HS-JEPA: A JEPA-Inspired Predictive Multi-View Framework for Human Life Logs

Byungsoo Kang†
Dept. of Computer Engineering
Myongji University
Yongin, Republic of Korea
kbsoo0620@gmail.com

Gwang-Min Park†
Dept. of Computer Engineering
Myongji University
Yongin, Republic of Korea
gwangminions@gmail.com

DongKyu Hyeon†
Dept. of Computer Engineering
Myongji University
Yongin, Republic of Korea
hyeondongkyu@gmail.com

† Equal contribution.

***Abstract*—A day is never recorded whole. Smartphones and wearables leave only partial observations of human state, scattered across time, behavior, physiology, and environment. Direct lifelog predictors aggregate these observations into a single feature vector and predict labels without explicitly preserving how observations explain one another. We propose the Human-State Joint-Embedding Predictive Architecture (HS-JEPA), a prototype that organizes each subject-day into nine label-agnostic semantic views, masks one view, and asks how much of it the rest of the day can explain. Each target is decomposed into a prediction, a residual unexplained by the fitted predictor, and its magnitude (surprise energy). Prediction improved on the train-fold mean-token baseline in all 36 original view-by-split combinations. Controls across 53 settings examined view construction; overlap-preserving random groupings predicted better, so error alone does not validate semantic design. External reconstruction on a fixed 114-participant College Experience Study (CES) holdout improved as training cohorts grew from 10 to 50 participants and sensing histories increased. In the original transductive 10-subject label evaluation, 9-D surprise energy achieved the numerically lowest overall Log Loss (0.6761), without statistically established superiority. We therefore present state decomposition as an exploratory analytical framework, not a claim of superior label prediction.**

I. Introduction

Sensors do not measure human state. They record only its partial traces, manifested in behavior, physiology, and environment. Smartphone and wearable sensing has made the regularities of sleep, mobility, social interaction, and mental well-being observable [1], [12], but the gap between observation and state remains. Subjective sleep satisfaction and sensor-based sleep indicators are not the same observation: the former carries personal appraisal, memory, expectation, and the response process, while the latter reflects physiological and behavioral processes together with measurement-device error [2]. In recurring-subject lifelog settings, this gap becomes problematic because heterogeneous observations are often collapsed into a single daily vector, obscuring which aspects of the day explain the reported state. When a few subjects are observed repeatedly and every log of a day is aggregated into a single vector for direct label prediction, a model can come to depend not on what changed that day but on who the person is. Lifestyle habits and device-use patterns are stable, and what is stable is easy to predict.

The Joint-Embedding Predictive Architecture (JEPA) offers an alternative: instead of reconstructing raw input, it predicts the representation of a masked target from observed context [3]. I-JEPA demonstrated representation prediction in images using informative target blocks and sufficient context [4]. In life logs, these units are not given. An image patch is given by spatial layout; the meaningful parts of a day must be defined. A morning rhythm, daytime mobility and physical load, evening social activity, nighttime sleep pressure—the temporal and behavioral relationships that compose a life come first, and only then do prediction units exist. Random column masking can still capture statistical dependencies among features, but it does not assign an explicit human-life interpretation to each prediction unit.

This is where HS-JEPA begins. It organizes each subject-day into nine label-agnostic human-life semantic views for cross-view prediction, masks one view at a time, and predicts its representation from the remaining views. The prediction represents the component explained by the other views; the residual captures target-view variation not explained by the current context; and surprise energy quantifies the magnitude of that residual. Unlike direct predictors that collapse all daily observations before label fitting, HS-JEPA preserves view identity and explicitly represents both cross-view predictability and unexplained view-specific variation. What cannot be explained is not discarded; it is measured. These components form candidate substates that retain shared structure alongside variation unexplained by the fitted predictor. Fig. 1 summarizes this workflow.

