

ConSensus: Multi-Agent Collaboration for Multimodal Sensing

Cited by 3
ACL 2026 Findings

Motivation: Challenges in Multimodal LLM Sensing

LLMs are being extended beyond text to reason about **human physiology** and the **physical world**.

Sensor data includes heterogeneous modalities: motion, biosignals, and environmental measurements.

Multimodal sensing is useful because sensors provide complementary information.

But real-world sensors can be noisy, missing, or unreliable.

This creates three challenges:

1. **Single-agent reasoning can miss evidence**
2. **Semantic fusion can be biased**
3. **Majority voting can fail under sensor failure**

Ground truth: amusement

User: Given the following sensor features, predict the emotional state from [baseline, stress, amusement].

Accelerometer: {ACC features}

ECG: {ECG features}

EMG: {EMG features}

EDA: {EDA features}

Temperature: {TEMP features}

Respiration: {RESP features} ...

Assistant: The *heart-rate* is elevated (≈ 75 bpm) compared to baseline. *EDA dynamic range* (0.12) less than amusement. *Respiratory stretch* is markedly increased (≈ 18 s), indicating stress-hyperventilation. Overall, the pattern aligns best with **stress**.

(a) Single-agent inference.

Ground truth: amusement

(after observing {ECG features})

ECG agent: Features show a very low high-frequency (HF) HRV component ($3.02e-03$) and a very high low-frequency to HF ratio (15.22), indicating **stress**.

ACC agent: ... is **amusement**.

EMG agent: ... is **amusement**.

EDA agent: ... is **amusement**.

TEMP agent: ... is **amusement**.

...

Judge agent: The most important modality is ECG. With its very low HF HRV component, the answer is **stress**.

(b) Modality agents with a judge agent.

Ground truth: amusement (missing)

ACC agent: ... is **baseline**.

ECG agent: ... is **stress**.

EMG agent: ... is **stress**.

EDA agent: ... is **amusement**.

TEMP agent: ... is **amusement**.

RESP agent: ... is **stress**.

→ Majority Voting: **stress**

Judge agent: The ECG and respiration features appear unreliable. In contrast, the EDA modality provides robust evidence, and the temperature features show consistent indicators. Based on these signals, the predicted emotional state is **amusement**.

(c) Data with 30% missing modalities.

Figure 3: Examples of LLM-based multimodal sensing on WESAD (Schmidt et al., 2018) using gpt-oss-20B.

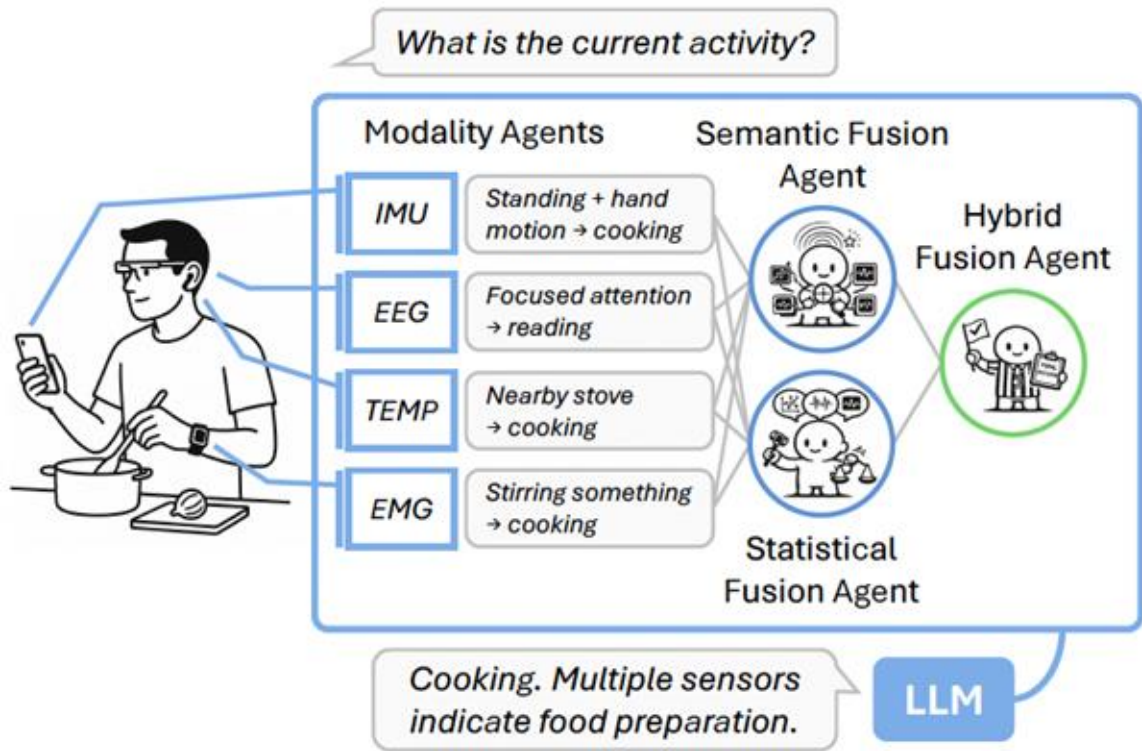
Observation 1: A single agent often produces an *incomplete* cross-modal interpretation in multimodal sensing tasks.

