subtraction is shorthand: rotations and articulated poses may require geometry-aware differences rather than naïve subtraction of parameter vectors.

This suggests an unusually concrete first experiment: **Can an existing multimodal or agentic reasoning model understand a short sequence of SMPL-X/FLAME parameters directly, without training a new task-specific motion representation?** For example, can it identify which facial or bodily region moved, localize the change in time, compare two subtle trajectories, and offer an evidence-based description? How does it perform when given raw numeric parameters, normalized trajectories, structured textual descriptions, rendered 3D motion, or the original RGB video?

This is a falsifiable capability study—not an assumption that explicit 3D will win. Parametric reconstruction may smooth away high-frequency facial motion, introduce fitting artifacts, or omit important cues such as vocal hesitation and interpersonal distance. Raw parameter sequences may also be difficult for a general-purpose language model to interpret without a meaningful schema. These limitations are part of the research question.

## 5 FROM RECOGNITION TO EVIDENCE-EVOLVING REASONING

Suppose the system observes a brief frown during “I’m fine.” A brittle classifier might return a single emotion label. An evidence-oriented system should instead maintain a distribution—or at least a calibrated set—of plausible explanations.

Let  $o_{1:t}$  denote observations so far,  $c_{1:t}$  conversational context,  $m_t$  a personalized memory state, and  $z_t$  an unobserved affective or communicative state. The desired inference can be expressed conceptually as

$$b_t(z) = P(z_t = z \mid o_{1:t}, c_{1:t}, m_t).$$

This is **an architectural objective, not a claim that an arbitrary neural model outputs well-calibrated probabilities**. The latent state is not directly observable, and different sources of evidence may disagree. A useful implementation might track a few candidate hypotheses, their supporting and contradictory evidence, and whether a clarification question is warranted.

For example:

- **Observed:** 450 ms pause; brief brow tightening; arm touch.
- **Possible hypotheses:** uncertainty, discomfort, concentration, habitual movement.
- **Context:** a sensitive question was just asked, but the person has also displayed similar self-touch during neutral topics.
- **Possible response:** “We can take our time—would you like to say more?”
- **Follow-up:** update the interpretation based on the reply instead of treating the first hypothesis as fact.

The distinction between **behavioral evidence** and **psychological inference** also improves explainability. A system should be able to say *which observation affected a response*, without claiming privileged access to the person’s thoughts. This is particularly important in psychological interviews and other consequential settings, where an AI should support, rather than substitute for, qualified human judgment.

## 6 A TWO-TIMESCALE SYSTEM: REAL-TIME PRESENCE AND SLOWER REFLECTION

I find a **small–large model collaboration** especially promising, although the exact division of work must be established experimentally.

The proposed architecture has four connected elements:

1. **Fine-grained multimodal perception.** Streaming audio, RGB, facial signals, gaze, and optional 3D motion parameters yield timestamped behavioral observations and confidence estimates.
2. **Fast interaction path.** A compact causal full-duplex model handles timing, turn-taking, backchannels, immediate verbal responses, and—where applicable—nonverbal generation. It should not wait for a long chain of psychological reasoning before every acknowledgment.
3. **Slow deliberative path.** A larger reasoning model examines accumulated evidence, conversational history, ambiguity, and competing interpretations. Its output is an update or recommendation, not an authoritative diagnosis.
4. **Evolving personal model.** Consented, revisable memory tracks interaction preferences and individual behavioral baselines; feedback helps update both future interpretations and response strategies.

One way to capture the interaction policy is

$$\pi(r_t \mid o_{\leq t}, b_t, m_t),$$

where  $r_t$  includes both **what** the system says and **how/when** it responds. The fast model may operate on a compressed version of  $b_t$ , updated asynchronously by the slower model. Crucially, slow reasoning should be allowed to revise earlier impressions without retroactively making the online agent appear omniscient.