We do not judge a human-state representation by view-prediction performance alone. Cross-view predictability, identity information associated with subject-specific habits, and label-readout utility are analyzed separately. Partition controls examine the nine-view design, and separate College Experience Study (CES) experiments assess masked-view reconstruction with larger training cohorts and longer sensing histories; neither establishes independent-cohort label generalization.

Our contributions are threefold. First, we formulate label-agnostic masked-view prediction for life logs using groups with explicit temporal and sensor-based interpretations. Second, we decompose each target representation into a pre-

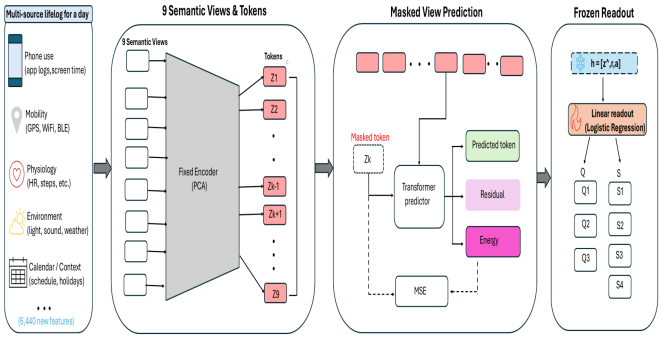


Fig. 1. HS-JEPA workflow. Nine semantic-view tokens are masked and predicted one at a time; predictions, residuals, and energies form the frozen state $\mathbf{h}_i = [\hat{\mathbf{z}}_i, \mathbf{r}_i, \mathbf{a}_i]$, evaluated by logistic label probes.

dicted component, a view-specific residual, and surprise energy. Third, we provide partition and view-removal controls alongside identity and label-readout evaluations, distinguishing the analytical use of state decomposition from predictive performance.

II. HS-JEPA Methodology

A. Human-Life Semantic Views

The nine label-agnostic views group daily features by time interval and sensor type to describe daily rhythm, mobility, physical load, environment, and departures from routine. They use app and screen activity, mobility, WiFi/BLE, heart rate, pedometer data, illumination, ambient sound, and calendar information. Views may share features; Section III-C examines this overlap and alternative groupings.

For subject-day i and view k, subject-wise residualization expresses features $\mathbf{x}_{i,k}$ relative to the subject's usual pattern:

$$\widetilde{\mathbf{x}}_{i,k} = \frac{\mathbf{x}_{i,k} - \boldsymbol{\mu}_{s(i),k}}{\boldsymbol{\sigma}_{s(i),k} + \epsilon}. \tag{1}$$

Here, $s(i)$ identifies the subject of day i, and $\boldsymbol{\mu}_{s(i),k}$ and $\boldsymbol{\sigma}_{s(i),k}$ are feature-wise means and standard deviations across that subject's days for view k. In the original ETRI setting, subject statistics and projections use all 700 unlabeled rows.

B. View-Token Predictive Core

Fixed view-specific encoders E_k use standardization and principal component analysis (PCA) to produce eight-dimensional tokens. For each target k, an independent predictor g_k receives the remaining tokens:

$$\begin{aligned} \mathbf{z}_{i,k} &= E_k(\widetilde{\mathbf{x}}_{i,k}) \in \mathbb{R}^8, \quad k = 1, \ldots, 9, \\ \hat{\mathbf{z}}_{i,k} &= g_k\big(M_k(\mathbf{z}_{i,1}, \ldots, \mathbf{z}_{i,9})\big), \end{aligned} \tag{2}$$

where M_k masks view k. Each g_k is a one-layer Transformer with hidden dimension 48, four attention heads, and view-position embeddings. Predictors are trained independently for 35 epochs using AdamW and token MSE; encoders remain fixed and no EMA teacher is used. Section III-C also tests a shared predictor conditioned on the target view.

