PROGRAMMING FOR DATA SCIENCE · FINAL PROJECT

CMI — Detect Behavior with Sensor Data

Team 4: Siun Park · Chanwoo Park · Kyuhwa Hong

May 28, 2026

Kaggle —Detect Behavior with Sensor Data (https://www.kaggle.com/competitions/cmi-detect-behavior-with-sensor-data/overview)

Detecting BFRBs from Wrist-Worn Sensors

What is a BFRB?

Body-Focused Repetitive Behaviors (e.g., hair pulling, skin picking, nail biting) are clinical markers of Obsessive-Compulsive Disorder (OCD) and anxiety, affecting 2–5% of the population.

Why it matters?

Cause real physical harm (e.g., skin damage, hair loss), stress, and psychosocial challenges.

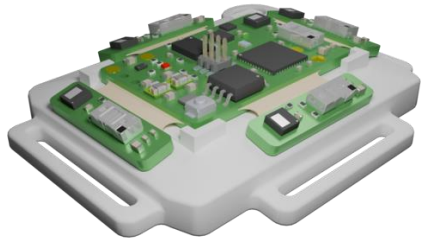
These behaviors represent the key indicators of mental health challenges.



Examples of BFRBs (skin/hair/nail-related).

Why wrist sensors + ML ?

- Current diagnosis relies on self-report, **no objective metric.**
- **Continuous monitoring** without clinical visits.
- Provide objective metrics to quantify treatment progress.



Wrist-worn device Sensor "Helios"

OUR TASK

Classify 18 gestures (8 BFRB-like + 10 non-BFRB) from time-series collected by the Helios wrist device.

1) BFRB-like gestures from non-BFRB-like gestures (binary)
2) The specific type of BFRB-like gestures (multiclass)

Helios Wrist Device: Three Sensor Modalities

Helios is designed to detect these Body-Focused Repetitive Behaviors
Sensors recorded at every timestep (~332 channels)

Child Mind Institute
Healthy Brain Network

IMU

Inertial Measurement Units

7 channels

Measure rotation and motion
Accelerometer (3) + fused quaternion rotation (4). How the wrist moves and where gravity points.

THERMO

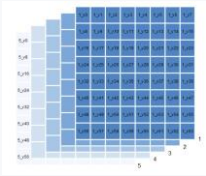
Thermopile

5 channels

Detecting body heat.
Infrared temperature in °C.
Distinguishes proximity to warm skin vs. cold ambient objects.

ToF

Time-of-Flight Sensor



320 channels

Measures the distance of objects.
5 × 8×8 depth grids. Encodes the spatial geometry of where it is in the space.



