

LEARNING MULTIMODAL SENSOR FUSION WITH JEPA REPRESENTATIONS FOR BREATHING ACTIVITY AND POSE RECOGNITION

Yunxiang Luo^{†*}, Constantino Álvarez Casado^{†‡*}, Xiaoting Wu[†], Nhi Nguyen[†],
Le Nguyen[†], Manuel Lage Cañellas^{†‡}, Miguel Bordallo López[†]

[†]Center for Machine Vision and Signal Analysis, University of Oulu, Finland

[‡]Candour Ltd, Oulu, Finland

ABSTRACT

Predictive video representations do not by themselves specify how a multimodal system should handle missing sensors. We transfer frozen V-JEPA2 features to a set of heterogeneous sensors comprised by RGB video, depth video, and radar time series. A shared fusion model uses modality queries, fixed sensor slots, and explicit availability to classify breathing activity and pose from all seven non-empty sensor subsets. The contribution is a controlled evaluation of subset supervision, pooling capacity, and representation transfer on 39 OMuSense-23 subjects with five subject-group folds and three seeds. With matched downstream parameter counts, the final model reaches recording-level activity macro-F1 of $75.23\% \pm 0.60$ averaged across seven subsets. Supervising all subsets in each batch improves the matched query-pooling control by 1.923 percentage points (conditional 95% subject-bootstrap interval [0.647, 3.209]), while query pooling shows no clear advantage over matched mean pooling under balanced supervision. Radar probes support transfer beyond pooled wavelet pixels, 64-token inference approximately preserves mean activity performance, and the tested Consensus-JEPA loss adds no measured benefit against its exact control.

Index Terms— sensor fusion, missing modalities, contactless sensing, V-JEPA2, activity recognition

1. INTRODUCTION

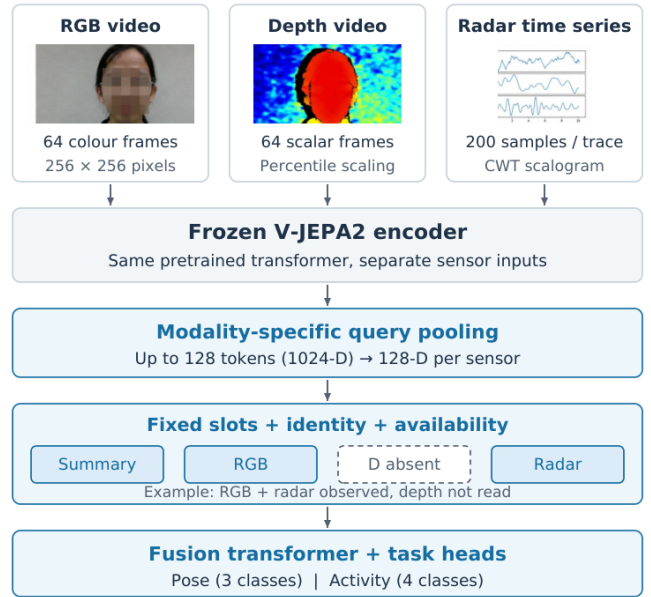
Multimodal contactless human sensing combines cameras and radar to extract physiological and behavioural information for health monitoring and human state assessment [1]. RGB, depth, and radar observe the appearance, geometry, and motion of the same activity and can provide complementary information about breathing and body pose, but the available sensor configuration may change between data collection and deployment.

Power, synchronization, and communication requirements can restrict which sensors are deployed [2]. Here, a missing modality denotes an unavailable sensor stream, rather than degradation within a stream that remains available. Modality dropout and adaptation of pretrained networks address this condition [2, 3], yet multimodal transformers remain sensitive to missing inputs [4]. Performance with complete inputs therefore does not establish performance with missing sensors.

We use V-JEPA2, a video model trained with a joint-embedding predictive architecture (JEPA) to predict target representations from

*These authors contributed equally to this work.