An important design choice is the interface between models. Instead of sending opaque advice such as “the user is sad,” the slow path could send a structured message: *possible discomfort; low confidence; avoid assuming cause; prioritize a gentle, optional follow-up*. That makes uncertainty actionable and creates a testable bridge from reasoning to behavior.

## 7 DATA AND AUTOMATIC ANNOTATION: SEPARATE MOVEMENT FROM MEANING

A central bottleneck is not just collecting more video. It is collecting **temporally precise, multi-level annotations** that do not confuse observation with interpretation.

High-frame-rate facial datasets such as **CASME II** (Yan et al., 2014), along with SAMM and SMIC, offer valuable benchmarks for short facial events. In contrast, natural dyadic datasets such as Seamless Interaction provide richer conversation but are not automatically annotated with every psychologically meaningful fine-grained event. The field needs bridges between these two regimes.

A practical annotation unit might be

$$\ell_i = (t_i^{\text{start}}, t_i^{\text{end}}, \text{channel}_i, \text{observable event}_i, \text{confidence}_i).$$

Only in a **separate interpretation layer** would annotators or models attach context-dependent hypotheses, possible communicative functions, and evidence quality. This separation matters because a system can be right about a hand movement while unjustifiably certain about the person’s feelings.

A scalable pipeline could combine (i) high-recall machine proposals from RGB, action-unit, gaze, 2D pose, or SMPL-X/FLAME trajectories; (ii) event segmentation and cross-modal alignment; (iii) model-assisted descriptive labeling; and (iv) expert or trained-human review of uncertain examples. Active learning can prioritize disagreements. For particularly sensitive labels, the appropriate gold standard may be multiple independent judgments, participant self-report, or experimentally validated context—not a single automated “emotion truth.”

## 8 EVALUATION: DOES THE SUBTLE SIGNAL ACTUALLY HELP?

This is perhaps the most important issue. A model can achieve high motion-recognition accuracy and still become a worse conversational partner by overreacting to incidental behavior.

Existing benchmarks already show how challenging interaction evaluation is. [Full-Duplex-Bench](#) (Lin et al., 2025) probes turn-taking, pause handling, interruptions, and backchannels; [TurnBench](#) (Jiang et al., 2026) examines spoken turn-taking dynamics across domains. NVIDIA’s [VideoFDB](#) (Mazumdar et al., 2026) adds an explicitly audiovisual view, with **237 expert-annotated clips covering 11 nonverbal conversational dynamics**. Its reported failure modes include **captioning collapse**, when an agent describes the image instead of participating naturally, and **visual-stream ignorance**, when visual information has little effect on the response. These are informative findings, not evidence that all models fail at every visual cue.

I would organize evaluation into three levels:

- **Perception:** Did the model detect the observable event?  
**Representative measurements:** Temporal localization, event F1, region attribution, cross-subject robustness, and annotation agreement.
- **Contextual Understanding:** Was the interpretation grounded and appropriately uncertain?  
**Representative measurements:** Evidence attribution, contradiction tests, hypothesis calibration, context sensitivity, and human/expert ratings.
- **Interaction Outcomes:** Did the signal improve the response?  
**Representative measurements:** Turn-taking latency, interruption behavior, response appropriateness, perceived support, controlled user preference, and task success.

The most revealing test is a **matched ablation**. Hold the conversational content, model capacity, latency budget, and evaluation protocol as constant as possible, then compare systems using (a) audio and conversation history, (b) audio plus RGB, (c) audio plus explicit 3D behavior, and (d) the same inputs plus contextual affective reasoning.

One simple target quantity is

$$\Delta_{\text{behavior}} = \mathbb{E}[S \mid \text{context} + \text{behavioral evidence}] - \mathbb{E}[S \mid \text{matched context only}],$$

where  $S$  is a preregistered interaction-quality measure. This difference should be estimated with uncertainty intervals and appropriate controls; by itself it does not prove causality unless the experimental design justifies that interpretation. Counterfactual tests—swapping a behavioral clip while preserving words, or introducing an irrelevant movement—can reveal whether a model truly uses the cues as evidence.