Prediction residuals measure cross-view error, unlike subject-wise centering; surprise energy summarizes their magnitude:

$$\mathbf{r}_{i,k} = \mathbf{z}_{i,k} - \hat{\mathbf{z}}_{i,k}, \qquad a_{i,k} = \tfrac{1}{8}\|\mathbf{r}_{i,k}\|_2^2. \tag{3}$$

The full state $\mathbf{h}_i = [\hat{\mathbf{z}}_i, \mathbf{r}_i, \mathbf{a}_i] \in \mathbb{R}^{153}$ concatenates $72+72+9$ dimensions, rather than forming a jointly learned latent. Components are also evaluated separately.

C. Linear Probe on Frozen Representations

To assess label utility, frozen representations undergo identical imputation, standardization, PCA (at most 24 dimensions), and L2-logistic probing per target. Regularization is selected by inner GroupKFold within each outer subject-heldout fold. Predictor-training scopes differ between the original and additional evaluations (Section III-A).

III. Empirical Evaluation and Analysis

A. Data and Evaluation Protocols

The ETRI dataset contains 450 training and 250 evaluation rows from the same 10 subjects. Probes use the 450 training rows and seven binary targets: Q1–Q3 (subjective responses) and S1–S4 (sleep/measurement targets). The original cross-view diagnostics use subject-heldout, cohort-heldout (clusters of 60-D event fingerprints), row-block (contiguous global rows), and chronological splits (the final 28% per subject), with at most two folds per family. Partition controls reuse seven stored folds.

Original label evaluations pretrain on all 700 unlabeled rows, holding out subjects only for downstream probes. Additional label evaluations also exclude evaluated subjects from predictor training. Both retain global label-free preprocessing and are therefore transductive with respect to unlabeled features. External CES reconstruction (Section III-E) is separate from Q/S label evaluation.

B. Cross-View Predictability

We first compare cross-view prediction with a constant baseline: the training-fold mean token. The original predictors reduced MSE in all 36 view-by-split-family combinations,

with mean variance-weighted $R^2 = 0.445$. Place/environment, routine break, night sleep pressure, and body fatigue had the strongest values. Variance-weighted R^2 uses component-wise validation means, whereas MSE is compared with the training-mean baseline; the latter can improve even when the former is negative.

C. View Construction and Information Sharing

To test whether this predictability depends on view construction, we evaluated 53 settings spanning 49 partition configurations, including five seeds for the semantic grouping, 24 overlap-preserving random partitions, overlap removal, time- and sensor-group partitions, single-view removals, and varying view counts. Ridge and a target-conditioned shared Transformer use identical preprocessing and the same seven folds (Table I).

TABLE I
CROSS-VIEW PREDICTION ERROR BY VIEW PARTITION

View partition	Views	NMSE ↓	
		Ridge	Shared Transformer
Semantic	9	0.266	0.479 ± 0.008
Random, overlap preserved	9	0.212 ± 0.012	0.391 ± 0.013
Semantic, overlap removed	9	0.527	0.668
Time-based	6	0.566	0.659
Sensor-group-based	14	0.714	0.770

NMSE is normalized by train-fold-mean baseline MSE. ±: standard deviation over five seeds (semantic shared Transformer) or 24 partitions (random). Other entries are single runs.

Across 5,803 prediction-evaluation cells, 5,788 (99.7%) beat the train-fold mean baseline. These are repeated, not independent, evaluations. Random partitions predicted better than semantic grouping, while removing semantic overlap raised Ridge NMSE from 0.266 to 0.527. Prediction error alone therefore does not validate semantic grouping; its role is to connect predictions and residuals to daily-life aspects.

For both predictors, errors in the remaining views increased most after removing routine break, night sleep pressure, place/environment, or body fatigue. View size and overlap also affect these changes; they do not measure behavioral-concept importance.

D. Subject-Identity Audit

Predictability may reflect stable personal patterns rather than daily changes. We therefore compare row-level linear subject-ID accuracy under matched preprocessing (Fig. 2(a)).

Accuracy fell from 0.964 for raw views to 0.078 for residualized tokens; predictions, residuals, and energies remained near the reported 0.127 reference. The reduction occurred primarily during residualization, not predictive training, and does not establish identity removal.