(a) One shared inference model



(b) All-view task supervision

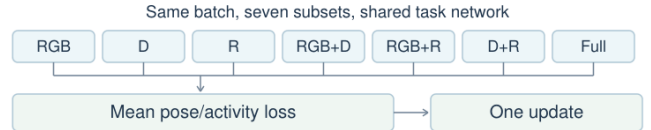


Fig. 1. (a) Shared inference architecture of the final query model and the earlier Subset-query (B3). Grey marks frozen feature extraction and blue trained pooling and fusion. The example declares depth absence, and insets show representative observations. (b) Final all-view task supervision evaluates one batch under all seven non-empty subsets, averages the pose/activity losses, and makes one update, without a latent prediction loss.

available context [5]. The central question is how downstream pooling and supervision affect the use of these frozen representations for breathing activity and pose classification when sensor availability changes. The contributions are as follows:

1. **Availability-aware multimodal fusion.** One classifier with query pooling, fixed sensor slots, and explicit availability covers all seven non-empty RGB, depth, and radar subsets, reach-

ing $75.23 \pm 0.60\%$ activity macro-F1 averaged across them.

2. **Controlled attribution of training gains.** With matched downstream parameters, supervising all subsets in each batch improves the query model by 1.923 percentage points (pp), while query pooling shows no clear advantage over matched mean pooling under balanced supervision.
3. **Evidence on representation use and compression.** Radar probes, token reductions, and an exact zero-weight control test transferred information, compression sensitivity, and a complete-to-subset prediction loss.

Predictive dynamics and planning remain outside the evaluated scope.

2. RELATED WORK

Multimodal learning joins representation, alignment, and fusion across sensor streams [6], and models trained on complete observations can deteriorate when an input stream is unavailable [3, 4]. Proposed responses include modality dropout [2, 7], feature reconstruction and transformer fusion [8, 9], shared and sensor-specific representations [10], logit distillation [11], availability-conditioned paths [12], and parameter-efficient adaptation [3]. They do not address frozen JEPA representations of heterogeneous RGB, depth, and radar inputs evaluated across all seven subsets.

Joint-embedding predictive architectures predict representations rather than input samples, from I-JEPA for images [13] to V-JEPA and V-JEPA2 for video [14, 5]. Predictive alignment has also been studied across modalities and time series [15, 16, 17], while V-JEPA2-AC adds action-conditioned latent dynamics for robotic planning [5]. We instead use V-JEPA2 as a frozen feature extractor and test a complete-to-subset prediction objective against directly supervised fusion.

3. PROPOSED METHODOLOGY

All models use frozen V-JEPA2 features with the same fusion transformer and task heads, as summarised in Table 1, and the final query/all-view model uses the B3 inference architecture, as depicted in Figure 1(a).

3.1. Sensor input construction

Let $\mathcal{M} = \{\text{RGB}, \text{D}, \text{R}\}$ denote the sensors and $\emptyset \neq S \subseteq \mathcal{M}$ the available subset. RGB and depth inputs contain 64 frames at 256×256 pixels. Depth is clipped at the 1st and 99th within-sample percentiles and scaled to the encoder range, and its frozen input convolution averages the pretrained colour kernels. Radar contributes 200 samples of phase displacement, filtered breathing, and filtered cardiac traces, which a complex Morlet wavelet transform maps to a stacked 256×256 scalogram replicated over three channels and two identical frames. Every modality uses the frozen `vjepa2-vit1-fpc64-256` encoder without its predictor. Encoded sequences are adaptively averaged to at most 128 tokens and cached. Radar time lies along a scalogram axis, so duplicated frames do not establish temporal alignment with video tokens.

3.2. Availability-aware fusion

Query pooling uses a learned, modality-specific query to attend to the cached tokens and produce a 128-dimensional vector, whereas mean pooling averages them before projection. The sensor slot is given by:

$$\mathbf{s}_m = \begin{cases} \text{pool}_m(\mathbf{T}_m) + \mathbf{e}_m + \mathbf{v}_1, & m \in S, \\ \mathbf{e}_m + \mathbf{v}_0, & m \notin S, \end{cases} \quad (1)$$

where $\mathbf{T}_m$ holds the tokens, $\mathbf{e}_m$ identifies the modality, and $\mathbf{v}_0, \mathbf{v}_1$ encode absence and presence. A summary token and the three slots enter a pre-norm transformer with two layers, four heads, width 128, and feed-forward width 256. Its summary output $\mathbf{h}_S$ feeds heads for three poses and four activities, whose decisions define 12 joint classes and whose mean cross-entropy is $\mathcal{L}_{\text{task}}$. Absent slots carry no reconstructed measurement.

