RETRACE: Resilience-Guided Trait-Conditioned Craving Estimation from Wearable Physiology in Opioid Use Disorder

YI XIAO, Arizona State University, USA
HARSHIT SHARMA, Arizona State University, USA
DESSA BERGEN-CICO, Syracuse University, USA
ASIF SALEKIN, Arizona State University, USA

Detecting opioid craving from wearable physiological signals is critical yet difficult, with the potential to support proactive interventions for individuals with opioid use disorder (OUD). This challenge is especially pronounced under subject-independent evaluation because craving is subjective, heterogeneous, and often physiologically entangled with stress. Our empirical analysis shows that stress elicits strong and reproducible autonomic responses, while craving-related signals are weaker, sparse, and largely embedded within stress-related physiology. We further show that psychological resilience, which shapes stress regulation and craving vulnerability, is not reliably observable from short-term wearable windows, but can be captured through reusable subject-level proxies, including post-stress heart-rate recovery and autobiographical memory recall.

Motivated by these findings, we introduce ***RETRACE***, a resilience-guided trait-conditioned framework for subject-independent craving estimation from wearable physiology. ***RETRACE*** reframes craving detection as trait-conditioned physiological interpretation: rather than assuming the same physiological pattern has the same meaning across individuals, it uses resilience-related subject context to guide inference. Technically, ***RETRACE*** introduces a novel dual-encoder design that separates generalizable stress physiology from subject-specific craving interpretation. It combines a frozen stress-pretrained encoder with a resilience-conditioned craving encoder, using feature-level gating and representation-level fusion to enable lightweight personalization without target-user craving labels or per-user retraining.

We evaluate ***RETRACE*** on a novel multimodal OUD dataset containing wearable physiology, stress and craving annotations, post-stress recovery dynamics, and autobiographical narratives. Under strict leave-one-subject-out evaluation, ***RETRACE*** outperforms classical machine learning, domain generalization, and multimodal conditioning baselines, achieving up to 7% absolute improvement over the strongest baseline. We further evaluate whether these gains generalize beyond a single benchmark by testing component contributions, comparing physiological and semantic resilience guidance, interpreting resilience-conditioned feature use, and evaluating stability across visits, wrist placements, runtime constraints, and an in-the-wild feasibility setting. These results show that robust wearable craving detection requires not only stronger physiological representations, but also a principled way to interpret ambiguous physiological evidence under human heterogeneity.

1 INTRODUCTION

Craving is a central driver of relapse in substance use disorders (SUDs) and a major barrier to sustained recovery [34, 100]. Among SUDs, opioid use disorder (OUD), in particular, remains a severe public health crisis marked by frequent relapse, high mortality, and substantial societal burden [53, 115, 128]. Timely detection of heightened craving may support early awareness, personalized coping, and recovery support before craving escalates into relapse risk [36, 123]. Continuous, non-invasive monitoring using wearable physiological sensors offers a promising pathway for scalable craving detection outside clinical settings [44, 91].

Despite this promise, physiological craving detection remains fundamentally challenging, particularly under subject-independent evaluation where models must generalize to unseen individuals [59]. Unlike many physiological sensing tasks with relatively consistent autonomic signatures, craving-related physiology is often subtle, heterogeneous, and *embedded within broader stress-related autonomic activity* [16, 17]. Consequently, similar physiological patterns may carry different psychological meanings across individuals. Most existing approaches adopt a state-level inference paradigm, directly mapping short windows of physiological signals to craving

Authors' Contact Information: Yi Xiao, Arizona State University, Tempe, USA, yxiao124@asu.edu; Harshit Sharma, Arizona State University, Tempe, USA, hsharm62@asu.edu; Dessa Bergen-Cico, dkbergen@syr.edu, Syracuse University, Syracuse, New York, USA; Asif Salekin, Arizona State University, Tempe, USA, Asif.Salekin@asu.edu.

labels using population-trained models [16, 19, 51]. This assumes a stable mapping between physiology and craving across individuals. Under substantial inter-individual variability, however, this assumption becomes fragile, making short-window physiological modeling alone insufficient for capturing the subject-dependent meaning of craving-related signals.

A key factor underlying this variability is psychological resilience that shapes how individuals regulate and respond to stress [122, 132, 138]. Rather than manifesting as an instantaneous physiological signal, resilience is a relatively stable trait-level construct that governs how physiological responses unfold and are interpreted across individuals. Prior work suggests that *resilience is associated with inter-individual differences in craving sensitivity* [26, 29, 97, 102]. This suggests that craving detection requires more than learning invariant short-window physiological representations; it also requires modeling how stable subject-level factors shape the interpretation of physiological evidence. We therefore conceptualize craving detection as a trait-conditioned inference problem, in which individual-level context guides how physiological signals are mapped to reported craving.

To operationalize this perspective, we design a wearable craving detection framework that conditions physiological inference on subject-level resilience. Rather than assuming that each physiological window has the same meaning across individuals, the framework uses reusable subject-specific context to guide how wearable signals are interpreted. A key challenge is that resilience is not directly observable from short-term physiological windows, as shown in Section 5.1, and direct measurement may be difficult to scale [32, 116, 133] in everyday sensing settings. We therefore explore practical resilience proxies that can be collected once and reused during inference. Specifically, we consider two complementary forms of trait-level guidance: heart-rate recovery dynamics, quantified using HR-AUC after stress exposure [32, 138], and semantic representations derived from autobiographical memory recall [43]. HR-AUC provides a physiology-grounded measure of post-stress recovery under controlled conditions, while autobiographical narratives provide a lower-burden source of subject-level context that does not require controlled stress induction.

Building on these insights, we propose ***RETRACE***—a **re**silience-guided **tra**it-conditioned **c**raving **e**stimation from wearable physiology in opioid use disorder. ***RETRACE*** combines a stress-pretrained encoder with a resilience-related subject-level representation. The stress encoder learns transferable autonomic structure from reliable stress labels that partially overlap with craving physiology, providing a stable backbone for interpreting weaker and more heterogeneous craving signals (Section 2.1). Importantly, ***RETRACE*** does not treat resilience as a direct physiological marker of craving. Instead, it uses resilience as contextual guidance for inference, controlling which physiological patterns are emphasized for each individual and how strongly generalizable stress physiology vs. subject-specific autonomic patterns contribute to craving prediction. This enables principled, low-capacity personalization under strict subject-independent evaluation.

1.1 Contributions

This work makes the following contributions:

- **Trait-conditioned inference for physiological craving detection.** We introduce ***RETRACE***, to our knowledge, the first resilience-conditioned, stress-pretrained framework for subject-independent wearable craving detection in OUD. ***RETRACE*** moves beyond a one-size-fits-all mapping from wearable physiology to craving by using trait-level context to guide how signals are interpreted for each individual. This enables subject-aware craving inference with low-capacity personalization, without requiring the target user's craving labels or user-specific model retraining.
- **Reusable resilience-related guidance for wearable craving sensing.** To support this trait-conditioned inference, we operationalize two forms of subject-level resilience guidance or proxies that can be collected once and reused during wearable craving inference: post-stress heart-rate recovery dynamics, quantified by HR-AUC after stress exposure, and semantic representations derived from autobiographical memory recall.

Together, these proxies capture a practical design trade-off: HR-AUC provides a physiology-grounded measure under controlled conditions, while autobiographical recall offers a lower-burden source of subject-level context that may better support scalability.

- **A multimodal dataset for studying stress, craving, and resilience in OUD.** We introduce, to our knowledge, the first dataset that jointly captures wearable physiological signals, craving labels, stress responses, and resilience-related attributes from the same participants, including both physiological recovery measures and autobiographical recall-based proxies (Section 3). This dataset enables direct empirical investigation of how stress, craving, and resilience interact, moving physiological craving research beyond state-level modeling. To support future work, we will publicly release the wearable physiological signals, derived physiological resilience values, and encoded autobiographical memory representations. Due to IRB restrictions raw autobiographical transcripts will not be publicly released.
- **Comprehensive validation and deployment-oriented analysis of *RETRACE*.** We provide a systematic analysis of the relationship between stress, craving, resilience, and physiological signals (Section 5), revealing fundamental challenges for subject-independent craving detection and motivating the need to incorporate resilience-aware inference. We evaluate ***RETRACE*** under strict subject-independent settings (Section 7) against classical machine learning, domain generalization, and multimodal conditioning baselines, demonstrating that neither invariant representation learning nor treating resilience guidance as auxiliary input is sufficient for robust craving estimation (Section 7.2). We further conduct ablation studies to quantify the roles of the stress-pretrained encoder, resilience-conditioned craving encoder, feature-level gating, representation-level fusion, and training losses (Section 7.3). We compare physiological and semantic resilience guidance to examine their trade-offs between physiological fidelity and practical scalability, and analyze positive and negative autobiographical recall to understand how different forms of subject-level context contribute to inference (Section 7.2 and 7.3). Through interpretability analyses, we examine how resilience guidance changes physiological feature use and supports subject-aware interpretation (Section 7.7). We assess deployment feasibility through cross-visit and cross-wrist robustness tests, temporal reusability of guidance embeddings (Section 7.6), runtime analysis for efficient edge deployment (Section 10), and a 14-day EMA-based real-world feasibility study (Section 8). Finally, we discuss deployment considerations (Section 9), including efficiency, usability, and trade-offs between physiological and semantic resilience guidance, along with broader implications and limitations (Section 11).

2 RELATED WORKS

2.1 Physiological Craving Detection and Its Limitations

Wearable physiological sensing is widely used to infer affective states, especially stress, from signals such as electrodermal activity (EDA), heart rate and heart rate variability (HR/HRV), skin temperature, and motion [79, 82, 87, 107]. Stress detection has become a mature sensing task, supported by standardized protocols, dense annotations, and reproducible physiological signatures. Most methods follow a *state-level inference* paradigm, mapping short physiological windows to momentary labels with population-level models [103, 126].

Wearable craving detection, in contrast, remains largely feasibility-oriented. Prior studies have shown that self-reported craving can be inferred from biosignals using classical machine learning models such as k-nearest neighbors and support vector machines [16, 17]. However, these studies often rely on within-cohort or event-level evaluation, limiting evidence for subject-independent generalization. Related work has also focused on detecting objective opioid use events from wearables [19, 50], which exhibit more stable pharmacological patterns and longer temporal dynamics than subjective craving. A recent scoping review further identifies the lack of standardized datasets, subject-independent protocols, and principled modeling frameworks as major barriers in this domain [59].

Stress is a well-established trigger of craving, and stress exposure can increase subjective craving across settings [83, 111]. Because directly eliciting opioid craving through drug-related cues can raise ethical and clinical concerns, especially for vulnerable populations, many studies use stress induction as a controlled physiological probe while measuring craving independently through self-report [40, 113]. However, craving-related physiology is often subtle, heterogeneous, and embedded within broader stress-related autonomic activity [131]. This creates inconsistent signal-label relationships across individuals and challenges standard state-level inference, which assumes a shared mapping between physiology and labels. We therefore use standard physiological classifiers as baselines and motivate models that explicitly account for inter-individual differences in how physiological signals relate to craving.

2.2 Modeling Inter-Individual Variability in Physiological Sensing

A central challenge in physiological sensing is the substantial inter-individual variability in how physiological signals relate to internal psychological states. To address this challenge, prior work has explored approaches to improve generalization across subjects, tasks, and environments.

Domain generalization. Domain generalization (DG) and invariant learning methods aim to learn representations that remain robust across subjects, tasks, or environments by enforcing invariance across training domains [10, 70, 99]. These approaches have been applied to physiological sensing tasks such as emotion recognition and workload estimation, where models seek subject-invariant features that generalize to unseen users [9, 94, 135]. However, DG methods typically assume that a shared mapping exists between physiological signals and labels across individuals. In craving detection, this assumption may not hold because similar autonomic responses can carry different meanings depending on individual differences in stress regulation. Our evaluation includes representative DG methods: IRM [10], GroupDRO[99], and V-REx[70], as baselines for subject-independent craving detection.

Multimodal and auxiliary-information approaches. Another line of work addresses inter-individual variability by incorporating additional modalities or auxiliary information, such as behavioral signals, audio-visual cues, or personality-related features [105, 109, 130, 134]. These methods often fuse information using cross-attention [125], mixture-of-experts [108], or feature-wise modulation methods such as FiLM [90]. While effective for improving prediction, they generally treat auxiliary information as additional input rather than as context that governs how physiological evidence should be interpreted. Moreover, continuous audio or visual sensing can also raise privacy, burden, and deployment concerns in sensitive populations [85, 140]. We therefore compare against representative multimodal fusion models, including Cross-Attention, Mixture-of-Experts, and FiLM, to test whether using resilience-related guidance as an auxiliary input is sufficient compared with explicitly conditioning physiological interpretation.

2.3 Resilience as a Missing Factor for Interpreting Physiological Craving Signals

Psychological resilience shapes how individuals cope with stress and adversity, particularly in substance use disorder (SUD) populations. Higher resilience is associated with lower perceived stress, better mental health outcomes, reduced relapse risk, and stronger recovery-related resources [97, 137, 138]. It has also been linked to craving intensity, suggesting a role in craving vulnerability [96]. Beyond a self-reported trait or outcome, resilience is increasingly viewed as a regulatory process that unfolds over time, reflecting how individuals recover from stress and how autonomic arousal resolves after challenge [6, 32, 47, 98, 122, 132]. Because stress is a well-established trigger of craving [14, 49, 93, 112], differences in stress regulation may change how physiological responses relate to craving: the same autonomic pattern may reflect effective recovery in one individual but elevated craving vulnerability in another.

This creates a key challenge for wearable craving assessment. Short physiological windows capture momentary autonomic activity, whereas resilience reflects regulation and recovery over time [32]. Standard wearable models map these short windows to labels using population-level relationships [63, 141], assuming that similar physiological patterns have similar meanings across individuals. This assumption becomes fragile when stress and craving physiology overlap and resilience changes how physiological evidence should be interpreted.