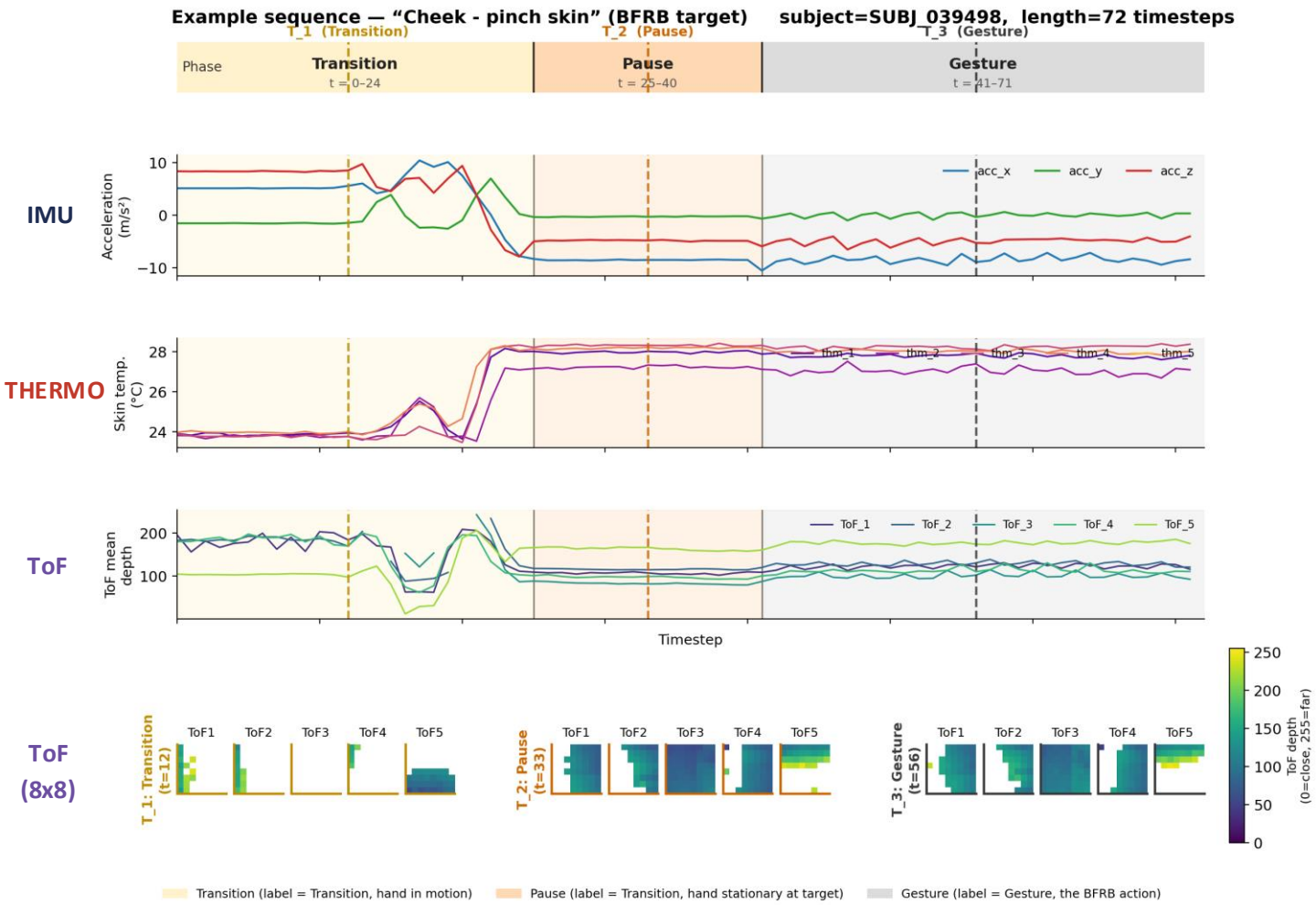
Labels of data & demographics

- **18 gestures:** 8 BFRB-like (hair pulling, skin pinching, scratching, …) + 10 normal (drinking, texting, waving, …) → binary/multiclass
- **4 body positions:** sitting, sitting leaning forward, lying on back, lying on side
- **Demographics:** tabular data for physical characteristics (subjects, adult/child, age, sex, shoulder to wrist cm …)

Test-set difference

50% of test sequences are IMU-only (THERMO and ToF are intentionally masked to test robustness)

Example Sequence — "Cheek pinch skin" (BFRB)



Three phases per sequence

Transition

Transition from “rest” position and moved to appropriate location

Pause

Followed this with a short pause wherein they did nothing

Gesture

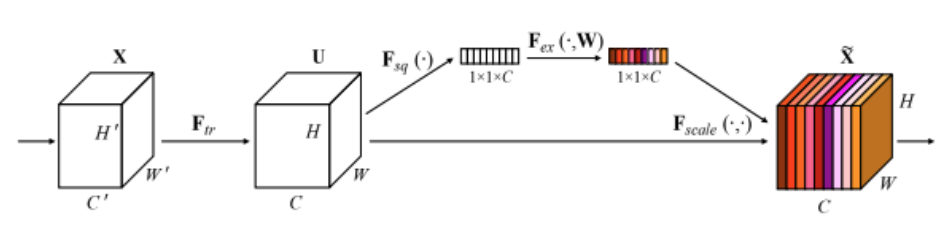
Perform a gesture from either BFRB-like or non-BFRB-like category of movements

Key signal

The model must look across all three phases.
The gesture alone is ambiguous, therefore the **transition pattern** is a key signal for highly subject- and BFRB-specific classification.

Related Works

Squeeze-and-Excitation (SE) Networks



SE-Net introduces a lightweight channel-attention module that recalibrates per-channel feature response based on global context.

SE block has two steps:

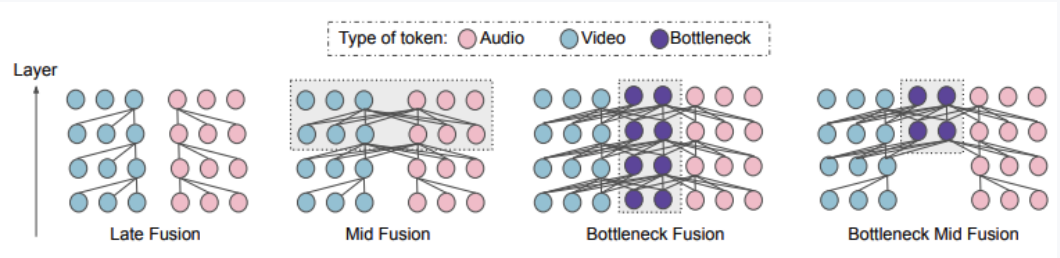
1) Squeeze compressing each channel into a single descriptor via global average pooling

2) Excitation passing this descriptors through a small two-layer bottleneck MLP (FC → ReLU → FC → sigmoid) to produce a per-channel gate.