3.3. Capacity and supervision controls

The factorial comparison crosses query and matched mean pooling with balanced and all-view supervision. Matched mean pooling adds a normalized residual $128 \rightarrow 256 \rightarrow 128$ perceptron per sensor, giving both downstream models 882,695 parameters against 683,399 for Subset-mean (B2), without equalizing computation or inductive bias. Balanced supervision draws one subset per batch, with selection counts across the seven subsets differing by at most one per epoch. All-view supervision evaluates every subset on the same batch and makes one update with:

$$\mathcal{L}_{\text{all}} = \frac{1}{7} \sum_{\emptyset \neq S \subseteq \mathcal{M}} \frac{\mathcal{L}_{\text{activity}}(S) + \mathcal{L}_{\text{pose}}(S)}{2}. \quad (2)$$

The final model uses this objective alone, which gives seven supervised views per example and so increases exposure.

3.4. Complete-to-subset prediction

Consensus-JEPA adds a training-only complete-input branch. For $S \subsetneq \mathcal{M}$, $\mathbf{z}_S = q_\phi(g_\theta(\mathbf{h}_S))$ predicts $\mathbf{z}_M = \text{sg}[g_\theta(\mathbf{h}_M)]$ from the same online network, where sg stops gradients through the complete-input target. The projector and predictor are two-layer perceptrons with hidden width 256 and output width 128. Training minimizes $\mathcal{L}_{\text{task}} + \lambda \mathcal{L}_{\text{cj}}$ with $\mathcal{L}_{\text{cj}} = 1 - \cos(\mathbf{z}_S, \mathbf{z}_M)$. The complete pass runs in evaluation mode without a moving-average teacher. Zero-loss (A0) and Consensus (C1) execute the same graph with $\lambda = 0$ and 0.003, and both auxiliary modules are removed at inference.

4. EXPERIMENTAL EVALUATION

Dataset and acquisition protocol. The complete aligned OMuSense-23 subset contains 39 subjects [1], each recorded standing, sitting, and lying in four nominally 30-second conditions separated by 20-second gaps: normal breathing, reading aloud, guided breathing, and breath holding. The 117 sessions give 468 activity recordings. Using TIMER-defined activity intervals, we extract 10-second windows at a nominal 1-second stride, appending a complete end-aligned window when the regular grid does not reach the interval end. This yields 10,530 aligned samples. Lying-pose RGB and depth are rotated by 180° using the pose field, so deployment would need label-independent orientation metadata, and the camera position and background of this pose may affect pose and joint scores. Activity is therefore the primary endpoint, although acquisition effects cannot be excluded.

Subject split, aggregation, and metrics. Five outer subject-group folds each reserve five training subjects for validation, with subjects disjoint across training, validation, and testing. Recording-level activity macro-F1 selects checkpoints, on the training subset for Full-only and specialists and on the seven-subset mean for universal models. Window logits are averaged per recording before

Table 1. Earlier controlled reference models and their out-of-fold (OOF) recording-level macro-F1 (%). All use frozen V-JEPA2 tokens, a fusion transformer, and supervised pose/activity heads, and universal models sample one subset per batch. Activity Avg-7 is mean $\pm$ sample SD, pose and joint are Avg-7 means, and Worst-7 averages the within-seed minimum.