Existing approaches to inter-individual variability, including domain generalization and multimodal fusion methods (Section 2.2), improve robustness by learning shared representations or adding auxiliary information as predictive input. However, they do not explicitly model how resilience guides the interpretation of physiological signals. This motivates our approach: rather than treating resilience as another predictive feature, we use resilience-related guidance as a conditioning signal for physiological interpretation. We operationalize this guidance through physiological recovery dynamics and autobiographical language representations, enabling subject-aware craving inference without subject-specific retraining.

3 User Study

Existing wearable-sensing datasets do not jointly capture physiological dynamics, stress and craving experiences, and subject-level factors such as resilience within a single controlled setting. To address this gap, we design a laboratory-based user study that collects synchronized multimodal data for craving detection under subject-independent evaluation. The resulting dataset includes: (1) continuous wearable physiological signals capturing multimodal autonomic responses; (2) task-level stress labels and self-reported craving annotations, enabling craving to be studied as a subjective state within stress-related conditions; and (3) subject-level autobiographical language used to derive resilience-related representations. This unified design enables, for the first time, a controlled analysis of how physiological dynamics, subjective craving, and resilience-related individual differences interact within the same experimental setting. The details and rationale of the study design are outlined below.

3.1 Participants and Recruitment

The study was approved by the Institutional Review Board (IRB) of **X University**, and all participants provided informed consent prior to participation. Participants were recruited from the local community and university population (control group) as well as through clinical and community partners (OUD group), and participated on a voluntary basis. Participants with OUD received a gift card valued at **XX USD** as compensation.

The dataset comprises 78 sessions, including 28 sessions from 24 individuals with OUD and 50 sessions from 50 control participants. Among OUD participants, data were collected either before or after medication for opioid use disorder (MOUD), referred to as *pre-dose* and *post-dose* sessions. While most participants contributed a single session, 4 individuals were recorded in both pre-dose and post-dose conditions, resulting in a total of 28 sessions from 24 individuals. To ensure strict subject-independent evaluation, these 4 individuals with repeated sessions were used exclusively for training and excluded from evaluation. The remaining participants each contributed a single session. This sample size is consistent with prior work in wearable sensing and HCI, where controlled user studies often require complex experimental protocols and multimodal data collection [15, 37, 81, 141]. These challenges are further amplified in vulnerable clinical populations such as individuals with substance use disorder, where recruitment and sustained participation are inherently more difficult [50, 51]. Participant demographics are summarized in Table 1.

3.2 Apparatus and Experimental Setup

All study sessions were conducted in a private room, either at a university research facility or within a hospital setting, with only the participant and a trained experimenter present. This controlled and private environment was designed to ensure participants' comfort and confidentiality, allowing them to express their emotions and

Group	Age	Gender	Race/Ethnicity
Control (n=50)	30.80 ± 9.60 [22–65]	Female: 35 (70.0%) Male: 15 (30.0%)	White: 30 (60.0%) Asian: 16 (32.0%) Black or African American: 3 (6.0%) Other: 1 (2.0%)
OUD (n=24)	35.67 ± 4.15 [29–47]	Female: 18 (75.0%) Male: 5 (20.8%) Non-binary: 1 (4.2%)	White: 14 (58.3%) Black or African American: 7 (29.2%) Other: 2 (8.3%) American Indian or Alaska Native: 1 (4.2%)

Table 1. Participant demographics. Values are reported as mean ± SD [range] or count (%).

subjective experiences related to stress and craving without concern for observation or interruption. Participants stood 8-12 ft from a display screen presenting visual stimuli while wearing an Empatica E4 wristband on both wrists. The E4 continuously recorded photoplethysmography (PPG), electrodermal activity (EDA), 3-axis accelerometer data (ACC), and skin temperature (TEMP).

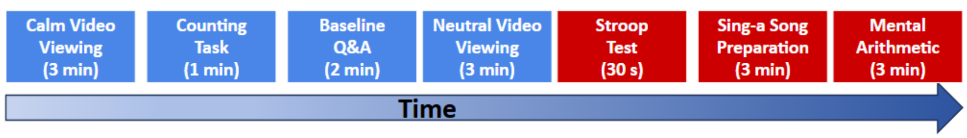

Fig. 1. Task sequence and timeline used in the user study. Blue blocks represent non-stress baseline tasks, and red blocks represent stress-inducing tasks.

3.3 Study Procedure

3.3.1 Study Protocol Justification: Our protocol is motivated by prior SUD laboratory studies [40, 113] showing that stress exposure can serve as a controlled cue for eliciting craving-related subjective and physiological responses [110]. Although in-the-wild craving data are important for ecological validity, relying on them can introduce substantial label noise. Craving episodes are often brief, self-reports may be delayed or missed, contextual triggers are uncontrolled, and physiological signals can be confounded. As a result, the temporal alignment between wearable physiology and reported craving can be imprecise, making it difficult to obtain reliable labels for model development. Reliable annotation is especially important in this work because, to our knowledge, this is the first study to examine whether trait-level subject information can modulate the interpretation of wearable physiological patterns for improved craving inference. To study this relationship carefully, we require a setting where physiological responses and craving reports can be collected with stronger temporal control.

Because prior studies [8] showed that direct opioid craving induction through a period of abstaining from opioid use may raise ethical, clinical, and standardization concerns, we use validated stress-induction tasks as reproducible physiological probes. At the same time, we do not treat stress as a proxy label for craving: stress is defined by task structure, while craving is measured independently through self-report after each task block. This design allows us to study craving-related physiology in a controlled and ethically appropriate setting while reducing label noise and preserving a clear distinction between stress exposure and subjective craving.

Notably, given the importance of ecological validity, we further conduct an in-the-wild feasibility study, described in Section 8.

3.3.2 Experimental Design and Task Protocol. Following prior work on laboratory-based affect and stress elicitation [7, 136, 141], the study protocol, as outlined in Figure 1, consisted of multiple phases or experimental blocks with interleaved intervals, beginning with non-stress baseline tasks and followed by stress-inducing tasks targeting cognitive, emotional, and social-evaluative stress systems. This structure allowed us to establish individual physiological baselines while exposing all participants to comparable stressors, keeping task-induced stress separate from the self-reported craving labels used in later analyses.

Non-Stress Tasks. Participants first completed a set of non-stress tasks designed to establish a relaxed autonomic baseline prior to stress induction:

(1) **Calm video viewing (3 minutes).** Participants watched a calming video depicting jellyfish floating in the ocean, which has been shown to elicit a relaxed physiological state [7, 136].
(2) **Counting task (1 minute).** Participants counted aloud from 0 to 59 at a self-selected, natural pace, a task commonly used to maintain engagement while preserving a non-stressful condition [107, 121].
(3) **Baseline question-and-answer task (2 minutes).** Participants responded to a sequence of neutral, everyday questions (e.g., their name, whether they had eaten breakfast, or their favorite restaurant) in a relaxed setting.
(4) **Neutral video viewing (3 minutes).** Participants watched a National Geographic documentary on pyramids, serving as a neutral, low-arousal baseline condition.

These tasks served as control conditions for comparison with physiological and behavioral responses observed during subsequent stress-inducing tasks. Calm video viewing is intended to actively induce a relaxed state, whereas neutral video viewing provides a low-arousal baseline without explicitly promoting relaxation.

Stress-Inducing Tasks. Participants then completed a set of validated stress-induction tasks adapted from prior psychology and behavioral research [27, 64, 101, 120]:

(1) **Stroop task [46, 60, 89] (3 minutes).** Participants viewed color words whose ink color differed from their semantic meaning and were instructed to verbally identify the ink color. The stimulus presentation rate gradually increased over the task duration, inducing cognitive stress through time pressure and interference.
(2) **Sing-a-Song Stress Test (30-second preparation).** Participants were unexpectedly informed that they would be asked to sing aloud while being observed and were given 30 seconds to prepare [136]. Consistent with prior work, only the preparation phase was analyzed, as it reliably elicits social-evaluative stress [7].
(3) **Mental arithmetic task (3 minutes).** Participants performed serial subtraction by counting backward from 1000 in increments of 17. Upon making an error, they were instructed to restart from 1000, following the protocol of the Trier Social Stress Test [64].

3.3.3 Stress and Craving Annotations. Stress labels were defined by task structure (baseline vs. stress blocks), leveraging established stress-induction protocols that elicit reliable and reproducible physiological responses [82, 87, 141]. Craving, in contrast, was assessed independently through self-report because it is a subjective motivational state that cannot be determined solely from task identity [40, 75, 110, 113]. Although stress-inducing tasks may create conditions in which craving can arise, craving is not assumed to occur during every stress block and may also occur during non-stress periods, as shown in Section 4.2. Following each task block, participants with OUD were asked to rate their perceived opioid *craving* on a 0–3 Likert scale (0 = not at all, 3 = extremely). For binary classification, craving labels were derived by thresholding self-reports, with ratings of 0 mapped to non-craving and ratings of 1–3 mapped to craving. Control participants were not asked about opioid craving.

Because craving can persist over short time intervals rather than occurring only at a single instant [11, 23, 48], each post-task craving report was treated as reflecting the participant's craving state over the corresponding task interval. This allowed us to assign craving labels to physiological windows within that interval while keeping craving defined as a participant-reported subjective state.

3.4 Autobiographical Memory and Context-neutral Speech Collection

As part of the study procedure, participants provided short spoken narratives to derive subject-level autobiographical language representations. At the beginning of the session, participants first completed a context-neutral speech task, responding to simple prompts (e.g., their name, daily routines, or general questions) for two minutes. This segment captures general linguistic characteristics without autobiographical or emotional content.

Participants then described one recent positive and one recent negative personal experience, each for two minutes, in a neutral, non-stressful setting independent of the stress-induction tasks. All speech segments were audio-recorded and transcribed. On average, context-neutral responses contained approximately 160 words, while autobiographical narratives contained approximately 210 words each (approximately 420 words per participant).

4 Dataset and Feature Construction

Physiological signals were processed using a unified pipeline to ensure consistent analysis across participants. We describe preprocessing, windowing, and feature extraction, followed by higher-level feature organization and dataset characterization used in subsequent analyses.

4.1 Physiological Signal Processing

Physiological signals collected from the Empatica E4 wristband [78], including EDA, PPG, 3-axis ACC, and skin TEMP, were temporally synchronized using device timestamps. EDA and PPG signals were processed using the NeuroKit toolkit [76], including denoising and decomposition into tonic and phasic components for EDA, and artifact removal with heart rate and inter-beat interval extraction for PPG. ACC and TEMP signals were aligned to ensure cross-modal consistency. Physiological recordings were segmented into overlapping windows of 30 seconds with a step size of 15 seconds [31, 107]. This window length balances temporal sensitivity with signal stability and is widely used in short-term autonomic state modeling. Each window is treated as an independent observation. For task-level analyses in Section 5.1, window-level features are aggregated within each task interval.

Feature Extraction. For each window, we extracted multimodal physiological features characterizing autonomic activity from EDA, PPG-derived cardiovascular signals, acceleration (ACC), and skin temperature (TEMP). EDA features include SCR-based measures (onsets, peaks, amplitudes, and recovery dynamics) together with tonic/phasic summary statistics, while cardiovascular features capture heart rate and HRV in time and frequency domains. Across modalities, additional FLIRT-style descriptors [39] were computed, including summary statistics (e.g., mean, median, standard deviation, variance, interquartile range, minimum, maximum, and percentile summaries), shape descriptors (e.g., RMS, mean absolute value, peak amplitude, waveform length, crest factor, and zero crossings), and spectral descriptors (e.g., total spectral power, spectral centroid, peak frequency, bandwidth, spectral entropy, and band-limited power features). Low-quality features were removed based on missingness and low variability, yielding a final 303-dimensional window-level representation used for all models. Additional details are provided in Appendix B.

4.2 Dataset Characteristics

This section characterizes the distribution of physiological windows and craving labels in the collected dataset. This analysis is important because craving is a subjective state: even under the same task protocol, participants may differ in whether craving occurs, how often it occurs, and during which task blocks it is reported.

Figure 2 summarizes subject-level variation in physiological window availability and craving labels. While the number of windows is broadly comparable across participants, the distribution of craving labels varies substantially.

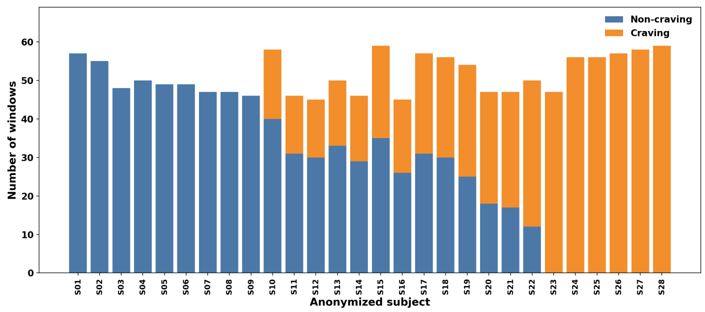

Fig. 2. Subject-level distribution of craving and non-craving windows among OUD participants. Each bar corresponds to one anonymized subject and is ordered by the proportion of craving windows.

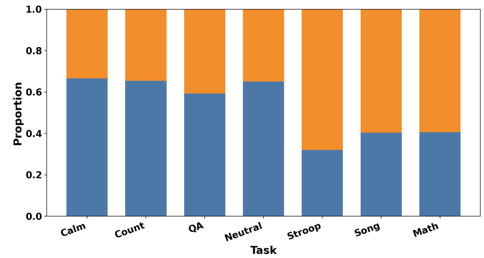

Fig. 3. Distribution of craving and non-craving window ratio among each tasks.