E. External Cohort Scaling and History

To test reconstruction with more participants and longer histories, we use 214 CES participants [13] with at least 90 observed days (215,871 participant-days), holding out 114 participants (113,569 rows) from predictor training. Dataset-specific fields map to nine semantic roles with eight subject-relative coordinates per view. Each predictor receives 500 updates; values are means and standard deviations over three seeds. Micro $R^2 = 1 - \mathrm{SSE}/\mathrm{SST}$ and $\mathrm{NMSE}_{\mathrm{CES}} = 1 - R^2$, unlike the train-mean-normalized NMSE in Table I.

With 10, 25, and 50 training participants (Fig. 2(b)), micro R^2 rose from 0.6522 ± 0.0105 to 0.6627 ± 0.0022 and 0.6669 ± 0.0034.

Fixing 100 training participants, the history experiment uses the first 14, 30, 60, and 90 observed days and full histories (Fig. 2(c)). From 14 days to full history, micro R^2 rose from 0.5598 ± 0.0093 to 0.6647 ± 0.0030, and $\mathrm{NMSE}_{\mathrm{CES}}$ fell from 0.4402 to 0.3353 (23.8%). These gains concern reconstruction, not cold-start transfer or future Q/S prediction.

F. Label-Readout Utility

Returning to ETRI, we ask whether the representations also support label prediction. Table II compares representations from one canonical HS-JEPA run under the original transductive nested five-fold subject-grouped probe. Q and S average their respective target families; Overall averages all seven targets. The 9-D surprise energy had the numerically lowest overall Log Loss.

Table III extends this comparison to six adapted JEPA variants [5], [6], [7], [8].

HS-JEPA beat JETS and the prior baseline in only three of five outer folds; fold-clustered sign-flip tests were not significant. Additional nested subject-grouped probes of the

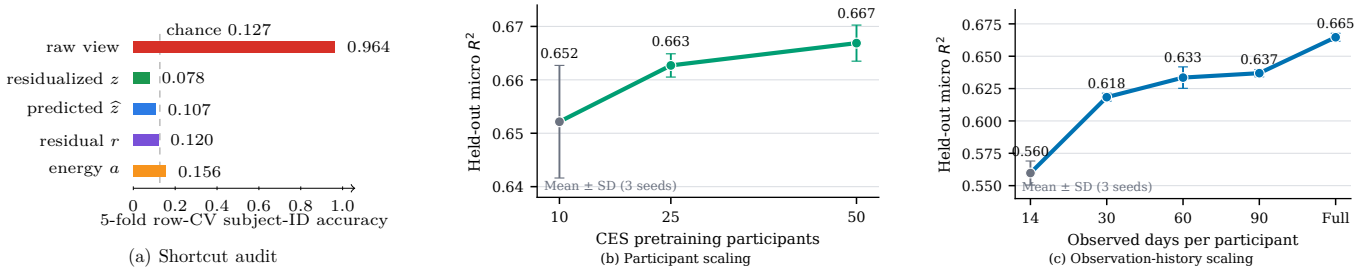


Fig. 2. Separate diagnostics: (a) subject-ID accuracy; (b) CES participant scaling (10–50 training participants); (c) CES history length (100 training participants). Panels (b,c) use 114 predictor-heldout participants; error bars denote standard deviations over three seeds.

TABLE II
LABEL-READOUT LOG LOSS OF HS-JEPA REPRESENTATIONS

Representation	Q	S	Overall
Prior only	0.6937	0.6660	0.6779
Raw semantic views (355-D)	0.9154	0.9834	0.9542
Residualized token $\mathbf{z}$	0.6953	0.6634	0.6771
Predicted token $\hat{\mathbf{z}}$	0.6940	0.6670	0.6785
Residual token $\mathbf{r}$	0.6922	0.6655	0.6770
Energy field $\mathbf{a}$	0.6915	0.6645	0.6761
Full $[\hat{\mathbf{z}}, \mathbf{r}, \mathbf{a}]$	0.6943	0.6631	0.6764

Lower Log Loss is better. One canonical run; original transductive five-fold subject-grouped probe.