Model (ID)	Nets	Training subsets	Pooling	Auxiliary target (weight)	Act. Avg-7	Pose	Joint	Worst-7
Full-only (B0)	1	Full input only	Mean	None	57.92 $\pm$ 2.67	78.99	48.82	22.35
Subset-mean (B2)	1	Balanced seven	Mean	None	71.88 $\pm$ 0.74	89.80	64.15	54.04
Subset-query (B3)	1	Balanced seven	Query	None	74.05 $\pm$ 0.36	90.09	66.64	55.22
Specialists (B1)	7	One fixed subset	Query	None	74.74 $\pm$ 0.55	90.29	67.63	60.00
Zero-loss (A0)	1	Balanced seven	Query	Full latent branch (0)	73.24 $\pm$ 1.56	90.07	65.92	54.26
Consensus (C1)	1	Balanced seven	Query	Full latent, predictor (0.003)	73.23 $\pm$ 1.16	90.06	65.91	54.73
Direct (A1)	1	Balanced seven	Query	Full latent, no predictor (0.003)	73.69 $\pm$ 0.82	89.82	66.08	54.93
Logit-KD (KD)	1	Balanced seven	Query	Full logits (0.5, temperature 2)	73.81 $\pm$ 1.32	89.91	66.55	54.49

$\hat{y}_r = \arg \max_c \bar{\ell}_{r,c}$, and the five test folds are pooled per seed, giving 468 predictions per subset. For task t with C_t classes,

$$F_{1,t}(S) = \frac{1}{C_t} \sum_{c=1}^{C_t} \frac{2 \text{TP}_c}{2 \text{TP}_c + \text{FP}_c + \text{FN}_c}, \quad (3)$$

$$\text{Avg7}_t = \frac{1}{7} \sum_{0 \neq S \subseteq \mathcal{M}} F_{1,t}(S).$$

Activity Avg-7 is primary, with pose and joint Avg-7 and activity Worst-7, the within-seed minimum subset score. Paired 95% intervals use 5,000 subject-cluster bootstrap draws shared between compared methods. They are conditional on the fitted models and folds, excluding retraining and new-cohort uncertainty, and exploratory without multiplicity correction.

Comparators and optimization. In Table 1, C1 must be compared with A0 to isolate the auxiliary weight, since B3 shares their inference architecture but not their training graph. Logit-KD targets complete-input logits of the same online network. These 14 configurations give 210 training tasks over five folds and three seeds, with AdamW at learning rate 3×10^{-4} , weight decay 10^{-4} , batch size 128, gradient clipping at 1, at most 120 epochs, and patience 15. The factorial experiment keeps these settings, seeds 20260816–20260818, and validation Avg-7 selection, covering 60 training cells, 50 of them new.

Representation probes and token reduction. Linear probes fit separate logistic-regression heads (L-BFGS, $C = 1$, at most 2,000 iterations) to seven fixed representations, giving 70 heads over five folds without tuning or repeated seeds, with standardization fitted on training windows only. Radar uses a 16×16 pooled scalogram, 135 spectral/statistical features (40 Hann-periodogram log-power bins over 0.1–4 Hz and five statistics per trace), mean V-JEPA2 features of 1,024 dimensions, and their 256-dimensional whitened PCA fitted on training data (seed 20260908). Mean RGB, mean depth, and full concatenation complete the set. Token reduction evaluates two 64-token variants of all 15 final checkpoints without retraining. The RGB/depth cache holds 32 temporal groups and four spatial stripes, so spatial aggregation gives 32×2 groups and temporal aggregation 16×4 , while radar uses the same reduction in both.

5. RESULTS AND DISCUSSION

Subset-aware training addresses the largest deficit. Full-only reaches 80.51% activity macro-F1 with all sensors but only 57.92% Avg-7, whereas Subset-mean reaches 71.88%, as shown in Table 1. This 13.96-point difference includes subset-aware checkpoint selection. It gains 31.69, 27.47, and 29.87 points for depth, radar, and depth+radar, but changes by -1.03 with complete input, so it mainly

corrects the mismatch between training and available sensors. Frozen predictive features support the task, but the downstream decision must still be trained for the observed subsets.