This highlights strong inter-subject variability in craving experience, even under controlled laboratory conditions. Figure 3 shows the distribution of craving and non-craving labels across tasks. Although stress-inducing tasks are associated with a higher prevalence of craving, craving does not occur deterministically under stress and is also observed during non-stress conditions. This confirms that stress and craving should not be treated as equivalent labels: stress is defined by task structure, whereas craving remains a participant-reported subjective state.

Together, these patterns show that craving labels are heterogeneous across both participants and tasks. The dataset, therefore, presents two key challenges for wearable craving detection: uneven craving prevalence across individuals and non-uniform relationships between stress exposure and reported craving. These challenges motivate subject-independent evaluation and modeling approaches that can account for individual differences in how physiological signals relate to craving.

5 Empirical Motivation: Physiological Craving Signals and Resilience Proxies

Before introducing the ***RETRACE*** framework, this section provides the empirical and literature-backed rationale for our modeling design. We first characterize how stress, craving, and resilience are reflected in short-term wearable physiology. The analysis shows that stress produces robust autonomic responses, craving-related signals are weaker and more heterogeneous, and resilience is not directly observable from short-term physiology. We then describe HR-AUC as a physiology-grounded resilience proxy and evaluate whether autobiographical recall can provide a lower-burden resilience-related representation. Together, these analyses and discussions motivate ***RETRACE***'s design: using stress-pretrained physiological structure as a backbone while conditioning craving inference on reusable subject-level guidance from resilience-related proxies.

5.1 Characterizing Stress, Craving, and Resilience in Short-Term Wearable Physiology

This section analyzes whether stress, craving, and resilience exhibit distinguishable physiological patterns in short-term wearable physiological signals. This question is central to wearable craving detection because deployed models typically operate on short physiological windows and must determine whether those windows contain craving-relevant information that can generalize across individuals. Stress is expected to produce robust autonomic responses and therefore serves as a useful reference condition. Resilience, however, is a relatively stable trait-level factor rather than a momentary state, so we examine whether it appears in instantaneous physiology or instead requires reusable subject-level descriptors.

5.1.1 Analysis overview. Following prior work on task-level physiological analysis from wearable sensors [63, 135, 141], we aggregate window-level features to the task-block level because stress and craving annotations were collected after each task rather than at individual window timestamps. Task-level features were baseline-corrected using each participant's neutral *Calm video viewing* task (Section 3.3.2) as a reference [20, 55, 56].

We then estimated task-level physiological associations using population-average generalized estimating equations (GEE), clustering by participant to account for repeated observations within individuals [74]. Stress and craving are defined based on the annotation protocol described in Section 3.3.3. Resilience is operationalized as a subject-level variable based on post-stress heart-rate recovery, quantified by HR-AUC, and discretized into high- and low-resilience groups using a median split for the present analysis. The rationale for this proxy and the grouping procedure are detailed in Section 5.2 and Appendix H. For each construct (stress, craving, and resilience), we fit separate GEE models with the construct as the main predictor, adjusting for subject group and the number of aggregated windows per task block. The estimated coefficients (β) represent population-average associations between the construct of interest and physiological features. To account for multiple comparisons across features, we apply the Benjamini–Hochberg false discovery rate (FDR) correction [127]. All reported significance results are based on FDR-corrected p-values. We analyze physiological modulation at two complementary levels.

At the system level, we capture coarse-grained modality-level effects by grouping features within four physiological systems—***Cardiovascular, Electrodermal, Movement, and Thermoregulation***—and computing PCA-based representations within each system, followed by GEE analysis on these representations (see Appendix C.1 for grouping details and rationale). *At the feature level,* we apply the same GEE framework directly to individual features (total 303) to capture fine-grained, feature-specific modulation. System-level results are summarized in Table 2, while Figure 4 visualizes feature-level effects across all features. In the figure, color intensity represents the magnitude of significant effects ($|\beta|$), and non-significant features are masked.

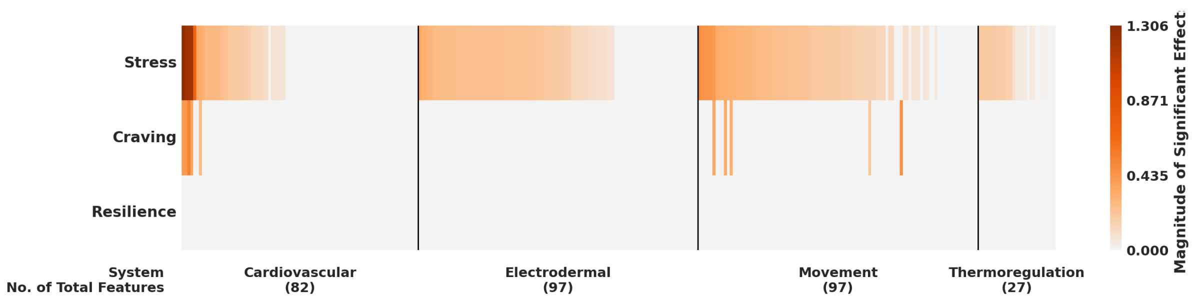

Fig. 4. Feature-level physiological effects across modalities. Each column corresponds to an individual physiological feature, grouped by system and ordered within each group by the magnitude of stress-related effects. Color intensity represents the magnitude of significant effects ($|\beta|$) estimated by GEE models, while non-significant features are masked. Rows correspond to different constructs (stress, craving, and resilience).

5.1.2 Stress-related physiological modulation. Stress elicited robust and widespread physiological responses.

At the feature level, stress was associated with increased variability and activation across multiple modalities. As shown in Figure 4, these effects are widespread, with a large number of features exhibiting significant modulation.

At the system level, these effects are consistent across all physiological modalities. As shown in Table 2, stress produces strong and statistically significant shifts across cardiovascular, electrodermal, movement, and thermoregulatory systems (all $p < 10^{-3}$).

Overall, these results confirm that the protocol elicited strong task-evoked autonomic modulation and that stress provides a reliable physiological reference condition.

Physiological System	Stress	Craving	Resilience
Cardiovascular	β=0.946, SE=0.284, $p<10^{-3}$	β=0.380, SE=0.685, p=0.579	β=−0.917, SE=1.173, p=0.434
Electrodermal	β=3.317, SE=0.673, $p<10^{-6}$	β=2.977, SE=1.816, p=0.101	β=3.525, SE=2.164, p=0.103
Movement	β=2.798, SE=0.471, $p<10^{-9}$	β=1.478, SE=1.462, p=0.312	β=1.504, SE=1.497, p=0.315
Thermoregulation	β=1.314, SE=0.289, $p<10^{-5}$	β=1.071, SE=0.484, p=0.0269	β=−0.962, SE=0.650, p=0.139

Table 2. System-level physiological modulation. Reported coefficients (β) summarize population-average task effects estimated by the GEE models. Positive β values indicate higher physiological activity under the task condition, while negative values indicate reduced activity. Larger absolute β values, when observed consistently across related features or physiological modalities, reflect stronger and more coherent task-evoked physiological modulation. Standard errors (SE) quantify uncertainty in these estimates, and p-values indicate whether the observed effects differ reliably from zero after accounting for within-subject dependence.

5.1.3 Craving-related physiological modulation. Compared with stress, craving-related physiological effects were weaker and more selective. At the feature level, craving-related modulation is sparse. As shown in Figure 4, only a small subset of features exhibit significant effects. After FDR correction [127], only 10 out of 303 features are significant, compared with 200 under stress. Notably, 9 out of these 10 craving-related features are also significant under stress, indicating a strong overlap between craving- and stress-related physiological responses, while a small number of features exhibit craving-specific effects.

At the system level, these effects are limited and less consistent across modalities. As shown in Table 2, only the thermoregulatory system shows a significant effect, while other systems do not reach statistical significance. Interestingly, although no individual thermoregulatory features are significant at the feature level, the system-level analysis reveals a significant effect. This discrepancy arises because aggregation captures multiple weak but consistent effects across features. While no single feature is strong enough to survive FDR correction, their combined variation forms a detectable pattern at the system level, indicating that craving-related thermoregulatory modulation is distributed and low-amplitude rather than driven by a single dominant feature.

Overall, these results indicate that craving has measurable physiological correlates, but these effects are weak, sparse, and largely embedded within broader stress-related activity, making craving difficult to distinguish from non-craving using short-term physiological signals alone.

5.1.4 Resilience-related physiological modulation. In contrast to stress and craving, resilience did not show a robust short-term physiological signature.

At the feature level, resilience-related effects do not exhibit a consistent pattern across features, as reflected in Figure 4. At the system level, resilience effects show substantial variability across modalities, with relatively large but highly uncertain estimates, and none reaching statistical significance (Table 2).

These results suggest that psychological resilience does not manifest as a stable task-evoked pattern that can be reliably detected from short-term wearable windows. Rather than treating resilience as an instantaneous physiological state, this motivates modeling it as a subject-level factor. This finding supports the use of reusable resilience-related descriptors, such as post-stress HR-AUC and autobiographical recall, to guide craving inference.

Summary and design implications. This section's analysis shows that craving has measurable population-level physiological correlates, supporting the feasibility of subject-independent wearable craving detection. However, these correlates are weak, selective, and largely embedded within broader stress-related activity, with limited distinct or craving-specific components, which reduces the effectiveness of a single population-level mapping from physiology to craving. In contrast, resilience is not directly detectable from short-term wearable windows. These findings motivate ***RETRACE***'s design: using stress-pretrained representations as a stable physiological backbone and using resilience-related subject-level guidance to interpret ambiguous craving-related signals across individuals. As outlined in Section 6.2, the stress-pretrained encoder provides a shared autonomic reference, while craving is learned separately through resilience-conditioned inference task that identifies craving-related deviations without treating stress itself as craving.

5.2 Post-Stress Heart-Rate Recovery as a Resilience Proxy

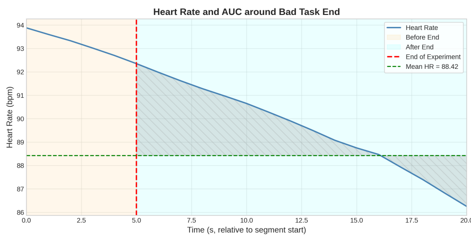

Fig. 5. Heart rate recovery AUC of participant 1. Smaller AUC indicates faster recovery (higher resilience).

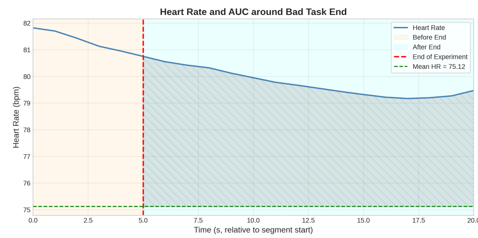

Fig. 6. Heart rate recovery AUC of participant 2. Larger AUC indicates slower recovery (lower resilience).

The preceding analysis confirms that resilience is not reliably detectable from short-term instantaneous wearable physiological signal windows. Following literature [32], we therefore use post-stress heart-rate recovery as a subject-level proxy for resilience-related regulatory capacity. Specifically, we compute the area under the curve (AUC) of heart rate during the post-stress recovery period, defined as the interval following the mental arithmetic task in the protocol described in Section 3.3.2. Smaller HR-AUC values indicate faster recovery toward baseline and are interpreted as higher physiological resilience, whereas larger HR-AUC values indicate slower recovery and are interpreted as lower physiological resilience.

Figures 5 and 6 illustrate two example recovery trajectories. In these examples, the participant with the smaller HR-AUC shows faster post-stress recovery, while the participant with the larger HR-AUC shows more prolonged physiological activation. HR-AUC is suitable for our analysis because it summarizes integrated autonomic recovery dynamics, including sympathetic withdrawal and parasympathetic reactivation [32]. We use this measure as a stable, interpretable, subject-level resilience proxy for subsequent analysis and resilience-guided craving inference.

5.3 Autobiographical Recall as a Lower-Burden Resilience Proxy

The previous section defines HR-AUC as a physiology-grounded proxy for resilience-related recovery. However, HR-AUC requires controlled stress exposure, a clearly defined recovery period, and stable baseline conditions. Although it can be collected once as a pre-deployment measure and reused during wearable-based craving inference for a target user, it still requires exposing the individual to a controlled stressor. This may add burden during practical deployments. We therefore examine whether autobiographical recall can provide a lower-burden representation of resilience-related individual differences.

5.3.1 Motivation for Autobiographical Recall. Autobiographical memory recall is relevant to resilience because it captures how individuals appraise, organize, and regulate emotionally meaningful experiences. Resilience is commonly conceptualized as the capacity to adapt to, recover from, and maintain functioning following adversity [32, 118]. This capacity is closely linked to emotion regulation, coping, affective flexibility, and stress-recovery processes [5, 26, 102, 117]. Because autobiographical narratives reflect how individuals interpret past experiences, construct personal meaning, and regulate emotional responses during recall, they may provide a stable subject-level window into resilience-related individual differences [4, 28, 77].

We consider both positive and negative autobiographical memories because they capture complementary aspects of resilience-related processing. Positive recall may reflect adaptive affective engagement, reward sensitivity, and access to positive self-relevant experiences, which are important for coping and psychological recovery [41, 119]. This is particularly relevant in OUD, where chronic opioid exposure can disrupt reward and stress systems, producing accumulated allostatic load, reduced sensitivity to non-drug rewards, and heightened stress reactivity [65–68]. Under this reward-deficit and stress-surfeit framework, positive autobiographical recall may not simply evoke pleasure; for some individuals, a mismatch between expected and experienced positive affect may instead produce ambivalence, anxiety, or physiological arousal [67, 68, 95, 104]. Thus, positive recall may encode individual differences in reward responsiveness, stress reactivity, and craving vulnerability.