In our work, every CNN block in IMU, THERMO encoder embeds an SE block, so that network can re-weight sensor channels per sample

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7132-7141).

Cross-modal fusion bottlenecks



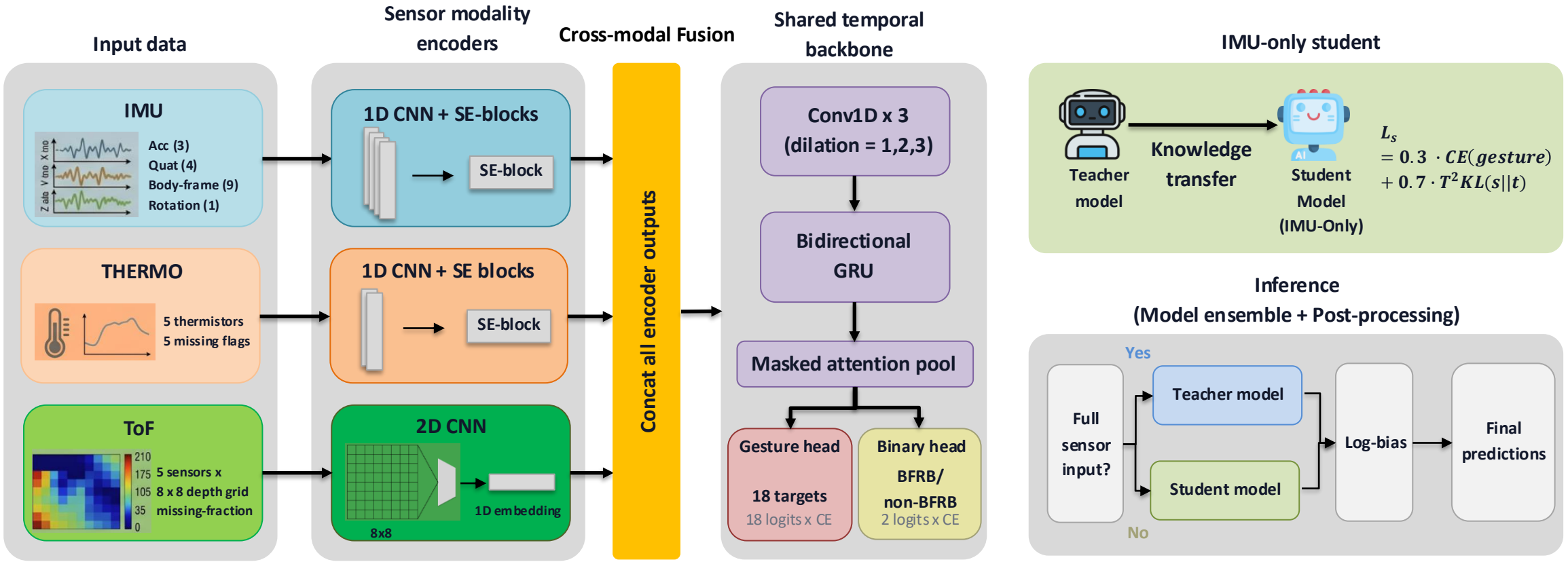
This study is how to fuse two modalities inside a transformer and argues against both late-fusion and full cross-attention.

The model insert **fusion bottleneck tokens** that are only path through which cross-modal information may travel within layer.

In our work, we adopt the same fusion bottleneck architecture to enhance the model performance. Each of IMU, THERMO, ToF encoder representations were share through fusion bottleneck to compose better global representation.

Nagrani, A., Yang, S., Arnab, A., Jansen, A., Schmid, C., & Sun, C. (2021). Attention bottlenecks for multimodal fusion. *Advances in neural information processing systems*, *34*, 14200-14213.

Overall Model Framework



INPUT
3 sensor modalities

ENCODERS
Per-modality Encoders

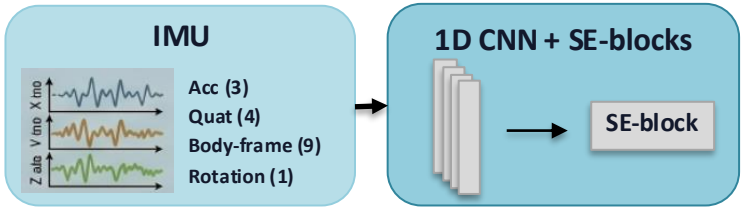
SHARED BACKBONE
1D-CNN + Bi-GRU + Attention
Gesture + Binary head