Matched capacity qualifies the pooling claim. The earlier 2.17 pp gain of Subset-query over Subset-mean also changes parameter count. With matched capacity, as shown in Table 2 and depicted in Figure 2(b), mean minus query is +0.206 pp [-1.083, 1.457] under balanced supervision, while query exceeds mean by 1.433 pp [0.024, 2.879] under all-view supervision, close to the interval boundary. The interaction $(Q_{\text{all}} - Q_{\text{bal}}) - (M_{\text{all}} - M_{\text{bal}})$ of 1.639 pp [-0.061, 3.350] does not establish a special synergy.

Table 2. Completed capacity-matched comparison. All four models have 882,695 downstream parameters. Scores are recording-level macro-F1 (%), with three-seed sample SD for activity Avg-7 and means for the remaining columns.

Pooling / supervision	Act. Avg-7	Pose	Joint	Worst-7
Query / balanced	73.31 $\pm$ 0.21	89.41	65.49	54.43
Mean / balanced	73.51 $\pm$ 0.99	89.61	65.58	56.09
Mean / all-view	73.80 $\pm$ 1.00	90.47	66.72	56.70
Query / all-view	75.23 $\pm$ 0.60	91.25	68.85	59.59

All-view supervision improves the matched query model. Query/all-view reaches 75.23% $\pm$ 0.60 activity Avg-7 against 73.31 $\pm$ 0.21 for its local query/balanced control, a paired gain of 1.923 pp [0.647, 3.209] that is positive for each seed. Pose and joint improve by 1.845 pp [1.069, 2.642] and 3.368 pp [1.919, 4.918]. The earlier B3 score of 74.05 records a different execution, not this same-environment reference. As depicted in Figure 2(a), the gain concentrates in incomplete inputs, with depth rising from 54.43 to 59.59% and radar from 67.39 to 71.45%. Depth remains the weakest subset, and the final model does not dominate every sensor condition.

One network compared with seven specialists. The final model is 0.490 pp [-0.607, 1.626] above the earlier specialist bank, which establishes neither superiority nor equivalence, and the bank was not rerun in the new environment. The specialists remain higher for depth and radar alone. Its 882,695 downstream parameters cover every availability state, against 4,381,745 for a bank pruned to its required branches, both excluding the shared frozen encoder of about 304 million parameters. This concerns storage and maintenance, not end-to-end speed or energy.

Radar probes test representation transfer. As shown in Table 3, with the same classifier type, mean V-JEPA2 features exceed pooled CWT pixels by 6.319 pp [2.751, 9.996] in activity F1, and the PCA control retains 3.951 pp [0.287, 7.729] at the same input dimension.