Observation 2: An LLM-based judge is prone to *knowledge bias*, whereby prior domain knowledge can mislead consensus even when robust evidence exists in other modalities.

Observation 3: Majority voting is prone to *sensor failure*, often producing unreliable votes and false consensus in multimodal sensing.

Sensor	Extracted features
ACC GYR MAG ANG	Mean, std, and absolute integral for each axis (x, y, z) and magnitude. Peak frequency per axis.
ECG PPG	Heart rate (HR): Mean and std of HR derived from inter-beat intervals. HR variability (HRV): RMSSD, pNN50, TINN, and std. Frequency domain: Power in ULF (0.01–0.04 Hz), LF (0.04–0.15 Hz), HF (0.15–0.4 Hz), and UHF (0.4–1.0 Hz). Total power, LF/HF ratio, relative powers, and normalized LF/HF components.
EDA	Mean, std, min, max, slope, and dynamic range (5 Hz low-pass filtered). Tonic (SCL): Mean, std, and correlation with time. Phasic (SCR): Mean, std, event count, sum of magnitudes, total duration, and area under curve (AUC).
EMG	Chain 1 (high-pass): Mean, std, dynamic range, absolute integral, median, 10th/90th percentiles. Mean/median/peak frequency and spectral energy across seven bands (0–350 Hz). Chain 2 (50 Hz low-pass): Peak count, mean/std/sum of peak amplitudes, and normalized sum of amplitudes.
RESP	Inhalation/exhalation durations (mean, std, ratio), stretch, inspiration volume, respiration rate, and average cycle duration (0.1–0.35 Hz bandpass).
TEMP	Mean, std, min, max, slope, and dynamic range.
EEG	Frequency bands: Delta, theta, alpha, beta, spindle, K-complex, and sawtooth. Band features: Mean, std, variance, dynamic range, peak count, zero-crossing rate, variance of first-order difference, and absolute power (Welch's method). Ratios: Delta/theta, theta/alpha, alpha/beta, and (delta+theta)/(alpha+beta).
EOG	Time Domain: Mean, std, variance, dynamic range, zero-crossings, and first-order difference variance. Eye movements: Large movement count ($> 120\mu V$ within 1.5s) and difference variance (clean signal). Spectral: Slow (0.5–2 Hz) and rapid (2–5 Hz) power ratios relative to total power (0.5–30 Hz).

Table 4: Summary of features extracted for different sensor types.