Negative autobiographical recall, in contrast, directly engages adverse experiences and therefore provides a natural context for examining emotion regulation and stress recovery. Prior work shows that regulation strategies such as cognitive reappraisal shape how negative memories are recalled and experienced, including their detail, emotional tone, and negative bias [25, 88]. The ability to suppress or control intrusive memory retrieval is also considered an important component of emotion regulation and can reduce the emotional impact of unwanted recollections [35, 61]. Therefore, negative autobiographical narratives may reveal how individuals organize, reinterpret, and regulate adverse experiences. A detailed neurophysiological context motivation discussion is further provided in Appendix E.1.

5.3.2 Representational Alignment Analysis. To test whether autobiographical language captures resilience-related information, we evaluate its alignment with HR-AUC, the physiology-grounded recovery proxy defined in Section 5.2. We compare *Autobiographical Language*, derived from participants' positive and negative memory recall transcripts, with *context-neutral speech*, collected at the beginning of the study as outlined in Section 3.4. Context-neutral speech serves as a control condition for general linguistic properties such as verbosity and speaking style without autobiographical or emotional content.

High-dimensional language embeddings extracted from these narratives were treated as fixed representations and summarized using Partial Least Squares (PLS) [18]. Because text embeddings capture heterogeneous semantic and stylistic information, unsupervised methods such as PCA may emphasize dominant linguistic variation rather than resilience-relevant dimensions [2]. PLS instead provides a supervised dimensionality reduction approach that identifies latent directions where language representations covary with an external outcome, making it suitable for representation–outcome alignment [1, 69]. In our analysis, embeddings were projected onto a single PLS component aligned with physiological resilience, operationalized as continuous HR-AUC in Section 5.2. We assessed the association between this PLS-derived dimension and HR-AUC using Spearman rank correlation, with significance evaluated by permutation testing using 5,000 permutations [33]. Technical details of PLS are provided in Appendix F.

Language Condition	Spearman r	p (Spearman)	p (Permutation)
Autobiographical Language	0.843	2.9×10^{-17}	< 0.001
Context-neutral Speech	0.132	0.313	0.316

Table 3. PLS-based representational alignment between *Autobiographical Language* and HR recovery. Spearman rank correlations and permutation-based p-values are reported for associations between a one-dimensional PLS latent representation derived from Autobiographical Language embeddings and HR area-under-the-curve (AUC) following stress cessation.

As shown in Table 3, autobiographical language exhibits a strong and statistically reliable association with HR-AUC ($r = 0.843$, $p < 0.001$), whereas context-neutral speech shows weak and non-significant alignment ($r = 0.132$, $p = 0.316$). This contrast suggests that resilience-related recovery information is selectively reflected in autobiographical recall rather than in general speech patterns alone. *These findings support autobiographical recall as a lower-burden, reusable proxy for resilience-related subject context.*

Overall, the findings in Section 5 directly motivate our modeling approach and establish HR-AUC and autobiographical recall as resilience-related subject-level proxies.

6 *RETRACE*—a **re**silience-conditioned, **s**tress-**p**retrained **i**nference framework

6.1 Problem Formulation

We present a resilience-guided framework for subject-independent craving detection from wearable physiological signals. Given a physiological window $\mathbf{x} \in \mathbb{R}^d$, the goal is to predict a binary craving label $y \in \{0, 1\}$ derived from self-reported ratings. In addition to window-level observations, each subject is associated with a fixed embedding $\mathbf{e}$ that captures stable individual differences through a resilience proxy.

Our analysis in Section 5.1 suggests that wearable physiology contains measurable subject-independent information related to craving, supporting the feasibility of physiological craving detection. At the same time, craving-related subject-independent signals are weaker and more selective than stress responses, and they are often embedded within broader stress-related autonomic activity. This indicates that craving detection is feasible, but achieving more effective and reliable inference requires models that can interpret subtle, subject-dependent physiological evidence rather than rely on a single population-level mapping.

We therefore formulate craving detection as a subject-conditioned inference problem. The physiological window $\mathbf{x}$ provides momentary autonomic evidence, while the subject-level embedding $\mathbf{e}$ provides a stable context for interpreting that evidence. Under this formulation, similar physiological patterns can contribute differently to craving prediction across individuals. This enables subject-independent inference that accounts for inter-individual variability without requiring per-user craving labels or user-specific model retraining.

6.2 Architecture Overview

RETRACE is designed to separate two complementary components of wearable craving inference: (1) generalizable autonomic structure and (2) craving-related patterns that vary from person to person. As illustrated in Figure 7, the model contains two encoder pathways. A frozen **stress-pretrained encoder** provides a stable physiological reference, while a **conditioned craving encoder** learns adaptive craving-related representations.

The ***resilience embedding*** $\mathbf{e}$ modulates the inference process at two levels. At the input level, it guides which physiological features should be emphasized for a given individual. At the representation level, it controls how much the final prediction should rely on the stress-pretrained representation versus the conditioned craving representation. Together, these components allow ***RETRACE*** to perform low-capacity personalization through $\mathbf{e}$ while preserving a shared model structure.

6.3 Dual-Encoder Design

The dual-encoder design is motivated by the empirical findings in Section 5.1: stress produces robust physiological modulation, whereas craving-related signals are weaker, more selective, and largely embedded within stress-related autonomic activity. A single encoder would need to learn both stable autonomic structure and subtle craving-related variation at the same time, which can entangle stress and craving representations and reduce subject-independent generalization. ***RETRACE*** instead separates these roles into two pathways.

Stress-pretrained encoder. The stress encoder f_{stress} is pretrained on a stress prediction task using cross-entropy loss and then kept fixed during craving training. Given an input physiological window $\mathbf{x}$, it produces a stress-pretrained representation: $\mathbf{z}_{\text{stress}} = f_{\text{stress}}(\mathbf{x})$. The encoder is trained using data that excludes the target subject used for evaluation, ensuring subject-independent pretraining. Because this encoder is trained using stronger and more reliable stress labels, it captures generalizable autonomic structure across individuals. Importantly, this pathway is not used to treat stress as craving. Rather, it provides a stable physiological reference that anchors the

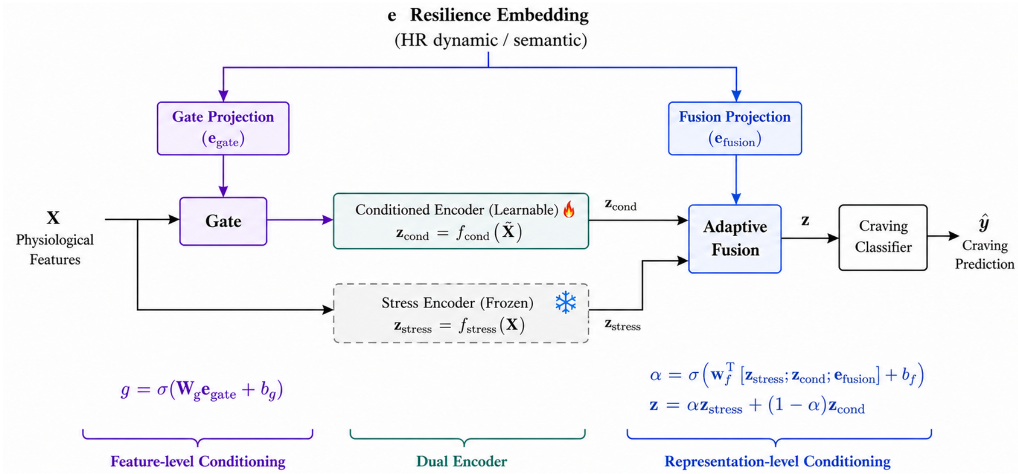

Fig. 7. Overview of the proposed resilience-guided architecture. A frozen stress encoder captures stable physiological structure, while a subject-conditioned encoder models adaptive variations. A resilience embedding modulates the input conditioning and representation fusion.

model in meaningful autonomic patterns, allowing weaker craving-related deviations to be interpreted relative to this reference.

Conditioned craving encoder. The conditioned encoder f_{cond} learns craving-related variation from resilience-conditioned inputs. Given a gated input $\tilde{\mathbf{x}}$, it produces: $\mathbf{z}_{\text{cond}} = f_{\text{cond}}(\tilde{\mathbf{x}})$. This pathway is trained with the craving objective and, through the 'orthogonality term objective' in Section 6.7, is designed to learn information distinct from $\mathbf{z}_{\text{stress}}$; thus, it focuses on physiological patterns that are subtle, heterogeneous, and potentially subject-dependent. Because input $\tilde{\mathbf{x}}$ is modulated by the resilience-related embedding, the encoder can emphasize different physiological evidence for different individuals. This allows the model to learn craving-related deviations that may not be fully captured by the stress-pretrained representation. In this way, craving is modeled as a separate resilience-conditioned inference task rather than as a direct extension of stress detection.

Together, the two encoders provide a structured representation space: the stress encoder anchors the model in generalizable autonomic physiology, while the conditioned encoder captures adaptive craving-related deviations.

6.4 Resilience-Guided Conditioning

To account for subject-dependent differences in how physiological signals relate to craving, ***RETRACE*** uses the resilience embedding $\mathbf{e}$ as a conditioning signal. Rather than treating $\mathbf{e}$ as an additional predictive feature, the ***RETRACE*** uses it to guide how physiological autonomic structure is interpreted and integrated.

As shown in Figure 7, the resilience embedding is projected into two conditioning signals: $\mathbf{e}_{\text{gate}}$ for feature-level conditioning and $\mathbf{e}_{\text{fusion}}$ for representation-level conditioning. These two pathways serve complementary roles. Feature-level conditioning helps the model determine which physiological features are more informative for a given individual. Representation-level conditioning controls how much the final prediction relies on the generalizable stress-pretrained representation versus the conditioned and adaptive craving-related representation.

6.4.1 *Feature-Level Conditioning.*

At the feature level, we introduce a gating mechanism that adaptively modulates the input representation based on the feature-level resilience embedding projection $\mathbf{e}_{\text{gate}}$. Specifically, we compute a gating vector: $\mathbf{g} = \sigma(W_g \mathbf{e}_{\text{gate}} + \mathbf{b}_g)$, and apply it to the input as $\tilde{\mathbf{x}} = \mathbf{x} \odot \mathbf{g}$, where $\sigma(\cdot)$ denotes the sigmoid function and $\odot$ represents element-wise multiplication. This mechanism assigns subject-specific importance weights to physiological features, enabling selective amplification or suppression of input dimensions. As a result, the same physiological signal can be interpreted differently across individuals, addressing variability in how signals relate to craving.

6.4.2 Representation-Level Conditioning. At the representation level, we adaptively combine different representations based on both physiological signal representations (Section 6.3) and the fusion projection of resilience embedding $\mathbf{e}_{\mathbf{fusion}}$. Given representations $\mathbf{z}_{\text{stress}}$ and $\mathbf{z}_{\text{cond}}$, we compute a fusion weight: $\alpha = \sigma\big(\mathbf{w}_f^\top[\mathbf{z}_{\text{stress}}, \mathbf{z}_{\text{cond}}, \mathbf{e}_{\mathbf{fusion}}] + b_f\big)$, and obtain the final representation as: $\mathbf{z} = \alpha\mathbf{z}_{\text{stress}} + (1-\alpha)\mathbf{z}_{\text{cond}}$.

This formulation allows the model to dynamically balance between a stable physiological reference of a broader activation and an adaptive representation. By conditioning the fusion process on $\mathbf{e}$, the model adjusts its decision behavior across individuals, accounting for individual differences in how physiological signals map to craving.

Together, these two forms of conditioning enable a resilience-guided inference process that adapts both feature utilization and decision behavior, without modifying the underlying physiological representation space.

6.5 Resilient Embedding Extraction

In this work, $\mathbf{e}$ can be instantiated in multiple ways. We consider two alternatives: (1) ***HR-dynamics guidance***, a physiological embedding derived from heart-rate (HR) dynamics, and (2) ***semantic guidance***, a representation obtained from autobiographical memory recall.

The ***HR-dynamics guidance*** is constructed from short (15s) heart-rate segments following calm and stress conditions. For each subject, heart-rate signals are normalized relative to a baseline period (Calm video viewing), and a single post-calm and post-stress segment is extracted after the corresponding task conditions (Section 5.2). These segments capture complementary baseline-like and stress-response recovery dynamics, enabling characterization of individual differences in recovery. Each segment is summarized using statistical and temporal descriptors (e.g., mean, standard deviation, signed and absolute AUC, slope). The resulting features from the post-calm and post-stress segments are concatenated into a fixed-length embedding (32-dimensional).

The ***semantic guidance*** is derived from autobiographical memory recall. Participants describe recent positive and negative experiences, which are transcribed and encoded using a pretrained language model into high-dimensional embeddings. The final representation is formed by concatenating the two narratives, yielding a fixed subject-level embedding (3072-dimensional).

Details on how these two guidance representations are extracted are provided in Appendix D.

6.6 Craving Prediction

The final representation $\mathbf{z}$ is passed to a prediction head to estimate the probability of craving: $\hat{y} = h_{\text{craving}}(\mathbf{z})$.

The model is trained using the binary craving labels derived from self-reported ratings. At inference time, ***RETRACE*** requires only the physiological window $\mathbf{x}$ and a one-time collected subject-level resilience embedding $\mathbf{e}$. Thus, it supports subject-aware craving inference without requiring craving labels from the target user or user-specific model retraining.

6.7 Training Objective

RETRACE is trained to predict craving using empirical risk minimization with additional regularization terms:

$$\mathcal{L} = \mathcal{L}_{\text{craving}} + \lambda_{\text{sparse}}\|\mathbf{g}\|_1 + \lambda_{\text{orth}}\|\mathbf{z}_{\text{stress}}^\top\mathbf{z}_{\text{cond}}\|_2 + \lambda_{\text{stress}}\,\mathrm{CE}\big(h_{\text{stress}}(\mathbf{z}_{\text{cond}}), y_{\text{stress}}\big).$$

The primary term, $\mathcal{L}_{\text{craving}}$, is the binary classification loss for predicting self-reported craving labels. The remaining terms serve as regularization to stabilize representation learning and support the proposed architecture.