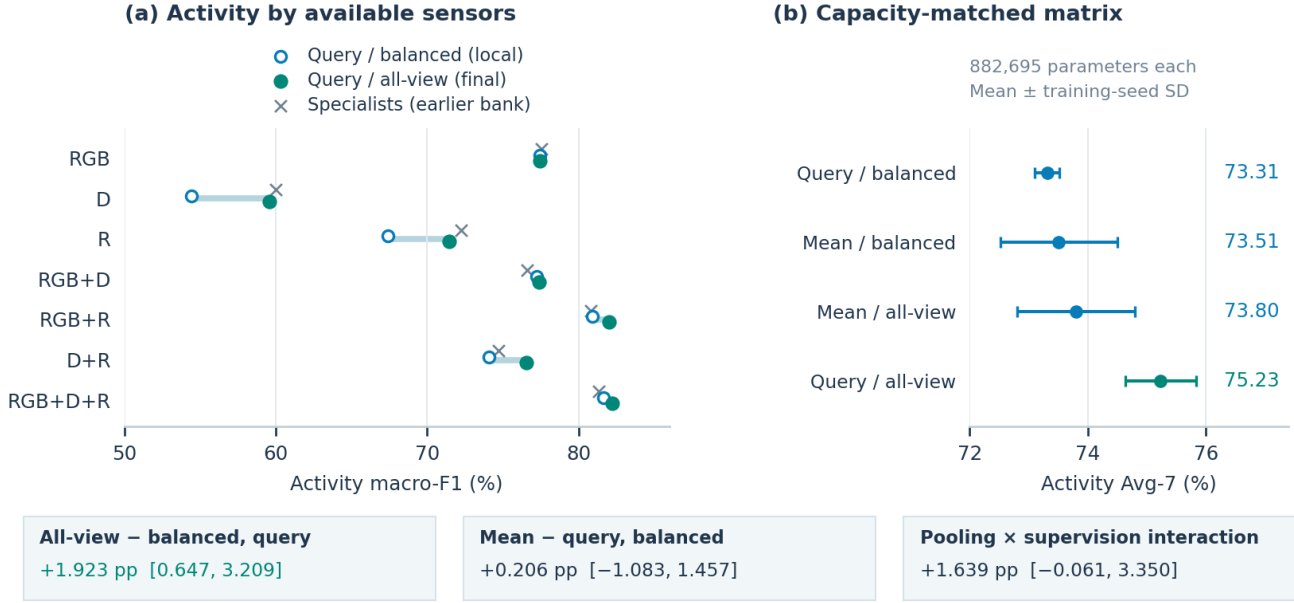


Fig. 2. (a) Activity macro-F1 per sensor subset (three-seed means) for the local query/balanced control, the final query/all-view model, and the earlier specialists, with D depth and R radar. (b) Activity Avg-7 of the four capacity-matched models with training-seed SD, not subject intervals. Lower panels give paired contrasts in pp with 95% subject-bootstrap intervals.

This supports transferred information beyond the pooled-image descriptor, although PCA fitting and pixel pooling prevent complete isolation of the encoder. The 2.143 pp [−1.217, 5.500] difference over spectral/statistical features leaves activity superiority uncertain. The radar probe also exceeds the radar-only shared model (76.31 against 71.45%), so feature information is not equally used by every downstream configuration, although the gap does not identify interference.

Table 3. Fixed radar linear probes, with classifier input dimension and radar-only recording-level macro-F1 (%), not seven-subset averages. Each row pools five OOF folds without repeated seeds.

Radar representation	Dim.	Activity	Pose	Joint
Pooled CWT pixels	256	69.99	71.30	45.06
Spectral/statistical	135	74.17	76.00	52.91
V-JEPA2 mean + PCA	256	73.95	82.22	60.49
V-JEPA2 mean	1,024	76.31	84.58	64.17

Post-training compression approximately preserves mean scores. At 64 tokens the final checkpoints give activity Avg-7 of $75.26\% \pm 0.34$ with spatial and $75.10\% \pm 0.20$ with temporal aggregation. Spatial 64 minus the 128-token reference is +0.025 pp [−0.431, 0.484], and temporal minus spatial −0.152 pp [−0.589, 0.272], proving neither equivalence nor a preferred geometry, so 128 tokens remain the reference. The FP32 payload halves from 1.50 to 0.75 MiB per window, although the 128-token cache is still read. An earlier B3 reduction to 32 tokens lowered Avg-7 from 74.05 to 59.24%. As reduction follows global encoder attention, neither experiment isolates local spatial or temporal information.

The matched predictive loss does not improve Avg-7. Consensus minus Zero-loss is −0.005 pp [−1.350, 1.278], a contrast that changes only the auxiliary weight. Direct alignment and Logit-KD do not exceed Subset-query, but change parameters or targets and do not show that such objectives generally fail. Over 15 Consen-

sus checkpoints and six incomplete subsets, the median weighted auxiliary/task gradient-norm ratio is 0.065%, and gradient cosine is negative in ten of 90 conditions. This indicates a weak auxiliary contribution at this weight, without establishing its cause or representation collapse, so the null result is specific to the tested target and protocol.

What this contributes to JEPA-based world models. The study separates transfer from a JEPA backbone, training an observation classifier for missing sensors, and learning predictive dynamics. This encoder could precede a separate transition model, but classification does not establish a physically sufficient or sensor-independent state. The transferable lesson is to test observation encoders across deployment subsets with controlled capacity, exposure, and auxiliary-loss baselines.