Improves accuracy by 7.1% and reduces fusion token cost by 12.7×

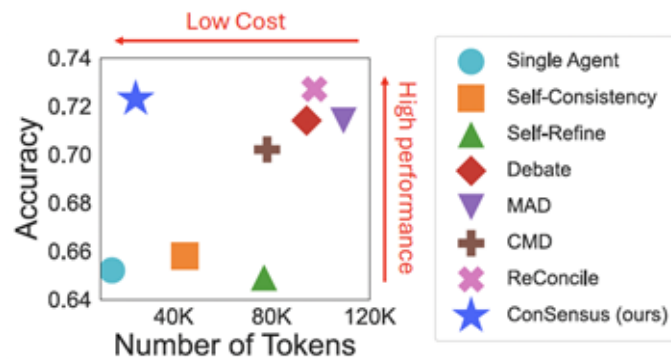
Training-free multi-agent framework for multimodal sensing

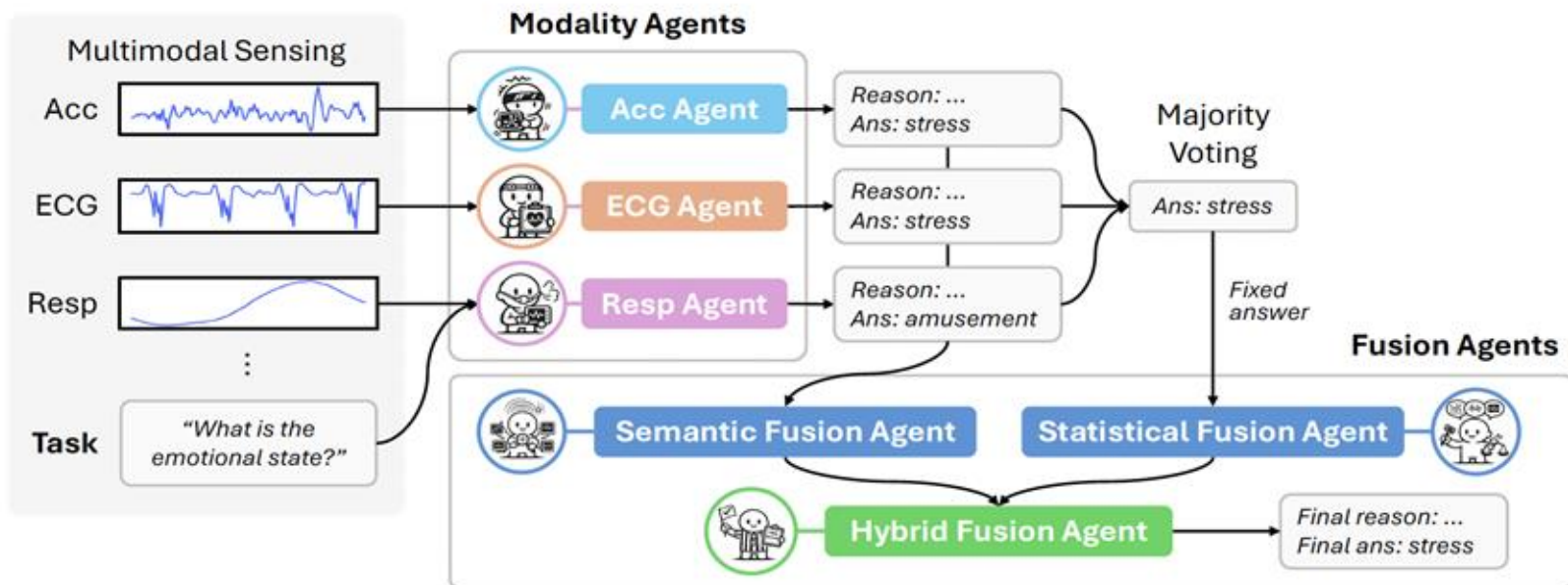
Each sensor is assigned to a **modality-specific LLM agent**

Two fusion paths:

- **Semantic fusion:** reasons over modality explanations
- **Statistical fusion:** anchors to majority vote

Hybrid fusion agent combines both to produce the final decision





Problem formulation

$$M = \{m_1, m_2, \dots, m_N\}$$

and task description T , predict:

$$y \in Y$$

Modality agents

$$A_i^{mod}(m_i, T) \rightarrow (\hat{y}_i, r_i)$$

$$(\hat{y}_{sem}, r_{sem}) = A_{sem}^{fuse}(\{(\hat{y}_i, r_i)\}_{i=1}^N)$$

$$\hat{y}_{vote} = \arg \max_{c \in Y} \sum_{i=1}^N \mathbf{1}[\hat{y}_i = c]$$

$$(\hat{y}_{stat}, r_{stat}) = A_{stat}^{fuse}(\{(\hat{y}_i, r_i)\}_{i=1}^N)$$

$$(\hat{y}, r) = A_{hyb}^{fuse}((\hat{y}_{sem}, r_{sem}), (\hat{y}_{stat}, r_{stat}))$$