The sparsity term $\|\mathbf{g}\|_1$ encourages the feature-level conditioning gate to emphasize only a subset of physiological features for each subject, rather than assigning high weights to all input dimensions. This supports selective, subject-aware feature use.

The orthogonality term $\|\mathbf{z}_{\text{stress}}^{\top}\mathbf{z}_{\text{cond}}\|_2$ encourages the stress-pretrained and conditioned craving representations to capture complementary information. This reduces redundancy between the two encoders and helps separate general autonomic structure from adaptive craving-related variation.

Finally, the auxiliary stress loss applies a lightweight stress prediction head $h_{\text{stress}}(\cdot)$ to the conditioned representation $\mathbf{z}_{\text{cond}}$. This term encourages the conditioned encoder to remain grounded in physiologically meaningful signals while learning craving-related patterns, reducing the risk of relying on spurious features.

The weights λ_{sparse}, λ_{orth}, and λ_{stress} control the strength of each regularization term. All auxiliary components are used only during training; at inference time, ***RETRACE*** requires only the physiological window $\mathbf{x}$ and the subject-level resilience embedding $\mathbf{e}$ for craving prediction.

Full architectural specifications for all components, layer configurations, and training hyperparameters for ***RETRACE*** are provided in Appendix A.

7 Evaluation

We evaluate the proposed ***RETRACE*** model on opioid craving detection, comparing against classical machine learning approaches, domain generalization (DG) and multimodal learning approaches as discussed in Section 2. We further conduct ablation and interpretability analyses to examine the role of subject-level conditioning.

7.1 Evaluation Protocol and Metrics

Evaluation Cohort. The OUD dataset contains 24 participants (28 sessions). We exclude participants with no within-subject craving variability and those with repeated sessions from evaluation, while retaining them for training. From the remaining eligible participants, 10 OUD subjects are randomly selected as the evaluation cohort.

Protocol and Data Statistics. We adopt a leave-one-subject-out (LOSO) protocol with 10 folds [24]. In each fold, one subject is held out for testing and the rest are used for training. For stress encoder pretraining, the stress encoder is trained separately within each fold using all control participants and the OUD training subjects in that fold, excluding the held-out test subject. For craving detection, only OUD participants are used. Across all OUD sessions used for craving modeling, the dataset contains 1,421 windows (785 non-craving and 636 craving). Under the LOSO protocol, the test folds collectively contain 494 windows (258 non-craving and 236 craving).

Metrics. We report micro-aggregated accuracy (Acc) and Macro-F1, computed on pooled predictions across test subjects, along with subject-wise mean ± standard deviation. Statistical significance is assessed using paired tests across subjects on accuracy (Acc), with p-values and confidence intervals (CI) reported relative to the proposed model with HR dynamics guidance.

7.2 Comparison with Baselines

We evaluate ***RETRACE*** with both resilience-related guidance representations introduced earlier: a physiology-grounded HR-AUC/HR-dynamics proxy from post-stress recovery and a semantic proxy from autobiographical memory recall. This comparison tests whether ***RETRACE***'s subject-level conditioning remains effective across both physiological and autobiographical forms of resilience-related guidance. Table 4 summarizes results across representative baselines for physiological sensing, as outlined in Section 2. These include classical machine learning models (MLP, SVM, XGBoost)[16, 17], domain generalization methods (IRM[10], GroupDRO[99], V-REx [70]), and multimodal conditioning architectures (FiLM [90], Cross-Attention[125], MoE[108]).

Model	**Guidance**	**Acc**	**Macro-F1**	**Mean ± Std**	ΔAcc	**p-value**	**CI**
MLP	–	0.581	0.578	0.579 ± 0.107	0.139	0.010	[0.063, 0.215]
RBF-SVM	–	0.630	0.624	0.623 ± 0.205	0.094	0.105	[0.009, 0.195]
XGBoost	–	0.638	0.632	0.629 ± 0.152	0.088	0.160	[0.003, 0.174]
IRM	–	0.619	0.616	0.619 ± 0.096	0.099	0.051	[0.015, 0.188]
GroupDRO	–	0.609	0.607	0.636 ± 0.114	0.108	0.050	[-0.0226, 0.1903]
V-REx	–	0.623	0.622	0.609 ± 0.115	0.094	0.066	[0.015, 0.173]
Cross-Attn	HR dynamics	0.650	0.647	0.646 ± 0.188	0.071	0.193	[-0.038, 0.189]
Cross-Attn	Semantic	0.607	0.607	0.610 ± 0.180	0.108	0.084	[0.006, 0.238]
MoE	HR dynamics	0.626	0.624	0.626 ± 0.095	0.092	0.037	[0.014, 0.167]
MoE	Semantic	0.621	0.613	0.621 ± 0.094	0.096	0.037	[0.024, 0.163]
FiLM	HR dynamics	0.630	0.625	0.631 ± 0.127	0.086	0.155	[-0.001, 0.181]
FiLM	Semantic	0.640	0.636	0.641 ± 0.157	0.077	0.375	[-0.033, 0.192]
RETRACE	**HR dynamics**	**0.721**	**0.720**	**0.718 ± 0.112**	–	–	–
RETRACE	**Semantic**	**0.704**	**0.694**	**0.699 ± 0.238**	–	–	–

Table 4. Baseline comparison on craving detection. Metrics are computed on pooled predictions across subjects. Mean ± std is computed across subject-wise folds. ΔAcc denotes the accuracy improvement relative to the proposed model with HR dynamics guidance (***RETRACE*** − Baseline). Confidence intervals (CI) are reported for accuracy.

Overall performance. ***RETRACE*** achieves the best performance across all metrics. In particular, the HR dynamics variant reaches an accuracy of 0.721 and Macro-F1 of 0.720, outperforming the strongest baseline (Cross-Attn with HR dynamics, 0.650 accuracy) by over 7%. The semantic variant also consistently exceeds all baselines (0.704 accuracy), demonstrating robustness across different forms of subject-level guidance. Notably, these improvements are achieved under a strictly subject-disjoint (LOSO) setting, highlighting the model's ability to generalize to unseen individuals.

Comparison with classical and domain generalization methods. Compared to classical models and domain generalization (DG) approaches, the proposed method consistently achieves higher performance. The best classical baseline (XGBoost) reaches 0.638 accuracy, while DG methods such as IRM and V-REx achieve 0.619 and 0.623, respectively, all substantially below our model (0.721). These methods rely on learning a shared mapping from physiological signals to labels across subjects, which breaks down under inter-subject heterogeneity. In contrast, our model explicitly conditions the interpretation of physiological signals on subject-level information, resulting in improved generalization.

Comparison with multimodal conditioning approaches. We further compare against multimodal baselines, including FiLM, Cross-Attention, and Mixture-of-Experts (MoE). While these methods incorporate subject-level information, their improvements remain limited (e.g., Cross-Attn with HR dynamics achieves 0.650 accuracy and FiLM reaches 0.640), still significantly below our model (0.721). This indicates that incorporating subject-level information alone is insufficient; how such information modulates physiological interpretation is critical. Our structured conditioning approach, combining feature-level gating and representation-level fusion, leads to consistent performance gains over these architectures.

Impact of guidance representation. The choice of subject-level guidance consistently affects performance across models. HR dynamics guidance outperforms semantic embeddings not only in the proposed model (0.721 vs. 0.704), but also across multimodal baselines (e.g., Cross-Attn: 0.650 vs. 0.607; MoE: 0.626 vs. 0.621). This suggests that representations aligned with physiological response patterns provide more effective conditioning signals. At the same time, the proposed model maintains strong performance under both guidance types, indicating that improvements are not solely driven by embedding choice, but by the conditioning mechanism itself.

Variant	HR dynamics Guidance			Semantic Guidance		
	Acc	F1	Acc Mean ± Std	Acc	F1	Acc Mean ± Std
(1) Component Ablations						
Remove fusion gate	0.619	0.617	0.616 ± 0.129	0.636	0.630	0.636 ± 0.157
Remove input gate	0.658	0.658	0.652 ± 0.147	0.654	0.649	0.648 ± 0.273
Remove frozen stress encoder	0.660	0.656	0.659 ± 0.170	0.623	0.620	0.628 ± 0.141
Remove learnable encoder	0.674	0.663	0.670 ± 0.217	0.654	0.645	0.663 ± 0.211
(2) Loss Ablations						
Remove sparsity loss	0.664	0.663	0.662 ± 0.108	0.666	0.667	0.676 ± 0.165
Remove orthogonality loss	0.658	0.657	0.657 ± 0.155	0.654	0.624	0.654 ± 0.179
Remove stress loss	0.650	0.647	0.642 ± 0.185	0.666	0.660	0.666 ± 0.131
Remove stress loss and orthogonality loss	0.692	0.691	0.686 ± 0.149	0.696	0.696	**0.692 ± 0.152**
(3) Embedding Variants						
Only post stress / bad memory recall	0.702	0.702	0.699 ± 0.140	0.646	0.646	0.648 ± 0.133
Only post calm / good memory recall	0.658	0.657	0.651 ± 0.150	0.674	0.660	0.675 ± 0.146
RETRACE: HR-dynamics / semantic	**0.721**	**0.720**	**0.718 ± 0.112**	**0.704**	**0.694**	**0.699 ± 0.238**

Table 5. Ablation study results under HR dynamics and semantic guidance.

Statistical significance. The proposed model consistently outperforms all baselines in terms of point estimates (ΔAcc > 0). While only a subset of comparisons reach statistical significance at the 0.05 level (e.g., MLP: $p = 0.010$), many baselines show marginal significance (e.g., IRM: $p = 0.051$, V-REx: $p = 0.066$), indicating a consistent performance trend across subjects despite variability.

7.3 Ablation study

We conduct ablation experiments to analyze the contributions of architectural components, auxiliary losses, and guidance representations. Table 5 summarizes results under both HR dynamics and semantic guidance.

Component Ablations. Removing individual components leads to consistent performance degradation, with the largest drop observed when the *adaptive fusion gate* is removed (e.g., from 0.721 to 0.619 under HR dynamics), indicating that it serves as the primary driver of performance. Other components, such as the *input gate* and *encoders*, also contribute (e.g., removing the input gate reduces accuracy from 0.721 to 0.658), but their impact is comparatively smaller. Overall, these results suggest that while architectural components are important, their effectiveness depends on the presence of meaningful guidance.

Loss Ablations. We further examine the effect of auxiliary losses. Removing the *auxiliary stress* loss decreases performance from 0.721 to 0.650, while removing the *orthogonality loss* leads to a drop to 0.658, indicating that both contribute to stable learning. Interestingly, removing both losses yields better performance than retaining only one (0.692), suggesting that these objectives interact and must be jointly optimized to effectively regularize the representation. Overall, the full model achieves the best performance, indicating that the combination of these losses improves representation quality and stability.

Embedding Variants. We analyze how different forms of resilience representation discussed in Section 6.5 contribute to performance. Under semantic guidance, using only one autobiographical recall component—positive or negative memory recall—consistently reduces performance. This suggests that the semantic guidance benefits from information distributed across both types of autobiographical narratives.

HR-dynamics guidance shows a different pattern. Using only the post-stress embedding results in a smaller drop in accuracy, from 0.721 to 0.702, whereas using only the calm-related embedding leads to a larger decrease, from 0.721 to 0.658. This is expected because the HR-dynamics proxy is intended to capture post-stress recovery,

Model	Neutral Video	Arithmetic	QA Baseline	Counting	Calm	Song Preparation	Stroop
HR dynamics Acc	0.707	0.714	0.923	0.714	0.768	0.600	0.658
Semantic Acc	0.724	0.684	1.000	0.633	0.659	0.800	0.763

Table 6. Task-level craving detection accuracy across models.

while the calm segment mainly provides a baseline reference for interpreting recovery. Together, these results suggest that HR-dynamics guidance is primarily driven by post-stress recovery dynamics, whereas semantic guidance captures a more distributed representation of subject-level context.

Summary. Overall, these results highlight that guidance is the primary factor driving performance improvements, while architectural components and auxiliary losses play supporting roles in stabilizing and refining representation learning. The full model achieves the best performance by combining subject-level guidance with structured conditioning and complementary regularization.

7.4 Model Behavior Across Task Conditions

The goal of this analysis is to examine whether ***RETRACE*** detects subjective craving rather than simply using task-induced stress as a shortcut. Because stress-inducing tasks are associated with higher craving prevalence, a model could appear effective by primarily detecting stress. We therefore evaluate performance separately across task conditions and compare aggregate performance between stress and non-stress segments.

As shown in Table 6, performance remains relatively consistent across tasks and does not systematically favor stress-inducing conditions. Aggregated results also show comparable accuracy across stress and non-stress segments: 0.678 vs. 0.743 under HR-dynamics guidance, and 0.725 vs. 0.693 under semantic guidance. If the model relied mainly on task-induced stress, we would expect substantially higher accuracy during stress tasks and degraded performance during non-stress tasks. This pattern is not observed, suggesting that the model does not simply exploit task or stress-related cues.

This result is consistent with the dataset characteristics in Section 4.2, where craving does not deterministically follow task structure. Although stress-inducing tasks show higher craving prevalence, craving also occurs during non-stress conditions and varies substantially across individuals. Together, these findings support the interpretation that ***RETRACE*** captures physiological patterns associated with participant-reported craving rather than merely detecting task-induced stress.

7.5 Effect of Guidance Quality and Specificity

The goal of this analysis is to test whether ***RETRACE*** benefits from resilience guidance because it provides meaningful subject-level context, rather than simply because it adds extra inputs or model parameters. To do so, we evaluate variants that systematically remove, weaken, or misalign the guidance signal: (1) *all-zero guidance*, which removes resilience-related information entirely; (2) *shared learnable guidance*, where a single global embedding is learned and shared across all samples to control for added parameterization; and (3) *shuffled guidance*, where resilience embeddings are randomly reassigned at test time to break the correspondence between each physiological signal and its true subject-level context.