Incremental evaluation and published sensing baselines. Table 4 places our stages beside published models that process engineered camera/radar or derived cardiorespiratory signals [1, 9, 18]. The initial distillation used an RGB–radar teacher and rolling leave-one-subject-out evaluation with three validation subjects and window metrics, whereas the controlled stage adds subject-group folds, recording aggregation, and corrected orientation and radar windowing. Because protocols differ, 75.23% Avg-7 cannot be interpreted as a controlled improvement over 70.43%. Likewise, the full-input 82.23% macro-F1 and published 87.2% hybrid accuracy cannot establish a performance ranking because both the metrics and evaluation protocols differ.

Scope and remaining evidence. We test declared absence, not failure detection or corrupted inputs, and all seven subsets appear in training. Unseen sensors, incomplete records, free-living activity, new sites, and other backbones are not evaluated. Training compute and specialist exposure are unmatched, pose may use acquisition context, and repeated development on this cohort limits the intervals. An exposure control with one update from seven indepen-

Table 4. Published OMuSense-23 baselines and stages of this work, as activity accuracy/macro-F1 (%). Published protocols are subject fivefold, except the 30/10/10 split of physiological SSL. Stages are not matched reruns, earlier F1 aggregation differs from pooled OOF scoring, and a dash marks an untabulated result.

Study/stage (representation)	Full input	Seven subsets
<i>Published, N = 50, window-level protocols</i>		
OMuSense-23 [1] (477 handcrafted)	81.0/81.0	–
Physiological SSL [9] (four 1-D)	74.0/–	–
Hybrid fusion [18] (handcrafted+deep)	87.2/–	–
<i>Initial stage, N = 39, rolling LOSO/window</i>		
V-JEPA2 (frozen RGB, depth, radar)	77.99/76.27	–
Distillation (experts + teacher)	81.33/79.80	72.39/70.43
<i>Controlled stage, N = 39, fivefold OOF/recording</i>		
Full-only (B0), corrected cache	–/80.51	–/57.92
Subset-query (B3), same cache	–/81.18	–/74.05
Query/all-view, final	–/82.23	–/75.23

dent batches differs from all-view by +0.140 and -0.488 pp in two validation cells, so a unique benefit from reusing examples is not established. Saved predictions and paired intervals are available. Full reproduction also requires model checkpoints, cached features, and subject splits. Other small fusion and auxiliary-loss screens remain exploratory.

6. CONCLUSION

We present a shared multimodal fusion model that recognizes breathing activity and pose from frozen V-JEPA2 representations of RGB, depth, and radar under all seven non-empty sensor subsets, without reconstructing missing measurements. Supervising all subsets in each batch yields an activity Avg-7 of $75.23\% \pm 0.60$ and improves the matched control by 1.923 pp, primarily for incomplete inputs. Under matched capacity, query pooling shows no clear advantage over matched mean pooling under balanced supervision, and the Consensus-JEPA loss yields no measurable gain. The evidence therefore supports direct supervision over the known availability states with explicit checks of representation, capacity, uncertainty, and compression, whereas predictive state modelling for JEPA-based world models remains a separate question.

ACKNOWLEDGMENT

The research was supported by the Research Council of Finland Smart Video Sensorization for Secure Healthcare Monitoring (SViSenS) (Grant Number: 370277), the Interreg Aurora ResilientEdge project (Grant Number: 20373282) and Profi7 Hybrid intelligence program (352788). The authors wish to acknowledge the use of AI-assisted tools for language editing, with a focus on grammar checking and improving readability, and coding support. In addition, the Elicit platform was used during the literature search as an AI-assisted discovery aid for candidate records. All inclusion decisions, exclusions, data extraction, synthesis, interpretation, and final writing decisions were performed by the authors.

COMPLIANCE WITH ETHICAL STANDARDS

This work analyses existing recordings, including RGB video, from the OMuSense-23 acquisition [1]. That study was conducted in accordance with the Declaration of Helsinki and institutional regula-

tions, and all participants provided written informed consent. No new data were collected and no participant intervention was performed. The face shown in Figure 1 is pixelated to prevent identification.