TABLE III
EXPLORATORY LABEL-READOUT COMPARISON WITH ADAPTED JEPA INDUCTIVE BIASES

Representation	Q	S	Overall
HS-JEPA surprise energy	**0.6915**	0.6645	**0.6761**
HS-JEPA full state	0.6943	**0.6631**	0.6764
Prior only	0.6937	0.6660	0.6779
JETS adapted	0.6973	0.6634	0.6780
CHARM adapted	0.6995	0.6635	0.6789
T-JEPA adapted	0.6992	0.6644	0.6793
TS-JEPA adapted	0.6999	0.6668	0.6810
PhysioJEPA proxy	0.7004	0.6751	0.6860
SleepJEPA proxy	0.6963	0.6807	0.6874

Pretraining uses the same 700 unlabeled rows; probes follow Table II. Bold denotes column minima. Adapted/proxy entries preserve selected biases, not the original modalities, scales, or systems.

partition-control states (Section III-C), excluding evaluated subjects from predictor training but retaining global preprocessing, likewise showed no significant gain over the prior. For the semantic partition, the shared Transformer's state yielded Log Loss of 0.6770–0.6816 across seeds. Thus, analytical state decomposition and label-prediction utility should not be conflated.

IV. CONCLUSION

HS-JEPA separates cross-view predictions, residuals, and surprise energies for life-aspect-specific analysis. Partition and view-removal controls reveal predictive relationships without establishing semantic-grouping superiority. On the fixed CES holdout, masked-view reconstruction improved as the training cohort increased from 10 to 50 participants and as the observation history expanded from 14 days to the full history. However, Q/S label evaluation remains limited to 10 recurring subjects and showed no significant advantage. Fold- and history-local preprocessing and independent-cohort label evaluation are needed to assess practical utility.

ACKNOWLEDGMENT

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2026-25480920).

REFERENCES

[1] R. Wang et al., "StudentLife: Assessing mental health, academic performance and behavioral trends of college students using smartphones," in *Proc. ACM UbiComp*, 2014.

[2] K. Niemeijer, M. Mestdagh, and P. Kuppens, "Tracking subjective sleep quality and mood with mobile sensing: Multiverse study," *J. Med. Internet Res.*, vol. 24, no. 3, 2022.

[3] Y. LeCun, "A path towards autonomous machine intelligence," *OpenReview*, 2022.

[4] M. Assran et al., "Self-supervised learning from images with a joint-embedding predictive architecture," in *Proc. IEEE/CVF CVPR*, 2023.

[5] H. Thimonier et al., "T-JEPA: Augmentation-free self-supervised learning for tabular data," in *Proc. ICLR*, 2025.

[6] S. Ennadir, S. Golkar, and L. Sarra, "Joint embeddings go temporal," in *NeurIPS Workshop on Time Series*, 2024.

[7] E. Xie et al., "JETS: A self-supervised joint embedding time series foundation model for behavioral data in healthcare," in *NeurIPS Workshop on Learning from Time-Series for Health*, 2025.

[8] U. Dutta et al., "Giving sensors a voice: Multimodal JEPA for semantic time-series embeddings," in *Proc. ICML*, 2026.

[9] H. Zou, T. Hastie, and R. Tibshirani, "Sparse principal component analysis," *J. Comput. Graph. Stat.*, vol. 15, no. 2, pp. 265–286, 2006.

[10] B. Kim et al., "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors," in *Proc. ICML*, 2018.

[11] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proc. EMNLP-IJCNLP*, 2019.

[12] S. Yao et al., "DeepSense: A unified deep learning framework for time-series mobile sensing data processing," in *Proc. WWW*, 2017.

[13] S. Nepal et al., "Capturing the college experience: A four-year mobile sensing study of mental health, resilience and behavior of college students during the pandemic," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 8, no. 1, 2024, doi: 10.1145/3643501.