Results in Table 7 reveal a consistent pattern. Removing guidance leads to a substantial drop in performance (Acc: 0.607), indicating that physiological signals alone are insufficient under subject-independent evaluation. Introducing a shared learnable embedding provides only marginal improvement (Acc: 0.621), suggesting that gains are not driven by additional parameters or global conditioning. In contrast, shuffled guidance yields intermediate performance (HR: 0.662, Semantic: 0.668), performing better than no guidance but worse than correctly aligned guidance. This indicates that the model benefits most when resilience guidance is correctly matched to the subject. Overall, these results support the role of resilience guidance as a meaningful subject-level conditioning signal. Correctly aligned resilience information helps ***RETRACE*** interpret physiological patterns in a subject-dependent way, rather than acting as a generic auxiliary input.

Variant	Acc	F1	Acc Mean ± Std
All zeros guidance	0.607	0.605	0.604 ± 0.123
Shared learnable embedding	0.621	0.620	0.616 ± 0.146
Shuffle (HR dyn.)	0.662	0.659	0.662 ± 0.128
Shuffle (Semantic)	0.668	0.668	0.672 ± 0.168

Table 7. Effect of different guidance strategies.

Setting	HR		Semantic	
	Acc	F1	Acc	F1
Cross visit				
Visit1 / Same-visit embedding	0.700	0.635	0.660	0.656
Visit1 / cross-visit embedding	0.700	0.670	0.681	0.679
Visit2 / Same-visit embedding	0.789	0.787	0.771	0.761
Visit2 / cross-visit embedding	0.771	0.764	0.789	0.769
Cross wrist				
Left (train & test)	0.721	0.720	0.704	0.694
Left → Right	0.679	0.679	0.714	0.711

Table 8. Cross-condition generalization under HR and semantic guidance.

7.6 Sensitivity Evaluation

We evaluate the robustness of the proposed framework under sensor and temporal distribution shifts, focusing on whether subject-level guidance can generalize across deployment conditions. Table 8 summarizes the results.

Cross-wrist Generalization. We examine robustness to sensor placement by training on left-wrist data and evaluating on right-wrist signals. Performance decreases under HR dynamics guidance (0.721 → 0.679), while semantic guidance remains stable (0.704 → 0.714). This suggests that semantic guidance provides a more invariant participant-level signal that is less sensitive to sensor-specific variations, whereas HR-based features depend more on consistent measurement conditions.

Cross-visit Generalization. We evaluate whether resilience representations collected at one-time can be reused across sessions by comparing same-visit and cross-visit embeddings. Performance remains stable across visits (e.g., 0.700 → 0.700 and 0.789 → 0.771), with no consistent degradation. This indicates that resilience embeddings capture subject-level characteristics that generalize over time, supporting their use as reusable conditioning signals.

Summary. Overall, the proposed framework demonstrates robustness to both sensor and temporal shifts. Semantic guidance shows stronger invariance to deployment-related changes, while HR dynamics guidance achieves slightly higher peak performance. Together, these results suggest that resilience-based conditioning provides a stable and reusable signal for real-world craving inference.

7.7 Empirical Justification of ***RETRACE***: Model Interpretation and Feature Attribution

To better understand how resilience-guided conditioning shapes physiological inference, we analyze feature attribution and representation weighting during inference. These analyses aim to verify whether the model adapts its interpretation of physiological signals in a subject-specific manner, as intended by the proposed design. Specifically, we use Integrated Gradients to estimate feature-level attribution and directly extract fusion gate weights from the forward pass to analyze how different representations are weighted during inference. To examine how this modulation varies across individuals, we aggregate attribution scores and gate weights at the subject level and group participants into high- and low-resilience based on their post-stress HR-AUC recovery, as defined in Appendix H. This grouping allows us to compare how the model behaves under different levels of resilience and to assess whether resilience systematically influences the interpretation of physiological signals.

Feature Attribution under HR dynamics Guidance. Figure 8 presents feature attribution results under HR dynamics guidance, comparing high- and low-resilience groups. Notably, only a small subset of features exhibits consistently high importance, primarily corresponding to AUC statistics of heart rate changes, including signed AUC, absolute AUC, and their positive and negative components. Furthermore, AUC features derived from post-stress segments generally show higher importance than those derived from post-calm segments. This pattern suggests that recovery dynamics following stress play a more prominent role in the model's inference process. In

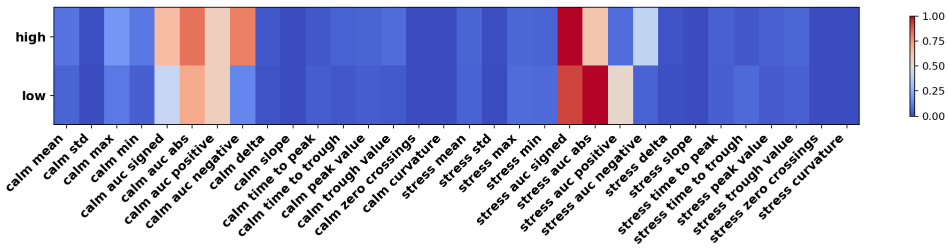

Fig. 8. Feature attribution under HR dynamics guidance for high- and low-resilience groups. AUC-based features (signed, absolute, positive, and negative) dominate attribution, while other temporal features show low importance.

particular, HR-based AUC features of post-stress recovery emerge as the most influential features, indicating that they capture key aspects of resilience-related variation in physiological responses. This interpretation is consistent with prior work that conceptualizes resilience as a recovery process and operationalizes it using AUC-based measures [32].

Fusion Gate Behavior. Figure 9 shows the fusion gate weights assigned to the learnable physiological encoder and the frozen stress encoder under both HR dynamics and semantic guidance. We observe a consistent pattern across guidance types: low-resilience individuals receive higher weights on the learnable encoder and lower weights on the frozen stress encoder. This behavior provides direct evidence that the model adapts its representation based on subject-level resilience, as intended by the conditioning mechanism. The frozen encoder captures shared stress-related structure, while the learnable encoder captures craving-specific deviations. For low-resilience individuals, the model relies more heavily on the learnable encoder, indicating greater sensitivity to craving-related physiological patterns. In contrast, for high-resilience individuals, physiological responses are more likely to reflect stress without escalating into craving, resulting in relatively greater reliance on the stress encoder. Importantly, the similarity of this pattern across both HR dynamics and semantic guidance suggests that different forms of participant-level guidance lead to consistent modulation of physiological interpretation. This indicates that both guidance mechanisms capture aligned representations of subject-specific resilience, despite being derived from distinct sources.

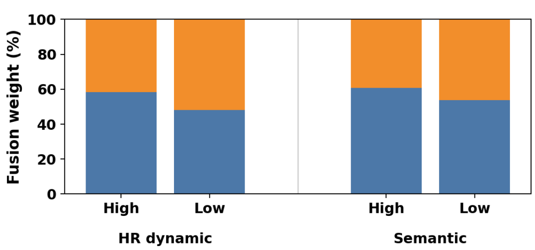

Fig. 9. Fusion gate weights under different guidance types. Blue denotes the frozen stress encoder, and orange denotes the learnable encoder.

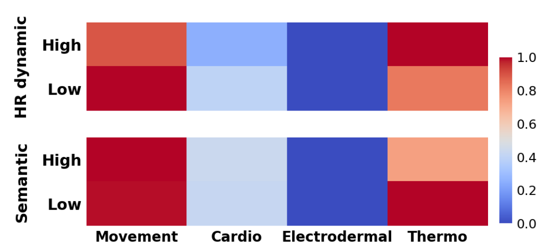

Fig. 10. Modality-level attribution under different guidance types.

Modality-Level Attribution. Figures 10 show modality-level attribution under HR dynamics and semantic guidance, respectively. Across both guidance types, the model assigns higher importance to thermoregulation and movement signals, while electrodermal and cardiovascular features receive comparatively lower weights.

This pattern is consistent with our statistical analysis in Section 5.1, which shows that thermoregulation features exhibit significant differentiation with respect to craving.

This finding is notable because electrodermal and cardiovascular signals are traditionally associated with stress detection, with electrodermal activity (EDA) widely regarded as a gold-standard biomarker for stress [136]. The reduced attribution of these canonical stress-related modalities suggests that the model is not simply relying on stress-related biomarkers, but instead learns from alternative physiological channels that better capture craving-related dynamics.

Summary. Overall, these analyses provide evidence that the model operates in a resilience-conditioned manner. Feature attribution shows that inference is driven by recovery-related dynamics rather than dominant stress responses. Fusion gate behavior demonstrates that subject-level guidance modulates the balance between shared and adaptive representations. Finally, the consistency between HR dynamics and semantic guidance indicates that both representations capture aligned notions of resilience.

Together, these findings support the central hypothesis that craving detection requires subject-specific interpretation of physiological signals, rather than a fixed mapping across individuals.

8 In-the-Wild Feasibility Study

To examine whether ***RETRACE*** can transfer beyond the controlled laboratory setting, we conducted a small in-the-wild feasibility study using real-world craving reports. This analysis is not intended as a definitive validation of real-world performance. Rather, it tests whether a model trained only on laboratory data can produce meaningful craving predictions when applied to naturally occurring daily-living wearable data from a held-out participant.

Data Collection. We collected in-the-wild data from one participant who also completed the laboratory study. Over a 14-day period, the participant completed ecological momentary assessments (EMA) three times per day, reporting craving levels over the preceding two-hour interval. This resulted in 32 EMA instances (26 non-craving, 6 craving) and 15,148 physiological windows. To ensure strict separation between training and evaluation, the model was trained on the Lab dataset, *excluding this participant,* and evaluated only on this participant's in-the-wild data.

Evaluation Protocol. ***RETRACE*** produces predictions at the physiological-window level. To align these predictions with EMA annotations, we aggregated all window-level predictions within the two-hour interval preceding each EMA report. Binary EMA-level predictions were obtained using majority voting across the corresponding physiological windows. This protocol preserves window-level inference while enabling evaluation against participant-reported craving over the EMA recall interval.

Results. Because the in-the-wild labels are highly imbalanced, with 26 non-craving and 6 craving instances, we report balanced accuracy (BAcc) and AUC rather than standard accuracy alone. As shown in Table 9, ***RETRACE*** achieves above-chance performance under both guidance types. Semantic guidance obtains the highest balanced accuracy (0.686), suggesting stronger binary classification performance, while HR-dynamics guidance obtains the highest AUC (0.679), suggesting a more stable ranking of craving likelihood across EMA instances.

Overall, this feasibility study provides initial evidence that resilience-guided craving inference can transfer from laboratory training to naturalistic wearable sensing. Because it includes only one participant and a few craving EMA reports, the findings are preliminary and require validation through larger longitudinal studies across participants, contexts, adherence patterns, and naturally occurring craving episodes.

Model	HR dynamics Guidance		Semantic Guidance	
	BAcc	AUC	BAcc	AUC
Cross-Attn	0.635	0.635	0.500	0.599
MoE	0.635	0.635	0.583	0.583
FiLM	0.625	0.448	0.562	0.410
RETRACE	**0.660**	**0.679**	**0.686**	0.564

Table 9. In-the-wild evaluation results under different guidance types. Balanced accuracy (BAcc) is reported as the primary metric due to class imbalance, with AUC provided for reference.

9 ***RETRACE***'s Deployment Considerations for Practical Use

RETRACE is designed for low-burden wearable craving sensing. At inference time, it requires only a short physiological window from a wearable device and a one-time collected subject-level resilience guidance representation. It does not require craving labels from the target user or user-specific model retraining, making it suitable for practical settings where continuous physiology can be collected passively and subject-level guidance can be obtained during onboarding.

Two deployment pathways. The two guidance variants offer different trade-offs between efficacy, usability, and deployment burden. The HR-dynamics variant uses post-stress heart-rate recovery as a physiology-grounded resilience proxy [32]. It may be most appropriate in supervised settings, such as clinical intake, rehabilitation programs, or structured onboarding, where a brief calm baseline and stress/recovery protocol can be administered. This guidance is closely tied to physiological recovery and may support stable probabilistic ranking of craving likelihood over time, but it requires controlled measurement conditions.

The semantic variant uses autobiographical memory recall as a lower-burden source of subject-level context. A user provides short descriptions of recent positive and negative personal experiences, which are transcribed and converted into a compact embedding for later inference. This approach does not require stress induction or physiological recovery measurement, making it easier to deploy remotely or in unsupervised settings. However, autobiographical narratives are privacy-sensitive [57]. In practical deployment, raw narratives should not be stored or reused after embedding generation; instead, only compact embeddings should be retained with appropriate encryption, access control, and deletion options. This reduces, but does not fully eliminate, privacy risks because embeddings may still encode sensitive information [114]. Thus, HR-dynamics guidance favors physiological specificity, while semantic guidance favors usability and scalability.

Use in wearable recovery support systems. In both deployment paths, ***RETRACE*** can operate on continuous wearable sensor streams and estimate craving risk without repeatedly prompting users for self-reports. This is important because mHealth systems that rely on sparse ecological momentary assessments may miss brief or acute periods of vulnerability [62]. By enabling continuous, passive, and subject-aware craving inference from wearable physiology, ***RETRACE*** could support timely and personalized digital support when physiological patterns indicate elevated craving risk [45, 52, 92].

10 Inference-time Efficiency and Runtime Analysis

To assess whether ***RETRACE*** can support real-time wearable and mobile health deployment, we benchmarked inference latency across ten heterogeneous computing platforms. Local inference is important for practical mHealth systems because it can reduce dependence on cloud computation, improve availability in low-connectivity settings, and better protect user privacy [71, 86, 139].