Objective and subjective measures need not agree. An agent might be fast yet intrusive, empathetic-sounding yet factually presumptuous, or behaviorally accurate yet socially awkward. Rather than forcing all scores into a single leaderboard number, evaluation should report the trade-offs and use blinded human judgments for constructs such as appropriateness and trust. The challenge is to measure whether the *additional behavioral information* improves the relationship between user and agent.

## 9 PERSONALIZATION AND CONTINUAL EVOLUTION—WITHOUT OVERCLAIMING

People differ in gaze habits, self-touch frequency, speaking rhythm, gesture size, and preference for emotional support. The same movement can carry different meanings across people and situations. A useful long-term agent should therefore learn **individual baselines**, not just generic mappings from motions to emotion labels.

But *evolving* should mean more than storing a longer transcript. It means updating beliefs when new evidence arrives, adapting response strategies based on feedback, and preserving distinctions between relatively stable preferences and momentary states. Long-term user modeling also creates privacy and safety obligations: informed consent, inspectable memory, user-controlled deletion, non-manipulative personalization, and explicit limits in high-stakes contexts.

The appealing idea of **psychological mirroring** should be treated as a research question rather than a default design goal. Some alignment in pacing, warmth, or gesture may help communication, but

---

indiscriminate emotional mirroring can be inappropriate. An agent must learn when to align, when to remain neutral, and when to ask instead of infer.

## 10 A CONCRETE RESEARCH AGENDA

The preceding literature suggests a progressive program rather than one monolithic model-training project.

**Step 1 — Test explicit 3D understanding.** Before designing a new representation learner, investigate what current multimodal and agentic models can do with raw or structured SMPL-X/FLAME sequences. Benchmark micro-motion spotting, parameter-to-language explanation, temporal comparison, and failure modes, against RGB and learned motion-token baselines. This establishes whether explicit parameters provide a useful interface to reasoning at all.

**Step 2 — Build a fine-grained behavioral evidence layer.** Develop machine-assisted temporal annotation and aligned multimodal evidence, with human review and uncertainty. The aim is not to label every gesture with a single emotion but to make observed behavior reliably queryable across time and context.

**Step 3 — Train context-aware affective reasoning.** Study whether combining observed behavior, speech, history, and multiple hypotheses yields better-grounded interpretations than event classification alone. Favor evidence attribution and uncertainty calibration over confidently phrased psychological guesses.

**Step 4 — Bring understanding into full-duplex generation.** Connect the evidence and reasoning layers to a streaming interaction policy. Evaluate whether they change *when* the agent speaks, *what* it says, and *how* it coordinates nonverbal responses. This is the point at which fine-grained understanding either proves its value or fails the ablation test.

**Step 5 — Personalize and evaluate longitudinally.** Test whether revisable memory and user feedback help across repeated interactions, while preserving user agency and privacy. Demonstrate the system first in constrained, consent-based dyadic settings—for example, a supportive interview or personal companion—rather than claiming diagnostic competence.

The eventual vision is an **affect-aware full-duplex conversational agent** with a fast online interaction path, slower contextual reasoning, and a continually revisable personal model. I tentatively call this direction *SocioVolve*, emphasizing that social-affective interpretation should *evolve with evidence and interaction*.

## 11 CLOSING: BEYOND SYNCHRONIZED SPEECH AND MOTION

It would be a mistake to underestimate what recent research has already accomplished. Full-duplex speech models have made turn-taking more natural. Dyadic motion datasets and models have brought expressive body behavior into the loop. DyaPlex shows that simultaneous speech–motion perception and generation can be modeled within a coherent streaming architecture. VideoFDB makes it harder to mistake visual question answering for genuinely audiovisual interaction.

Yet there is still a meaningful distance between **observing a subtle action, interpreting it with appropriate uncertainty, and using it to improve a live human conversation**. Crossing that distance requires psychology-informed behavioral definitions, explicit or testable representations, context-sensitive reasoning, responsive generation, personal adaptation, and evaluation centered on human outcomes.