Knowledge Distillation
Full sensor model → IMU-only model

Per-Modality Encoders

IMU encoder

7 raw → +9 derived → 17 channels



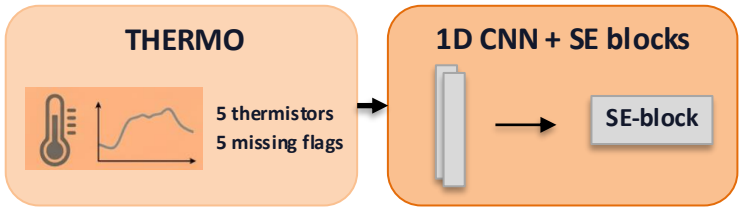
- 1×1 projection
- 1D Conv ×4 (dilated 1,2,3,4)
- SE (squeeze-and-excitation) block in every layer
- Output: (Batch, T, 96)

Why it matters

IMU captures rapid motion directly at high frequency; its dynamics differ from other sensors. Physical biases, gravity, wrist orientation, requires preprocessing.

Thermo encoder

5 raw → +5 missing-mask → 10 channels



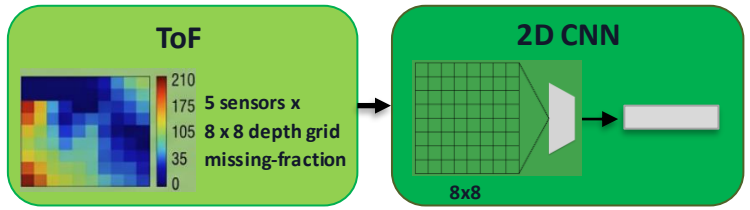
- 1×1 projection
- 1D Conv ×2 + SE
- Lightweight — small signal
- Output: (Batch, T, 64)

Why it matters

Thermo is sparse and noisy; we expose a per-channel missing-value mask so the model can route around dropouts.

ToF encoder

5 sensors × 8×8 depth grid

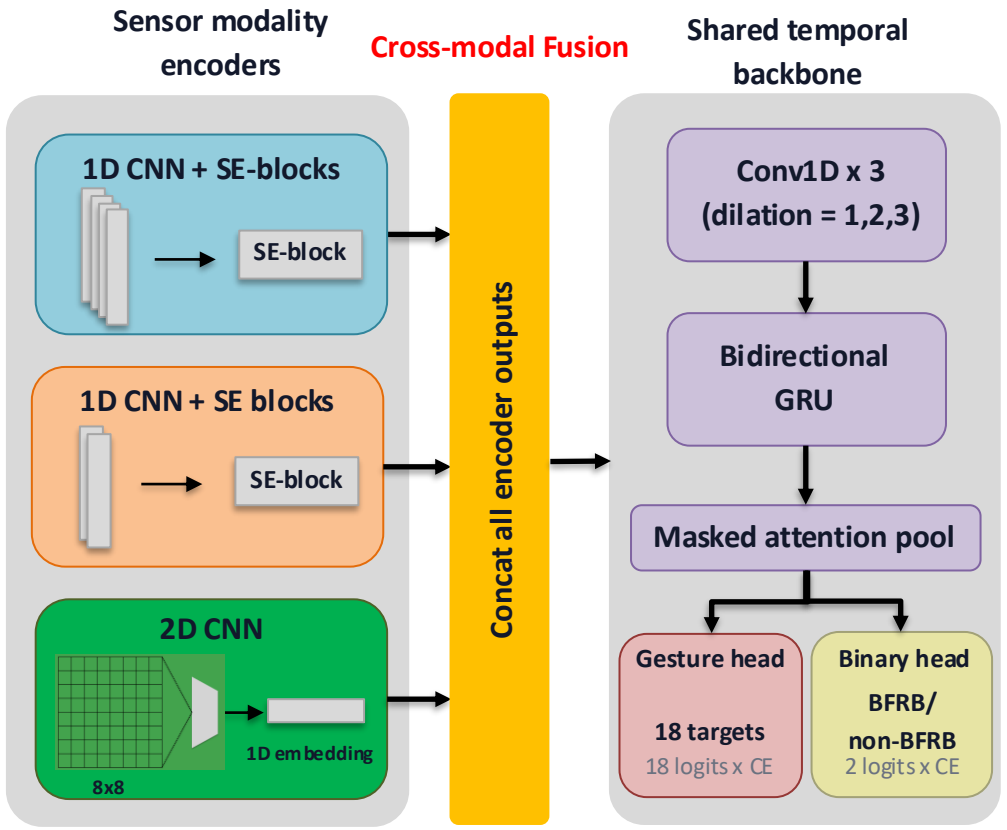


- Per-frame 2D CNN (8→4→2→1)
- 5 input channels (sensors)
- Frame → 96-d embedding
- Output: (Batch, T, 96)

Why it matters

2D CNN preserves spatial structure of each 8×8 grid before flattening it into the temporal stream.