Dataset

Dataset	Task	#Classes	Sensor modalities
WESAD	Affective state recognition	3	ACC ECG EMG EDA TEMP RESP (Chest) ACC PPG EDA TEMP (Wrist)
SleepEDF	Sleep stage classification	5	EEG (Fpz-Cz) EEG (Pz-Oz) EOG EMG RESP
ActionSense	Kitchen activity recognition	5	ACC ANG EMG (Left arm), ACC ANG EMG (Right arm)
MMFit	Gym exercise recognition	10	ACC GYR × (Ear Left wrist Right wrist Waist) MAG (Waist) HR × (Left wrist Right wrist)
PAMAP2	Daily activity recognition	12	ACC GYR × (Hand Chest Ankle)

Table 3: Summary of multimodal sensing datasets and their respective sensor modalities.

Prompts

1. Single-Agent Baseline Prompt

This is **not the proposed method**. It is the baseline. Here, one LLM receives **all modalities together**, using one-shot example.

Single-Agent Baseline — System prompt

You are multimodal sensing agent that solves a sensing task. You have the following information about the task:

Task: Classify the user's sleep stage: `<classes>`, based on physiological signals collected from wearable sensors.

Classes: `<description of the classes>`

You will receive sensor features from multiple modalities, and you have the following information about the modality: `{<modality 1>: {"Data collection": <data collection protocol>, "Feature extraction": <feature extraction methods>}, <modality 2>: ..., (repeated for all modalities)}`

Your goal is to analyze the features and provide a reasoned answer using your knowledge.

Single-Agent Baseline — User prompt

You have received sensor features from multiple modalities:

Examples:

Sensor values might not always align with your inherent knowledge due to differences in data collection or processing. So, we included a few labeled examples to help your interpretation:

Example of `<class 1>`:

- `"<feature name>: <value> "`
- `" (repeated for all modality features) "`

Example of `<class 2>`: ...

`(repeated for all examples)`

Current sample features: ...

Please provide your answer for the task among `<classes>` and the reasoning for your answer. Note that the sensor features might be wrong due to the data collection or processing. You can evaluate the quality of the features by checking the examples you have.

Respond in the following strict JSON format: `{"REASON": "<Reasoning for the answer>", "ANSWER": "<Answer among <classes>>"}`

Do not include any additional text outside of the JSON.

LLM response — Single-agent

```
{"REASON": "The ratio of (delta+theta) to (alpha+beta) for both Fpz-Cz (9.34) and Pz-Oz (3.75) falls between the typical values for wakefulness ( $\approx 6.1$  and  $1.5$ ) and stage N2 ( $\approx 10.6$  and  $4.5$ ). The delta power (86.75) is higher than in N1 but much lower than in deep sleep (N3). Alpha and beta powers are lower than in N1/N2 but higher than in wakefulness, indicating a transition toward sleep. The k-complex power and EMG activity are low, suggesting a lighter sleep stage. EOG shows many large eye movements and a high slow-movement ratio, which is more consistent with wake or N1 rather than REM. Overall, the feature pattern best matches stage N2, the intermediate sleep stage.", "ANSWER": "N2"}
```

2. Modality Agent Prompt

The paper gives an example for an **EEG-Pz-Oz agent**

Modality Agents — System prompt

You are EEG-Pz-Oz agent that solves a sensing task.

```
<task description prompt>
```

Modality Agents — User prompt

You have received sensor features from EEG-Pz-Oz modality:

```
<modality-specific example prompt>
```

```
<modality-specific sample prompt>
```

```
<instruction prompt>
```

```
<formatting prompt>
```