7. REFERENCES

- [1] Manuel Lage Cañellas, Le Nguyen, Anirban Mukherjee, Constantino Álvarez Casado, Xiaoting Wu, Nhi Nguyen, Praneeth Susarla, Sasan Sharifipour, Dinesh B. Jayagopi, and Miguel Bordallo López, “OMuSense-23: A multimodal dataset for contactless breathing pattern recognition and biometric analysis,” *IEEE Internet of Things Journal*, pp. 1–14, 2026.
- [2] Sébastien de Blois, Mathieu Garon, Christian Gagné, and Jean-François Lalonde, “Input dropout for spatially aligned modalities,” in *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2020, pp. 733–737.
- [3] Md Kaykobad Reza, Ashley Prater-Bennette, and M. Salman Asif, “Robust multimodal learning with missing modalities via parameter-efficient adaptation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, no. 2, pp. 742–754, 2025.
- [4] Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng, “Are multimodal transformers robust to missing modality?,” in *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*. IEEE, 2022, pp. 18156–18165.
- [5] Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, et al., “V-JEPA 2: Self-supervised video models enable understanding, prediction and planning,” *arXiv preprint arXiv:2506.09985*, 2025.
- [6] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency, “Multimodal machine learning: A survey and taxonomy,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2019.
- [7] Natalia Neverova, Christian Wolf, Graham W. Taylor, and Florian Nebout, “ModDrop: Adaptive multi-modal gesture recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 8, pp. 1692–1706, 2016.
- [8] Sangmin Woo, Sumin Lee, Yeonju Park, Muhammad Adi Nugroho, and Changick Kim, “Towards good practices for missing modality robust action recognition,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023, vol. 37, pp. 2776–2784.
- [9] Manuel Lage Cañellas, Constantino Álvarez Casado, Le Nguyen, and Miguel Bordallo López, “A self-supervised multimodal framework for 1D physiological data fusion in remote health monitoring,” *Information Fusion*, vol. 124, pp. 103397, 2025.
- [10] Hu Wang, Yuanhong Chen, Congbo Ma, Jodie Avery, Louise Hull, and Gustavo Carneiro, “Multi-modal learning with missing modality via shared-specific feature modelling,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 15878–15887.
- [11] Qi Wang, Liang Zhan, Paul Thompson, and Jiayu Zhou, “Multimodal learning with incomplete modalities by knowledge distillation,” in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2020.

- [12] Sijie Li, Chen Chen, and Jungong Han, “SimMLM: A simple framework for multi-modal learning with missing modality,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 24068–24077.
- [13] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas, “Self-supervised learning from images with a joint-embedding predictive architecture,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 15619–15629.
- [14] Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas, “Revisiting feature prediction for learning visual representations from video,” *arXiv preprint arXiv:2404.08471*, 2024.
- [15] Hongyang Lei, Xiaolong Cheng, Qi Qin, Dan Wang, Huazhen Huang, Qingqing Gu, Yetao Wu, and Luo Ji, “M3-JEPA: Multimodal alignment via multi-gate MoE based on the joint-embedding predictive architecture,” in *Proceedings of the 42nd International Conference on Machine Learning*, 2025, vol. 267 of *Proceedings of Machine Learning Research*, pp. 33902–33917.
- [16] Abanoub M. Girgis, Alvaro Valcarce, and Mehdi Ben-nis, “Time-series JEPA for predictive remote control under capacity-limited networks,” *arXiv preprint arXiv:2406.04853*, 2024.
- [17] Utsav Dutta, Gerardo Pastrana, Sina Khoshfetrat Pakazad, and Henrik Ohlsson, “Giving sensors a voice: Multimodal JEPA for semantic time-series embeddings,” *arXiv preprint arXiv:2605.31580*, 2026.
- [18] Manuel Lage Cañellas, Constantino Álvarez Casado, Miika Malin, Nicoletta Prencipe, Sasan Sharifipour, and Miguel Boddallo López, “Fusion of handcrafted and deep-learned cardiorespiratory features for breathing pattern classification,” in *Proceedings of the 33rd European Signal Processing Conference (EUSIPCO)*, Palermo, Italy, 2025, pp. 1627–1631, IEEE Signal Processing Society.