Cross-modal Fusion: Beyond simple concatenation



Limitations of simple concatenation

- Each modalities are **correlated;** e.g., “hand near face” is related with “temperature rising”
- Without cross-attention, the model must rediscover these correlations in the shared backbone, after losing per-modality detail

Cross-modal Fusion Bottleneck

- Before concatenating encoder outputs, all features passes through a **bottleneck token,** information from other modalities.
- **Stage 1:** Bottleneck token reads info from modalities

$$B \leftarrow B + MHA(B, M, M), \qquad M = [m^{IMU}, m^{Thm}, m^{ToF}]$$

- **Stage 2:** Each modality reads bottleneck, absorbs cross-modal info

$$m^{(s)} \leftarrow m^{(s)} + MHA\left(m^{(s)}, B, B\right), \qquad \forall s \in [IMU, Thm, ToF]$$

Shared Temporal Backbone + Two Heads

Concat
IMU + THERMO + ToF
(Batch, T, 257)

→

1D CNN ×3
dilated, residual, SE
→ (Batch, T, 256)

→

Bi-GRU
2 layers, hidden 192
→ (Batch, T, 384)

→

Masked Attn Pool
masked attention
→ 384-d vector

Two classification heads

Gesture head (18 target) *Main
18 logits
1.0 x Cross-Entropy

Binary head (BFRB / non-BFRB)
2 logits
Use auxiliary head, easy binary head pulls the shared backbone to meaningful features in early training

Why CNN → GRU → Attention?

- **CNN** captures local multi-scale temporal patterns.
- **Bi-GRU** integrates context across the full Transition → Pause → Gesture, where it captures long-term patterns.
- **Attention pool** weights on important vectors for classification.

Total training loss

$$\mathcal{L} = 1.0 \times CE(gesture) + 0.3 \times CE(binary)$$

Model ensemble

Single-branch (SB) Stack 348 features flat → 1D CNN backbone.

Multi-branch (MB) Per-modality encoders, concatenated late.

Final teacher = 50 % SB + 50 % MB averaged across **5 folds cross-validation.**

Knowledge Distillation for IMU-only Robustness

Challenge: 50 % of the hidden test set is IMU-only. The full-sensor model has no THERMO/ToF to predict.

STAGE 1

Teacher model (full sensor ensemble)



Teacher model

Train on full IMU + THERMO + ToF

Produce soft probabilities for every training sample.

soft probs

STAGE 2

Student model (IMU-only network)

$$L_{student} = 0.3 \cdot CE(gesture) + 0.7 \cdot T^2 KL(student||teacher)$$

Student sees only IMU input but mimics the teacher's soft outputs.



Student Model (IMU-Only)

Final predictions:

- **If full-sensor data:** argmax of averaged logits from the teacher model.
- **If IMU-only data:** argmax of averaged logits from the student model.

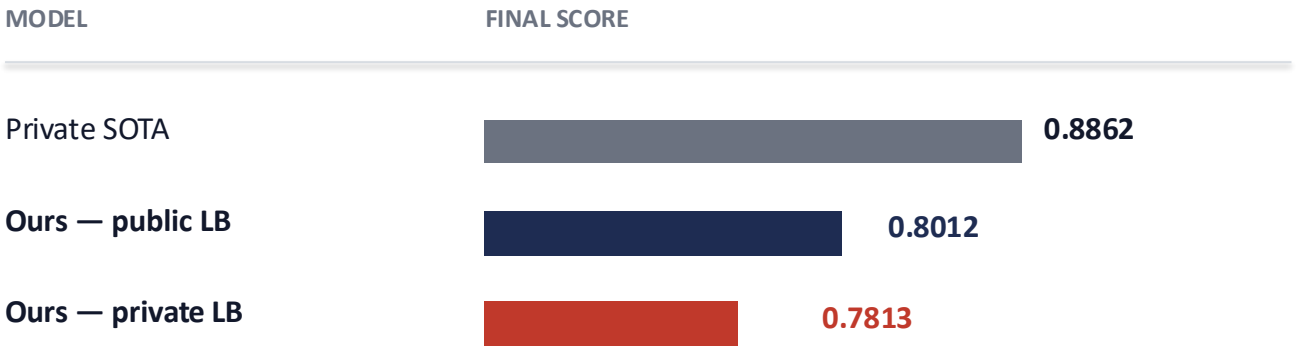
Metrics & Final Scores vs SOTA (Kaggle best score)