LLM response — Respiration modality agent

```
{"REASON": "The current sample shows an inhale duration longer than exhale (ratio 1.39) and an inhale mean of 2.49 s versus exhale mean of 1.79 s, which aligns with the REM example where the inhale is longer than the exhale (ratio 1.33). Other REM characteristics—such as a relatively high respiration rate (~13 breaths/min), moderate duration (~4.6 s), and moderate stretch—are also reflected in the current data. In contrast, the other stages (W, N1, N2, N3) all have inhale/exhale ratios below 1 and different mean durations. Therefore, the most consistent match is REM.", "ANSWER": "REM"}
```

3. Semantic Fusion Agent Prompt

It does **not receive raw sensor features directly**. It receives the outputs from the modality agents

Semantic Fusion Agent — System prompt

You are a fusion agent that solves a multimodal sensing task based on interpretations from multiple sensing agents.

`<task description prompt>`

You will receive reasonings and answers from multiple agents based on their interpretations of different modalities. Your goal is to provide a final reasoned answer for the task.

Semantic Fusion Agent — User prompt

You have received responses from multiple sensing agents:

`{"EEG-Fpz-Cz": <modality agent output>, "EMG-submental": ..., (repeated for all modality agents)}`

Using your own knowledge and expertise, analyze the reasonings and answers and provide a final reasoned answer.

`<formatting prompt>`

Case 1: REM case

LLM response — Semantic fusion agent

```
{"REASON": "The most reliable indicator of sleep stage is the EEG. While the frontal channel (Fpz-Cz) shows a high (delta+theta)/(alpha+beta) ratio typical of N2, the posterior channel (Pz-Oz) displays a ratio (3.75) that lies between the typical REM value ( $\approx 2.6$ ) and the N2 value ( $\approx 4.5$ ), and its other features are more consistent with REM than with deeper stages. The respiratory pattern shows an inhale/exhale ratio  $>1$  and a respiration rate ( $\sim 13$  breaths/min) characteristic of REM. EMG amplitude is low compared with wakefulness, which aligns with REM muscle atonia. Considering the stronger evidence from EEG and respiration, the overall pattern points to REM sleep. Therefore the most plausible classification is REM.", "ANSWER": "REM"}
```

Case 2: N1 case

Semantic Fusion Agent — Failure Case

```
{"REASON": "The most reliable signals for sleep staging are the EEG and respiration measures. Both EEG channels report low-amplitude, low-power activity, which is inconsistent with deep sleep (N3) and also inconsistent with REM. The EOG channel's lack of large eye movements rules out REM and is compatible with both light and intermediate stages, but does not distinguish N1 from N2. The respiration data shows a moderate stretch, a respiration rate of 13.4 bpm, and exhale/inhale durations around 2.2 s—values that fall squarely within the typical range for stage N2. The EMG signal is intermediate between wakefulness and deeper stages, again pointing to a transitional stage rather than deep sleep. Given that the majority of modalities lean toward an intermediate sleep stage and the EOG does not provide evidence for REM, the most consistent classification is stage N2.", "ANSWER": "N2"}
```

4. Statistical Fusion Agent Prompt

Statistical Fusion Agent — System prompt

You are a fusion agent that solves a multimodal sensing task based on interpretations from multiple sensing agents.

`<task description prompt>`

You will receive reasonings and answers from multiple agents based on their interpretations of different modalities. Your goal is to provide a final reasoned answer for the task.

Statistical Fusion Agent — User prompt

You have received responses from multiple sensing agents:

`<responses from modality agents>`

You are on the side that the correct answer is `<majority-voted answer>` which is the majority answer. Based on the given reasonings and answers, provide a final reasoned answer for the task.

For agents who provided different answers from `<majority-voted answer>`, explain why their reasoning is likely affected by noise or unreliable signal interpretation. Stay consistent with the position that the correct answer is likely `<majority-voted answer>`.