We evaluated both guidance variants: HR dynamics and semantic guidance. For the semantic variant, language representations are pre-computed during onboarding and frozen before deployment. Therefore, inference does not require running a large language model or text encoder on-device. For both variants, runtime computation consists only of processing the physiological feature window, combining it with the static subject-level guidance representation, and producing the craving prediction.

Because wearable sensing operates on incoming windows sequentially, we measured forward-pass latency with batch size $B = 1$, reflecting real-time processing of a single physiological window [21, 54, 106]. Table 10 reports three deployment-relevant metrics: mean latency, 99th percentile (P99) latency to capture worst-case runtime variation, and Resident Set Size (RSS) to measure peak memory footprint. Despite the HR dynamics guidance model containing more parameters (899,828) than the semantic-guidance model (469,236), both remain exceptionally lightweight. On the Apple M4 CPU, forward inference completes in 0.132 ms for semantic guidance and 0.154 ms for HR dynamics guidance. While the Apple MPS backend is slower at $B = 1$—an expected outcome for small, unbatched models where kernel launch and device transfer overheads dominate—larger CUDA devices like the RTX 4090 and NVIDIA GB10 maintain highly stable, sub-millisecond forward latency.

Crucially, results on consumer and edge devices further support practical deployment. On a Google Pixel 6, inference completes in 14.730 ms for semantic guidance and 16.708 ms for HR dynamics guidance, with RSS below 250 MB for both variants. On a Raspberry Pi, inference takes 1.616 ms and 2.654 ms for semantic and HR dynamics guidance, respectively. Even on the slower Jetson Nano, the model remains responsive, with mean latency of 14.531 ms for the HR dynamics branch. These results indicate that ***RETRACE*** can run locally on common mobile and edge devices within the timing constraints of wearable sensing pipelines [22, 42, 72].

(a) Semantic guidance branch

Device Class	Hardware Specification	Inference		
		Mean (ms)	P99 (ms)	RSS (MB)
Server/Workstation	AMD Ryzen 9 7950X (CPU)	0.294	0.369	405.7
	NVIDIA RTX 4090 (GPU)	0.315	0.330	787.4
	Intel CPU (20 logical cores)	1.721	2.356	436.9
Mobile/Consumer	Apple M4 (CPU)	0.132	0.204	206.7
	Apple Silicon (MPS)	1.226	1.456	365.6
	Google Pixel 6	14.730	21.588	191.5
Edge/IoT	NVIDIA GB10 platform (CPU)	5.741	18.901	507.3
	NVIDIA GB10 (CUDA)	0.329	0.342	987.4
	Raspberry Pi (aarch64 CPU)	1.616	5.648	209.8
	NVIDIA Jetson Nano (CPU)	9.618	16.254	268.9

(b) HR dynamics guidance branch

Device Class	Hardware Specification	Inference		
		Mean (ms)	P99 (ms)	RSS (MB)
Server/Workstation	AMD Ryzen 9 7950X (CPU)	0.339	0.345	533.1
	NVIDIA RTX 4090 (GPU)	0.337	0.352	971.7
	Intel CPU (20 logical cores)	2.181	2.752	499.3
Mobile/Consumer	Apple M4 (CPU)	0.154	0.203	307.1
	Apple Silicon (MPS)	1.378	1.604	468.1
	Google Pixel 6	16.708	27.515	241.5
Edge/IoT	NVIDIA GB10 platform (CPU)	0.981	1.440	628.7
	NVIDIA GB10 (CUDA)	0.376	0.386	1241.9
	Raspberry Pi (aarch64 CPU)	2.654	2.964	272.8
	NVIDIA Jetson Nano (CPU)	14.531	31.019	277.9

Table 10. Training and forward-pass runtime across hardware platforms for batch size $B = 1$. Mean and P99 latency are reported in milliseconds. RSS memory denotes resident set size during execution. The Google Pixel 6 row is left blank because evaluation is ongoing.

11 Discussion and limitation

This section discusses the main implications of ***RETRACE*** for wearable craving sensing and outlines the limitations that define the scope of the current study.

11.1 Key Implications

- **Trait-conditioned interpretation without per-user training. *RETRACE*** enables passive personalization by using resilience-related context to guide physiological inference, without per-user retraining or labeled craving data. This supports practical wearable deployment where user-specific craving labels are difficult to collect.
- **Stress as a physiological reference for craving.** Stress and craving share overlapping physiological responses, but stress signals are stronger and more consistent, whereas craving signals are weaker and more heterogeneous. ***RETRACE*** uses a frozen stress-pretrained encoder as a physiological reference, while learning craving as a separate resilience-conditioned task. This allows the model to interpret craving-related deviations within broader stress-related physiology without treating stress itself as craving.

- **Trade-offs between physiological and semantic guidance.** The two guidance options offer different deployment advantages. HR-dynamics guidance is closely tied to post-stress autonomic recovery and may be preferable in clinical or supervised settings where standardized stress recovery protocols are feasible. Semantic guidance from autobiographical recall is easier to collect, can be reused over time, and may be more suitable for remote or in-the-wild deployment. The strong performance of both variants suggests that the benefit comes from trait-conditioned interpretation rather than from a single guidance modality.
- **Ethical and privacy considerations.** Autobiographical narratives may contain sensitive personal information, so deployment requires careful privacy protection and user control. In ***RETRACE***, raw narratives are not needed at inference time; only compact subject-level embeddings are used. This can reduce exposure of sensitive information, especially if embeddings are computed locally, stored securely, and made deletable by the user. However, embeddings may still encode private information, so misuse, stigmatization, and loss of user agency remain important concerns.

This study also has important limitations:

- **Dataset and generalization.** Our evaluation uses one controlled laboratory dataset and a 14-day in-the-wild feasibility study with one participant. While collecting reliable craving annotations from clinical populations is challenging and comparable sample sizes are common in controlled IMWUT/CHI physiological studies [16, 37, 50, 51, 141], broader validation is needed to establish clinical deployability. Future work should evaluate ***RETRACE*** on larger, more diverse, longitudinal real-world datasets to test generalization across populations, environments, and daily-life craving patterns.
- **Proxy-based representation of resilience.** Both HR dynamics and autobiographical language are proxies for resilience-related individual differences, not direct measures of psychological resilience. HR dynamics depends on structured stress recovery measurement, while autobiographical recall may not fully capture the complexity of resilience-related traits. Future work should compare these proxies with validated resilience scales, clinical outcomes, and more structured elicitation methods, such as prompt-guided narratives or multimodal assessments.

12 Conclusion

This work establishes wearable craving detection in OUD as a problem of subject-aware physiological interpretation, rather than only short-window state classification. Our findings show that craving is physiologically measurable but difficult to isolate because it is weak, heterogeneous, and intertwined with stress-related autonomic activity. They also show that resilience is not a momentary wearable signal, but a subject-level factor that helps explain why similar physiological patterns may carry different meanings across individuals. By using resilience-related guidance to condition physiological interpretation, ***RETRACE*** enables subject-aware craving estimation without requiring per-user craving labels or retraining. More broadly, this work shows that wearable health systems for heterogeneous and clinically sensitive populations can benefit from treating personal context not only as additional input, but as a mechanism for interpreting ambiguous physiological evidence. While further validation in larger, more diverse, and longer-term real-world OUD cohorts is needed, ***RETRACE*** provides a practical step toward scalable, subject-aware wearable sensing to support OUD recovery.

References

[1] Hervé Abdi. 2010. Partial least squares regression and projection on latent structure regression (PLS Regression). *Wiley interdisciplinary reviews: computational statistics* 2, 1 (2010), 97–106.

[2] Hervé Abdi and Lynne J Williams. 2010. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics* 2, 4 (2010), 433–459.

[3] Jonathan M Adler. 2012. Living into the story: agency and coherence in a longitudinal study of narrative identity development and mental health over the course of psychotherapy. *Journal of personality and social psychology* 102, 2 (2012), 367.

[4] Jonathan M Adler, Jennifer Lodi-Smith, Frederick L Philippe, and Iliane Houle. 2016. The incremental validity of narrative identity in predicting well-being: A review of the field and recommendations for the future. *Personality and Social Psychology Review* 20, 2 (2016), 142–175.

[5] Amelia Aldao, Susan Nolen-Hoeksema, and Susanne Schweizer. 2010. Emotion-regulation strategies across psychopathology: A meta-analytic review. *Clinical psychology review* 30, 2 (2010), 217–237.

[6] Michael T Allen and Michael D Crowell. 1989. Patterns of autonomic response during laboratory stressors. *Psychophysiology* 26, 5 (1989), 603–614.

[7] Mohammed A Almazrouei, Ruth M Morgan, and Itiel E Dror. 2023. A method to induce stress in human subjects in online research environments. *Behavior research methods* 55, 5 (2023), 2575–2582.

[8] Emily Anderson and Lindsay McNair. 2018. Ethical issues in research involving participants with opioid use disorder. *Therapeutic innovation & regulatory science* 52, 3 (2018), 280–284.

[9] Andrea Apicella, Pasquale Arpaia, Giovanni D'Errico, Davide Marocco, Giovanna Mastrati, Nicola Moccaldi, and Roberto Prevete. 2024. Toward cross-subject and cross-session generalization in EEG-based emotion recognition: Systematic review, taxonomy, and methods. *Neurocomputing* 604 (2024), 128354.

[10] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2019. Invariant risk minimization. *arXiv preprint arXiv:1907.02893* (2019).

[11] Emmanuelle Baillet, Marc Auriacombe, Cassandre Romao, Hélène Garnier, Christophe Gauld, Chloé Vacher, Joël Swendsen, Mélina Fatseas, and Fuschia Serre. 2024. Craving changes in first 14 days of addiction treatment: an outcome predictor of 5 years substance use status? *Translational Psychiatry* 14, 1 (2024), 497.

[12] Cauã Ferreira Barros, Bruna Borges Azevedo, Valdemar Vicente Graciano Neto, Mohamad Kassab, Marcos Kalinowski, Hugo Alexandre D Do Nascimento, and Michelle CGSP Bandeira. 2025. Large language model for qualitative research: A systematic mapping study. In *2025 IEEE/ACM International Workshop on Methodological Issues with Empirical Studies in Software Engineering (WSESE)*. IEEE, 48–55.

[13] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)* 57, 1 (1995), 289–300.

[14] George R Breese, Kathleen Chu, Christopher V Dayas, Douglas Funk, Darin J Knapp, George F Koob, Dzung Anh Lê, Laura E O'Dell, David H Overstreet, Amanda J Roberts, et al. 2005. Stress enhancement of craving during sobriety: a risk for relapse. *Alcoholism: Clinical and Experimental Research* 29, 2 (2005), 185–195.

[15] Kelly Caine. 2016. Local standards for sample size at CHI. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 981–992.

[16] Stephanie Carreiro, Keerthi Kumar Chintha, Sloke Shrestha, Brittany Chapman, David Smelson, and Premananda Indic. 2020. Wearable sensor-based detection of stress and craving in patients during treatment for substance use disorder: A mixed methods pilot study. *Drug and alcohol dependence* 209 (2020), 107929.

[17] Stephanie Carreiro, Pravitha Ramanand, Melissa Taylor, Rebecca Leach, Joshua Stapp, Sloke Sherestha, David Smelson, and Premananda Indic. 2024. Evaluation of a digital tool for detecting stress and craving in SUD recovery: An observational trial of accuracy and engagement. *Drug and Alcohol Dependence* 261 (2024), 111353.

[18] Jaesung Cha. 1994. Partial least squares. *Adv. Methods Mark. Res* 407 (1994), 52–78.

[19] Brittany P Chapman, Bhanu Teja Gullapalli, Tauhidur Rahman, David Smelson, Edward W Boyer, and Stephanie Carreiro. 2022. Impact of individual and treatment characteristics on wearable sensor-based digital biomarkers of opioid use. *NPJ Digital Medicine* 5, 1 (2022), 123.

[20] Chariklia Chatzaki and Manolis Tsiknakis. 2025. An overview of stress analysis based on physiological signals: Systematic review of open datasets and current trends. *Sensors* 25, 23 (2025), 7108.

[21] Erh-Chung Chen, Pin-Yu Chen, I Chung, Che-Rung Lee, et al. 2024. Overload: Latency attacks on object detection for edge devices. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 24716–24725.

[22] Jiasi Chen and Xukan Ran. 2019. Deep learning with edge computing: A review. *Proc. IEEE* 107, 8 (2019), 1655–1674.

[23] Valentine Chirokoff, Maud Dupuy, Majd Abdallah, Melina Fatseas, Fuschia Serre, Marc Auriacombe, David Misdrahi, Sylvie Berthoz, Joel Swendsen, Edith V Sullivan, et al. 2023. Craving dynamics and related cerebral substrates predict timing of use in alcohol, tobacco, and cannabis use disorders. *Addiction neuroscience* 9 (2023), 100138.

[24] Shefali Arora Chouhan. 2026. Supervised information gain-based feature selection for multimodal physiological signals in stress prediction. *Scientific Reports* (2026).

[25] Desirée Colombo, Silvia Serino, Carlos Suso-Ribera, Javier Fernández-Álvarez, Pietro Cipresso, Azucena García-Palacios, Giuseppe Riva, and Cristina Botella. 2021. The moderating role of emotion regulation in the recall of negative autobiographical memories. *International Journal of Environmental Research and Public Health* 18, 13 (2021), 7122.

[26] Bruce E Compas, Sarah S Jaser, Alexandra H Bettis, Kelly H Watson, Meredith A Gruhn, Jennifer P Dunbar, Ellen Williams, and Jennifer C Thigpen. 2017. Coping, emotion regulation, and psychopathology in childhood and adolescence: A meta-analysis and narrative review. *Psychological bulletin* 143, 9 (2017), 939.