Metrics

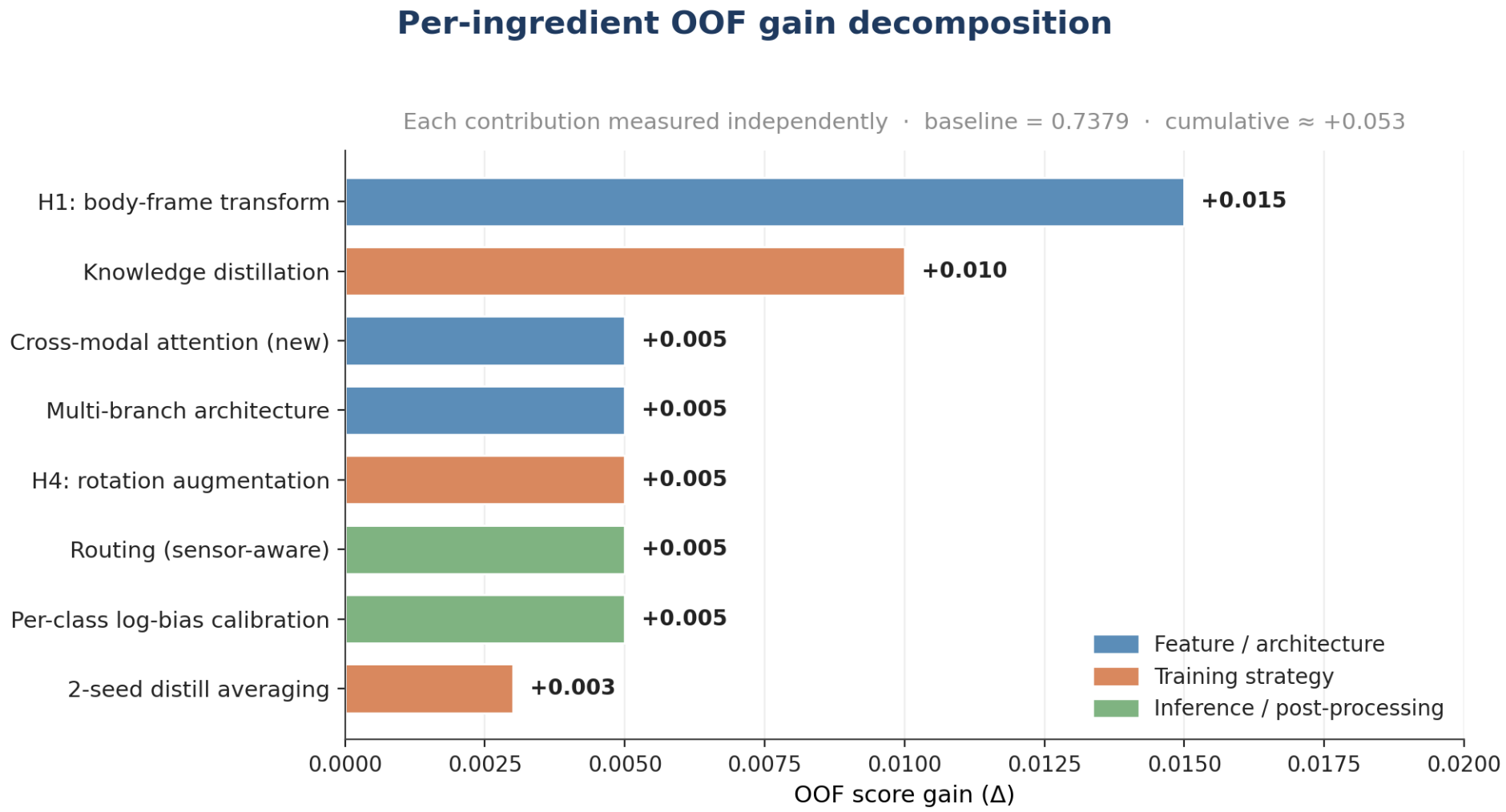
Binary F1 : BFRB vs. non-BFRB (2 classes).

Macro F1 : correct BFRB gestures (8 + 1 collapsed non-BFRB).