The goal is not an AI that claims to read minds. It is an AI that **notices more, assumes less, responds better, and learns responsibly**.

**Beyond synchronized speech and motion, toward psychologically grounded, fine-grained socio-affective intelligence that understands, responds, and evolves through interaction.**

---

---

## REFERENCES

- Vasu Agrawal et al. Seamless interaction: Dyadic audiovisual motion modeling and large-scale dataset, 2025. URL <https://arxiv.org/abs/2506.22554>. See arXiv for the complete author list.
- Lisa Feldman Barrett, Ralph Adolphs, Stacy Marsella, Aleix M. Martinez, and Seth D. Pollak. Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest*, 20(1):1–68, 2019. doi: 10.1177/1529100619832930. URL <https://doi.org/10.1177/1529100619832930>.
- 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, 2024. URL <https://arxiv.org/abs/2410.00037>.
- Freeman Jiang, Ramon Sanabria, Soham Deshmukh, Bandhav Veluri, Simon Michael Vuch Williams, Elliott K. Suen, Garreth Lee, Kevin Yoonho Choi, Takuya Umeki, Riku Kubo, Sathvik Udupa, Chien-yu Huang, Shih-Yun Shan Kuan, Zhuoyan Tao, Satyapriya Krishna, Sefik Emre Eskimez, Yu Tsao, Hung-yi Lee, and Shinji Watanabe. TurnBench: A multi-domain benchmark for turn-taking dynamics in spoken dialogue, 2026. URL <https://arxiv.org/abs/2608.25218>. Accepted at IEEE SLT 2026.
- Philipp Kanske. The social mind: Disentangling affective and cognitive routes to understanding others. *Interdisciplinary Science Reviews*, 43(2):115–124, 2018. doi: 10.1080/03080188.2018.1453243. URL <https://doi.org/10.1080/03080188.2018.1453243>.
- Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics*, 36(6):194:1–194:17, 2017. doi: 10.1145/3130800.3130813. URL <https://flame.is.tue.mpg.de/>.
- 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, 2025. URL <https://arxiv.org/abs/2503.04721>. Accepted at IEEE ASRU 2025.
- Amrita Mazumdar, Seonwook Park, Rajarshi Roy, Nikhil Srihari, Shengze Wang, Yuhao Zhou, Julia Wang, Koki Nagano, and Shalini De Mello. VideoFDB: Evaluating full-duplex vision-speech capabilities in conversational agents, 2026. URL <https://arxiv.org/abs/2605.30256>.
- Koki Nagano, Hongyu Liu, Seonwook Park, Tianye Li, Amrita Mazumdar, Christian Jacobsen, Shengze Wang, Michael Stengel, Rajarshi Roy, Ka Chun Cheung, Simon See, and Shalini De Mello. DyaPlex: Full-duplex speech-motion model for dyadic interaction, 2026. URL <https://arxiv.org/abs/2606.03874>.
- Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10975–10985, 2019. doi: 10.1109/CVPR.2019.01123. URL <https://smpl-x.is.tue.mpg.de/>.
- 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>.
- Thinking Machines Lab. Interaction models: A scalable approach to human-AI collaboration. *Thinking Machines Lab: Connectionism*, May 2026. doi: 10.64434/tml.20260511. URL <https://thinkingmachines.ai/blog/interaction-models/>. Research article on the author’s laboratory website.
- Wen-Jing Yan, Xiaobai Li, Su-Jing Wang, Guoying Zhao, Yong-Jin Liu, Yu-Hsin Chen, and Xiaolan Fu. CASME II: An improved spontaneous micro-expression database and the baseline evaluation. *PLOS ONE*, 9(1):e86041, 2014. doi: 10.1371/journal.pone.0086041. URL <https://doi.org/10.1371/journal.pone.0086041>.