`<formatting prompt>`

LLM response — Statistical fusion agent

```
{"REASON": "The two modalities that are most directly linked to sleep stage classification—central EEG (Fpz–Cz) and submental EMG—both point to light sleep (N1). The Fpz–Cz feature vector shows low-amplitude, low-power characteristics that match the prototypical N1 profile far better than any other stage. The submental EMG is in the intermediate activity range between wakefulness and deeper stages, with mean, standard deviation, percentiles, and peak statistics that align with published N1 values. These two signals are the gold standards for distinguishing wake, N1, N2, N3, and REM, so their agreement strongly supports an N1 classification. The other modalities provide weaker or conflicting evidence. Given the central EEG and EMG consensus and the known unreliability of single-modality cues in isolation, the most reliable conclusion is that the subject is in the N1 stage.", "ANSWER": "N1"}
```

5. Hybrid Fusion Agent Prompt

Hybrid Fusion Agent — System prompt

You are a coordinator agent that solves a multimodal sensing task based on interpretations from multiple sensing agents.

<task description prompt>

You will receive reasonings and answers from multiple agents, each interpreting different modalities, as well as two fusion agents that have already aggregated these responses in different ways. Your goal is to coordinate the responses from the fusion agents and provide a final reasoned answer for the task.

Hybrid Fusion Agent — User prompt

You have received the following information.

1. "Original responses from individual modality agents:"

<responses from modality agents>

2. "Responses from fusion agents:"

<response from the semantic fusion agent>

<response from the statistical fusion agent>

Your task is to determine the responses from the fusion agents and provide a final reasoned answer for the task.

<formatting prompt>

LLM response — Hybrid fusion agent

```
{"REASON": "The most reliable evidence comes from the central EEG (Fpz-Cz) and submental EMG, both of which are the gold standards for distinguishing sleep stages. The Fpz-Cz channel shows low-amplitude, low-power activity with variance and delta power values that match the prototypical N1 profile far better than any other stage. The submental EMG is in the intermediate activity range between wakefulness and deeper stages, with mean, standard deviation, and peak statistics that align closely with published N1 values. These two signals agree on light sleep (N1). Other modalities provide conflicting or weaker cues: the Pz-Oz channel suggests REM but its intermediate (delta+theta)/(alpha+beta) ratio and higher alpha/beta ratio point to noise; the EOG horizontal channel indicates minimal rapid eye movements, which is typical of both N1 and N2 and not decisive; the respiration signal is consistent with N2 but respiration alone is a poor discriminator. Given the consensus of the most reliable modalities and the known unreliability of single-modality cues in isolation, the most robust conclusion is that the subject is in the N1 stage.", "ANSWER": "N1"}
```

Evaluation

Backbone LLM: gpt-oss-20B

Additional LLMs: gpt-oss-120B, Llama-3.1-8B-Instruct, Llama-4-Scout-17B-16E

Temperature: 0 for deterministic outputs

Metric: Accuracy; macro-F1 reported in Appendix B

Single-agent baselines:

Single-Agent, Self-Consistency, Self-Refine

Multi-Agent Debate Baselines:

Compare ConSensus with debate-style multi-agent methods

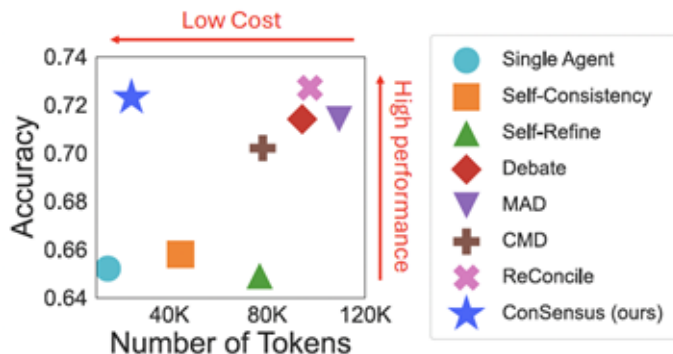
Debate: agents revise after seeing others' outputs

MAD: debate + judge agent

CMD: group-based communication

ReConcile: confidence-weighted voting

Key difference: debate methods use **two rounds**; ConSensus uses **single-round fusion**