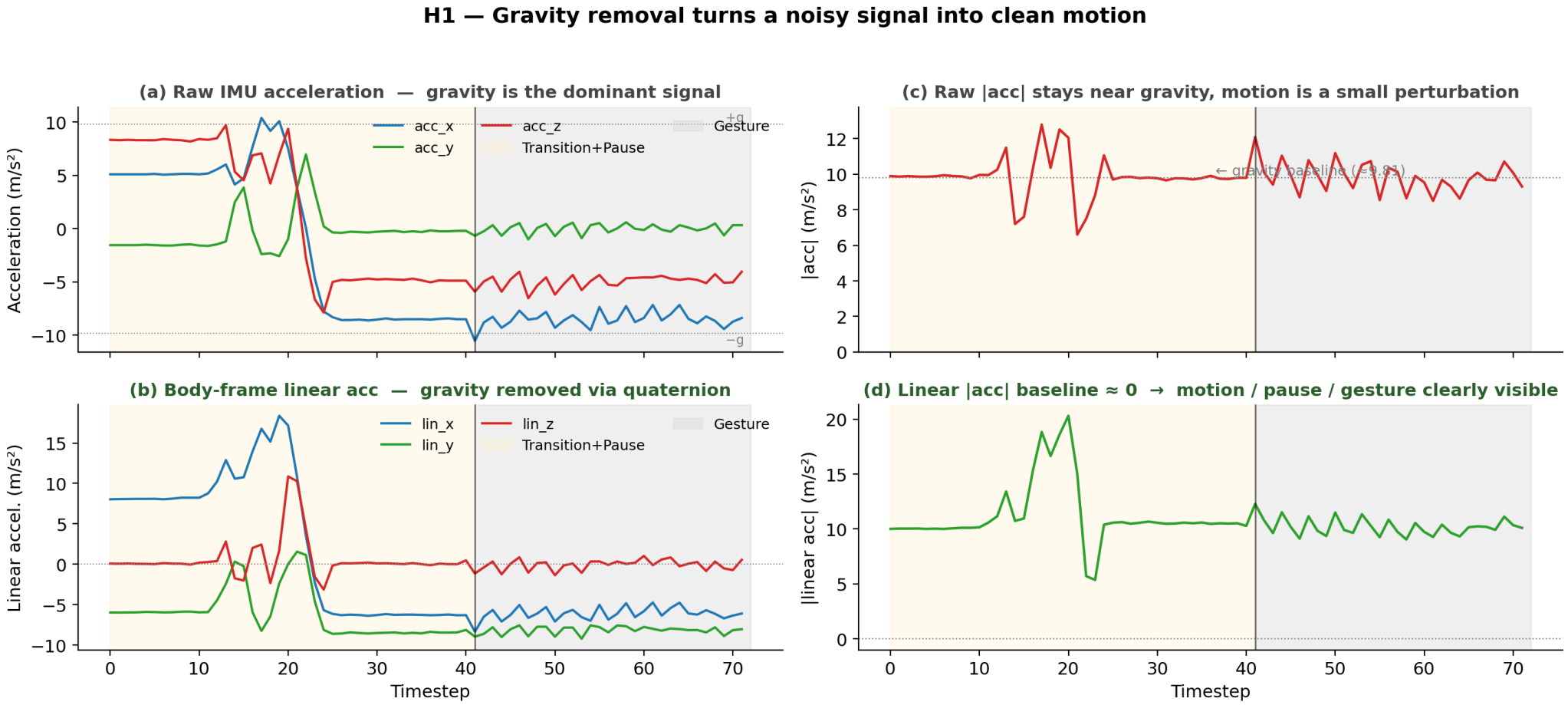
Final score = mean(Binary F1, Macro F1).



Ablation Study

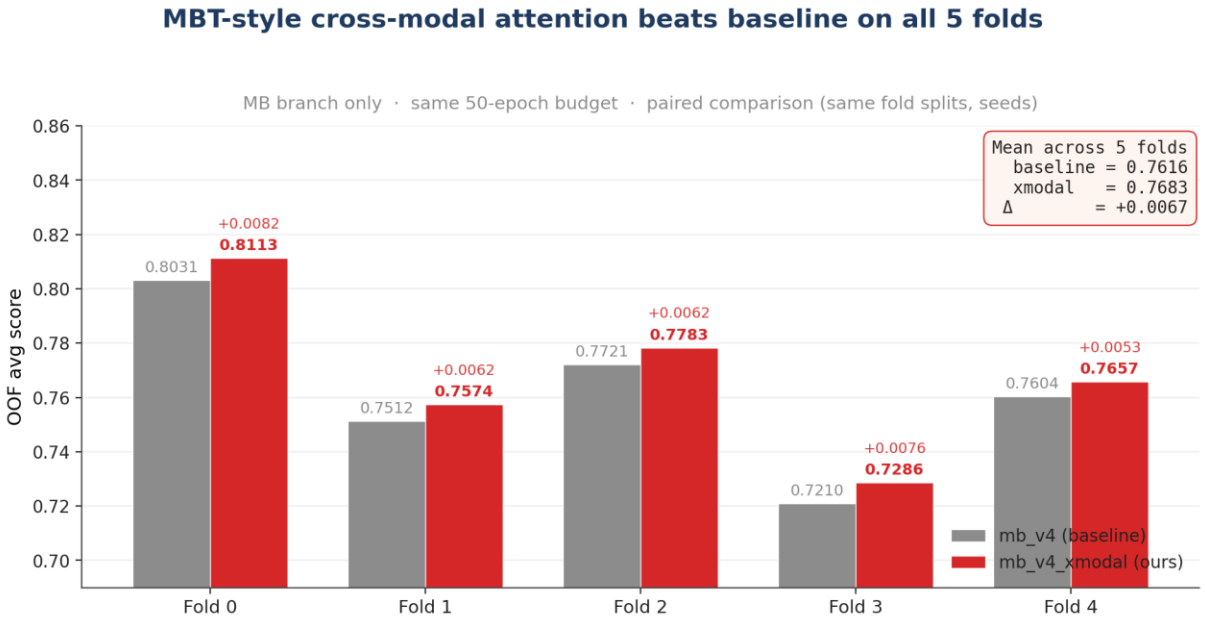


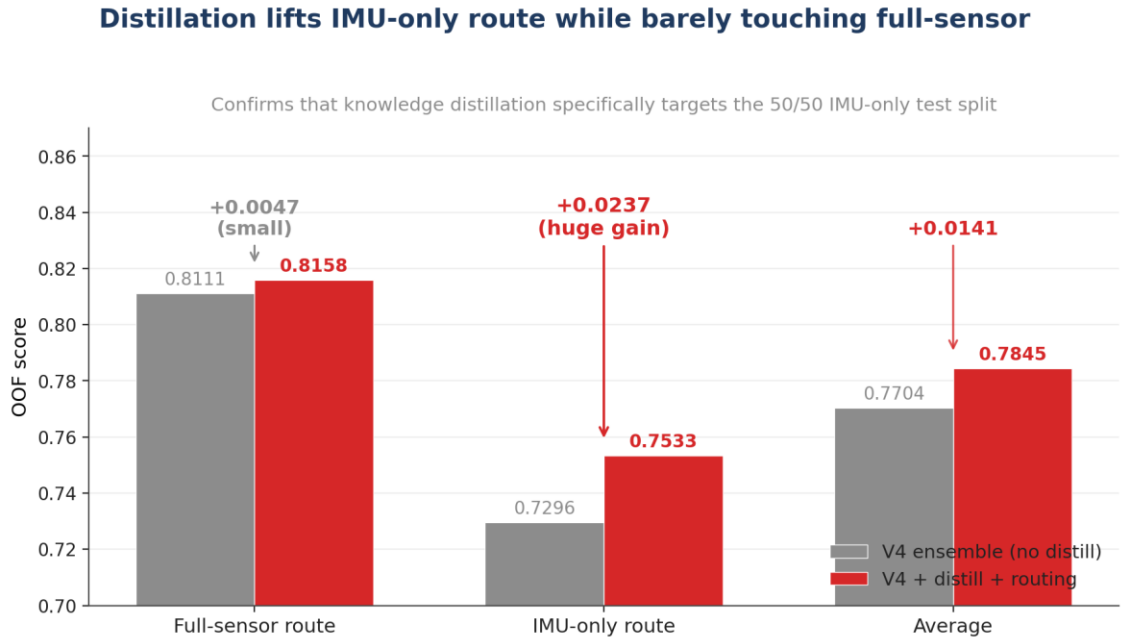
Ablation Study



Same sequence as the overview slide. Yellow = Transition + Pause (labelled Transition), Gray = Gesture. H1 adds (lin_acc, |lin_acc|, jerk, pitch/roll) — physics-informed motion features.

Ablation Study





- Cross-modal attention consistently **improves all 5 folds (+0.0067 avg).**
- Knowledge distillation mainly boosts **the IMU-only route (+0.0237).**
- This supports our sensor-aware strategy: use multimodal fusion when full sensors exist, and use the distilled student when sensors are missing.

CONCLUSION

CMI — Detect Behavior with Sensor Data

Programming for Data Science · 2026

01 Sensor-Aware Model Architecture

We built **a multi-branch model** with separate encoders for IMU, Thermopile, and ToF sensors. Each modality was processed according to its own signal structure, then combined through a shared temporal backbone.

Cross-attention further improves performance by allowing the model to fuse complementary information across sensor modalities.

02 Knowledge Distillation

We built **a teacher-student distillation** framework to handle **IMU-only test samples**.

The full-sensor teacher transfers knowledge from IMU, Thermopile, and ToF signals to an IMU-only student model.

03 Final Results and Future Work

Our model achieved **a competitive score**, although it did not reach the Kaggle best result.

Further improvements may come from stronger cross-modal fusion, better IMU-only augmentation, and **more refined calibration or ensemble strategies**.

Thank you.

Individual Contributions

S — Siun Park

Model development & Presentation

- Overall model development
- Multi-branch architecture & CNN+ SE blocks
- Knowledge distillation idea and development
- Presentation preparation

C — Chanwoo Park

Data preprocessing & Validation

- Visualization of the dataset with different modality sensors
- Data preprocessing for the IMU-only, and full-sensor dataset
- 5-fold cross validation splits and metric implement
- Ablation experiments for each components

K — Kyuhwa Hong

Model development & Submitting

- Overall model development
- Present idea of cross-modal fusion and sensor modality encoders to enhance model performance
- Kaggle submission notebook + packaging
- Presentation preparation