Results

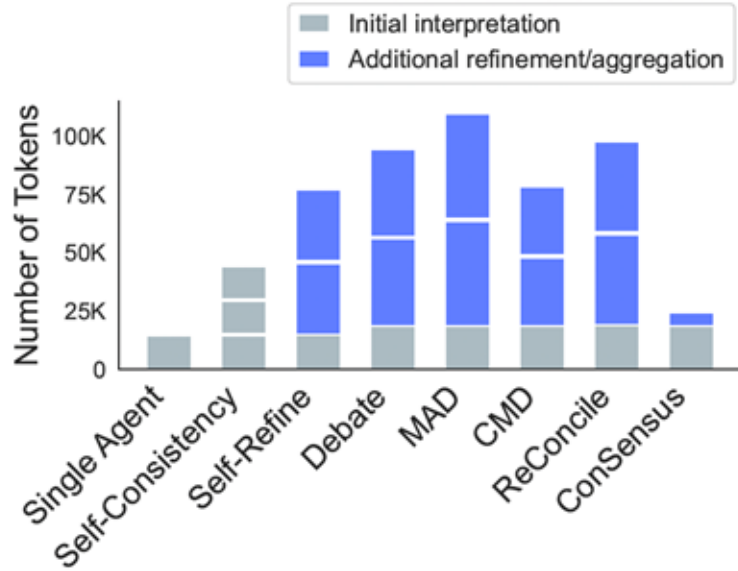
Method	WESAD	SleepEDF	ActionSense	MMFit	PAMAP2	Avg.
<i>Single-agent baselines</i>						
Single-Agent	0.793 ± 0.033	0.519 ± 0.031	0.577 ± 0.032	0.819 ± 0.022	0.551 ± 0.027	0.652 ± 0.027
+ Self-Consistency	0.786 ± 0.035	0.541 ± 0.031	0.555 ± 0.031	0.862 ± 0.019	0.547 ± 0.027	0.658 ± 0.027
+ Self-Refine	0.747 ± 0.035	0.551 ± 0.031	0.566 ± 0.031	0.822 ± 0.022	0.563 ± 0.026	0.650 ± 0.026
<i>Modality agents + multi-agent debate baselines, requiring 76K additional tokens per sample</i>						
+ Debate	0.873 ± 0.027	0.548 ± 0.032	0.609 ± 0.031	0.984 ± 0.007	0.561 ± 0.028	0.715 ± 0.028
+ MAD	0.847 ± 0.029	0.562 ± 0.031	<u>0.613</u> ± 0.031	0.960 ± 0.011	0.589 ± 0.026	0.714 ± 0.026
+ CMD	0.840 ± 0.030	0.578 ± 0.031	0.589 ± 0.034	0.962 ± 0.011	0.539 ± 0.025	0.702 ± 0.025
+ ReConcile	<u>0.880</u> ± 0.027	0.571 ± 0.030	0.640 ± 0.031	0.964 ± 0.011	<u>0.579</u> ± 0.026	0.727 ± 0.026
<i>Modality agents + fusion agents (ours), requiring 6K additional tokens per sample</i>						
+ Semantic Fusion	0.825 ± 0.031	0.580 ± 0.031	0.605 ± 0.033	0.964 ± 0.011	0.559 ± 0.026	0.707 ± 0.026
+ Statistical Fusion	0.927 ± 0.021	<u>0.592</u> ± 0.032	0.597 ± 0.033	0.960 ± 0.011	0.534 ± 0.026	0.722 ± 0.026
ConSensus	<u>0.880</u> ± 0.029	0.600 ± 0.031	0.611 ± 0.031	<u>0.967</u> ± 0.010	0.558 ± 0.026	<u>0.723</u> ± 0.026

Modality agents help because they preserve sensor-specific evidence.

Hybrid fusion helps because semantic and statistical fusion have different strengths.

ConSensus is efficient because it avoids expensive multi-round debate while keeping competitive accuracy.

Token Efficiency



- **Gray = Initial interpretation**
 - system prompts
 - task/data descriptions
 - sensor feature descriptions
 - initial modality-agent predictions
- **Blue = Extra aggregation cost**
 - refinement steps
 - debate rounds
 - fusion prompts
- ConSensus: **6K** aggregation tokens vs. ReConcile: **78.6K**
- ConSensus reduces aggregation cost by **12.7×**

Robustness to Sensor Failure

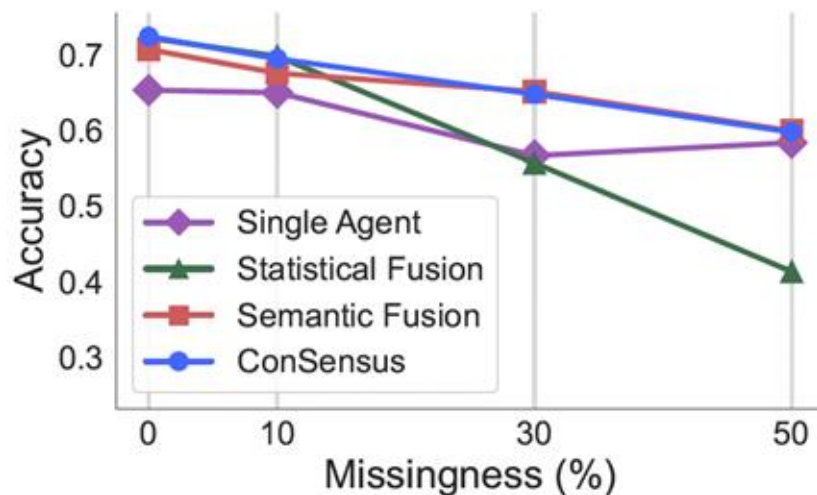
Randomly omit **10%, 30%, and 50%** of sensor modalities

Statistical fusion drops sharply with missing sensors

At **50% missingness**:

- Statistical fusion: **41.4%**
- Semantic fusion: **59.9%**

Hybrid fusion shifts toward semantic reasoning when voting becomes unreliable



Effect of LLM

Tested ConSensus across multiple open-source LLMs




Models: **gpt-oss-120B**, **Llama-3.1-8B-Instruct**,
Llama-4-Scout

ConSensus consistently improves over **Single-Agent**

ReConcile performs best on **gpt-oss-120B**

ConSensus is strongest on smaller/Llama models

Key point: useful for **resource-constrained LLMs**

Method	WE-SAD	Sleep EDF	Action Sense	MM-Fit	PA-MAP2	Avg.
 gpt-oss-120B						
<i>Single agent</i>						
Single-Agent	0.807	0.580	0.564	0.730	0.592	0.654
<i>Modality agents</i>						
+ ReConcile	0.887	0.588	0.604	0.980	0.636	0.739
+ Semantic Fusion	0.793	0.572	0.600	0.967	0.633	0.713
+ Statistical Fusion	0.880	0.580	0.572	0.960	0.572	0.713
ConSensus	0.833	0.572	0.584	0.967	0.611	0.713
 Llama-3.1-8B-Instruct						
<i>Single agent</i>						
Single-Agent	0.493	0.292	0.288	0.167	0.225	0.293
<i>Modality agents</i>						
+ ReConcile	0.673	0.304	0.296	0.390	0.231	0.379
+ Semantic Fusion	0.773	0.368	0.288	0.493	0.344	0.453
+ Statistical Fusion	0.787	0.416	0.324	0.543	0.342	0.482
ConSensus	0.780	0.368	0.296	0.487	0.347	0.456
 Llama-4-Scout-17B-16E						
<i>Single agent</i>						
Single-Agent	0.660	0.416	0.268	0.293	0.386	0.405
<i>Modality agents</i>						
+ ReConcile	0.853	0.556	0.572	0.877	0.497	0.671
+ Semantic Fusion	0.833	0.536	0.588	0.900	0.525	0.676
+ Statistical Fusion	0.860	0.568	0.604	0.910	0.531	0.695
ConSensus	0.860	0.548	0.592	0.903	0.542	0.689

Limitation

Experiments used **dataset subsets** due to LLM inference cost

Evaluation limited to **classification tasks**

Did not combine ConSensus with SC, SR, or ReConcile-style confidence

No fine-tuning; only prompt-based LLM agents

Future work: scalability, RAG, fine-tuning, sensor-specialized LLMs

Thanks

Questions ?