[27] Samantha L Connolly and Lauren B Alloy. 2018. Negative event recall as a vulnerability for depression: Relationship between momentary stress-reactive rumination and memory for daily life stress. *Clinical Psychological Science* 6, 1 (2018), 32–47.

[28] Martin A Conway and Christopher W Pleydell-Pearce. 2000. The construction of autobiographical memories in the self-memory system. *Psychological review* 107, 2 (2000), 261.

[29] Sheila E Crowell, Chloe R Skidmore, Holly K Rau, and Paula G Williams. 2013. Psychosocial stress, emotion regulation, and resilience in adolescence. In *Handbook of adolescent health psychology*. Springer, 129–141.

[30] Reza Dastranj and Martin Kolar. 2024. Forecasting Mortality Rates: Unveiling Patterns with a PCA-GEE Approach. *arXiv preprint arXiv:2407.01547* (2024).

[31] Akbar Dehghani, Omid Sarbishei, Tristan Glatard, and Emad Shihab. 2019. A quantitative comparison of overlapping and non-overlapping sliding windows for human activity recognition using inertial sensors. *Sensors* 19, 22 (2019), 5026.

[32] Ruud den Hartigh and Yannick Hill. 2022. Conceptualizing and measuring psychological resilience: What can we learn from physics? *New Ideas in Psychology* 66 (2022), 100934.

[33] Edgar Dobriban. 2020. Permutation methods for factor analysis and PCA. (2020).

[34] Khaled Draghmeh, Mark S Gold, and Brian Fuehrlein. 2025. A Review of Relapse in Opioid Use Disorder. *Current Addiction Reports* 12, 1 (2025), 58.

[35] Haakon G Engen and Michael C Anderson. 2018. Memory control: A fundamental mechanism of emotion regulation. *Trends in Cognitive Sciences* 22, 11 (2018), 982–995.

[36] Matthew C Enkema, Kevin A Hallgren, Elizabeth C Neilson, Sarah Bowen, Elizabeth R Bird, and Mary E Larimer. 2020. Disrupting the path to craving: Acting without awareness mediates the link between negative affect and craving. *Psychology of Addictive Behaviors* 34, 5 (2020), 620.

[37] Talia Sofia Ezer, Jonathan Giron, Hadas Erel, and Oren Zuckerman. 2024. Somaesthetic Meditation Wearable: Exploring the Effect of Targeted Warmth Technology on Meditators' Experiences. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–14.

[38] Ronald Fischer and Diana Boer. 2015. Motivational basis of personality traits: A meta-analysis of value-personality correlations. *Journal of personality* 83, 5 (2015), 491–510.

[39] Simon Föll, Martin Maritsch, Federica Spinola, Varun Mishra, Filipe Barata, Tobias Kowatsch, Elgar Fleisch, and Felix Wortmann. 2021. FLIRT: A feature generation toolkit for wearable data. *Computer Methods and Programs in Biomedicine* 212 (2021), 106461.

[40] Helen C Fox, Kwang-Ik A Hong, Kristen Siedlarz, and Rajita Sinha. 2008. Enhanced sensitivity to stress and drug/alcohol craving in abstinent cocaine-dependent individuals compared to social drinkers. *Neuropsychopharmacology* 33, 4 (2008), 796–805.

[41] Barbara L Fredrickson. 2001. The role of positive emotions in positive psychology: The broaden-and-build theory of positive emotions. *American psychologist* 56, 3 (2001), 218.

[42] Chang Gao, Antonio Rios-Navarro, Xi Chen, Shih-Chii Liu, and Tobi Delbruck. 2020. EdgeDRNN: Recurrent neural network accelerator for edge inference. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems* 10, 4 (2020), 419–432.

[43] Elvira García-Bajos and Malen Migueles. 2013. An integrative study of autobiographical memory for positive and negative experiences. *The Spanish Journal of Psychology* 16 (2013), E102.

[44] Eric L Garland, Bhanu T Gullapalli, Kort C Prince, Adam W Hanley, Mathias Sanyer, Mark Tuomenoksa, and Tauhidur Rahman. 2023. Zoom-based mindfulness-oriented recovery enhancement plus just-in-time mindfulness practice triggered by wearable sensors for opioid craving and chronic pain. *Mindfulness* 14, 6 (2023), 1329–1345.

[45] Eric L. Garland, Bhanu T. Gullapalli, Kort C. Prince, Adam W. Hanley, Mathias Sanyer, Mark Tuomenoksa, and Tauhidur Rahman. 2023. Zoom-Based Mindfulness-Oriented Recovery Enhancement Plus Just-in-Time Mindfulness Practice Triggered by Wearable Sensors for Opioid Craving and Chronic Pain. *Mindfulness* 14 (2023), 1329–1345.

[46] Charles Golden, Shawna M Freshwater, and Zarabeth Golden. 1978. Stroop color and word test. (1978).

[47] David S Goldstein. 2013. Differential responses of components of the autonomic nervous system. *Handbook of clinical neurology* 117 (2013), 13–22.

[48] Sergej Grunevski, Julia Kong, and Anna Konova. 2024. Predictive Relationship Between Different Timescales of Opioid Craving and Opioid Use. *Drug and Alcohol Dependence* 260 (2024), 109974.

[49] Sabine Miriam Grüsser, Chantal Patricia Mörsen, Klaus Wölfling, and Herta Flor. 2006. The relationship of stress, coping, effect expectancies and craving. *European Addiction Research* 13, 1 (2006), 31–38.

[50] Bhanu Teja Gullapalli, Stephanie Carreiro, Brittany P Chapman, Deepak Ganesan, Jan Sjoquist, and Tauhidur Rahman. 2021. Opitrack: a wearable-based clinical opioid use tracker with temporal convolutional attention networks. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 5, 3 (2021), 1–29.

[51] Bhanu Teja Gullapalli, Annamalai Natarajan, Gustavo A Angarita, Robert T Malison, Deepak Ganesan, and Tauhidur Rahman. 2019. On-body sensing of cocaine craving, euphoria and drug-seeking behavior using cardiac and respiratory signals. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 2 (2019), 1–31.

[52] Jazmin Hampton, Reynalde Eugene, Nirzari Kapadia, Emily Caggiano, Amanda Geagea, C. James Watson, and Stephanie Carreiro. 2025. Digital detection of craving and stress for individuals in recovery from substance use disorder: A qualitative study. *Drug and Alcohol Dependence Reports* 15 (2025), 100336.

[53] Kim A Hoffman, Javier Ponce Terashima, and Dennis McCarty. 2019. Opioid use disorder and treatment: challenges and opportunities. *BMC health services research* 19, 1 (2019), 884.

[54] Runxi Huang, Mingxuan Yu, Mingyu Tsoi, and Xiaomin Ouyang. 2025. MMEdge: Accelerating On-device Multimodal Inference via Pipelined Sensing and Encoding. *arXiv preprint arXiv:2510.25327* (2025).

[55] Francis KC Hui, Samuel Muller, and Alan H Welsh. 2025. Adjusted predictions for generalized estimating equations. *Biometrics* 81, 3 (2025), ujaf090.

[56] Talha Iqbal, Andrew J Simpkin, Davood Roshan, Nicola Glynn, John Killilea, Jane Walsh, Gerard Molloy, Sandra Ganly, Hannah Ryman, Eileen Coen, et al. 2022. Stress monitoring using wearable sensors: a pilot study and stress-predict dataset. *Sensors* 22, 21 (2022), 8135.

[57] Avnish Singh Jat and Tor-Morten Grønli. 2023. Harnessing the digital revolution: a comprehensive review of mHealth applications for remote monitoring in transforming healthcare delivery. In *International Conference on Mobile Web and Intelligent Information Systems*. Springer, 55–67.

[58] Sarah Jenner, Dimitris Raidos, Emma Anderson, Stella Fleetwood, Ben Ainsworth, Kerry Fox, Jana Kreppner, and Mary Barker. 2025. Using large language models for narrative analysis: a novel application of generative AI. *Methods in Psychology* 12 (2025), 100183.

[59] Ashani Jeyadevan and Jodie Grigg. 2024. A Time-Limited Scoping Review Investigating the Use of Wearable Biosensors in the Substance Use Field. *Current Addiction Reports* 11, 5 (2024), 928–939.

[60] P Karthikeyan, M Murugappan, and S Yaacob. 2014. Analysis of stroop color word test-based human stress detection using electrocardiography and heart rate variability signals. *Arabian Journal for Science and Engineering* 39, 3 (2014), 1835–1847.

[61] Yuta Katsumi and Sanda Dolcos. 2020. Suppress to feel and remember less: Neural correlates of explicit and implicit emotional suppression on perception and memory. *Neuropsychologia* 145 (2020), 106683.

[62] Zachary King, Zoe Setiadi, Liana Hamdan, Hajar Ahmed, Bishal Lamichhane, Ashutosh Sabharwal, Ramiro Salas, Nidal Moukaddam, and Akane Sano. 2025. Predicting Craving-Related Emotions among Opioid Use Disorder Patients: Preliminary Results. In *2025 IEEE 21st International Conference on Body Sensor Networks (BSN)*. IEEE.

[63] Zachary D King, Judith Moskowitz, Begum Egilmez, Shibo Zhang, Lida Zhang, Michael Bass, John Rogers, Roozbeh Ghaffari, Laurie Wakschlag, and Nabil Alshurafa. 2019. Micro-stress EMA: A passive sensing framework for detecting in-the-wild stress in pregnant mothers. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 3, 3 (2019), 1–22.

[64] Clemens Kirschbaum, Karl-Martin Pirke, and Dirk H Hellhammer. 1993. The 'Trier Social Stress Test'–a tool for investigating psychobiological stress responses in a laboratory setting. *Neuropsychobiology* 28, 1-2 (1993), 76–81.

[65] George Koob and Mary Jeanne Kreek. 2007. Stress, dysregulation of drug reward pathways, and the transition to drug dependence. *American journal of psychiatry* 164, 8 (2007), 1149–1159.

[66] George F Koob. 2013. Addiction is a reward deficit and stress surfeit disorder. *Frontiers in psychiatry* 4 (2013), 72.

[67] George F Koob, Cara L Buck, Ami Cohen, Scott Edwards, Paula E Park, Joel E Schlosburg, Brooke Schmeichel, Leandro F Vendruscolo, Carrie L Wade, Timothy W Whitfield Jr, et al. 2014. Addiction as a stress surfeit disorder. *Neuropharmacology* 76 (2014), 370–382.

[68] George F Koob and Michel Le Moal. 2008. Addiction and the brain antireward system. *Annu. Rev. Psychol.* 59, 1 (2008), 29–53.

[69] Anjali Krishnan, Lynne J Williams, Anthony Randal McIntosh, and Hervé Abdi. 2011. Partial Least Squares (PLS) methods for neuroimaging: a tutorial and review. *Neuroimage* 56, 2 (2011), 455–475.

[70] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. 2021. Out-of-distribution generalization via risk extrapolation (rex). In *International conference on machine learning*. PMLR, 5815–5826.

[71] Nicholas D Lane, Emiliano Miluzzo, Hong Lu, Daniel Peebles, Tanzeem Choudhury, and Andrew T Campbell. 2010. A survey of mobile phone sensing. *IEEE Communications magazine* 48, 9 (2010), 140–150.

[72] En Li, Zhi Zhou, and Xu Chen. 2018. Edge intelligence: On-demand deep learning model co-inference with device-edge synergy. In *Proceedings of the 2018 workshop on mobile edge communications*. 31–36.

[73] Fang Li and Dan Zhang. 2025. Multimodal physiological signals from wearable sensors for affective computing: A systematic review. *Intelligent Sports and Health* 1, 4 (2025), 210–222.

[74] Kung-Yee Liang and Scott L Zeger. 1986. Longitudinal data analysis using generalized linear models. *Biometrika* 73, 1 (1986), 13–22.

[75] R Ross MacLean, Jessica L Armstrong, and Mehmet Sofuoglu. 2019. Stress and opioid use disorder: A systematic review. *Addictive behaviors* 98 (2019), 106010.

[76] Dominique Makowski, Tam Pham, Zen J Lau, Jan C Brammer, François Lespinasse, Hung Pham, Christopher Schölzel, and SH Annabel Chen. 2021. NeuroKit2: A Python toolbox for neurophysiological signal processing. *Behavior research methods* (2021), 1–8.

[77] Dan P McAdams. 2001. The psychology of life stories. *Review of general psychology* 5, 2 (2001), 100–122.

[78] Cameron McCarthy, Nikhilesh Pradhan, Calum Redpath, and Andy Adler. 2016. Validation of the Empatica E4 wristband. In *2016 IEEE EMBS international student conference (ISC)*. IEEE, 1–4.

[79] Bruce S McEwen and Huda Akil. 2020. Revisiting the stress concept: implications for affective disorders. *Journal of Neuroscience* 40, 1 (2020), 12–21.
[80] Riley Q McNaboe, Youngsun Kong, Wendy A Henderson, Xiaomei Cong, Aolan Li, Min-Hee Seo, Ming-Hui Chen, Bin Feng, and Hugo F Posada-Quintero. 2025. Optimizing Sensor Locations for Electrodermal Activity Monitoring Using a Wearable Belt System. *Journal of sensor and actuator networks* 14, 2 (2025), 31.
[81] Lakmal Meegahapola, Marios Constantinides, Zoran Radivojevic, Hongwei Li, Michael Eggleston, and Daniele Quercia. 2026. Stress Mindset Matters: Rethinking Mental Stress Detection with Multimodal Wearable Sensors. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. 1–26.
[82] Luca Menghini, Evelyn Gianfranchi, Nicola Cellini, Elisabetta Patron, Mariaelena Tagliabue, and Michela Sarlo. 2019. Stressing the accuracy: Wrist-worn wearable sensor validation over different conditions. *Psychophysiology* 56, 11 (2019), e13441.
[83] Ethan H Mereish and Robert Miranda Jr. 2024. Vicarious heterosexism-based stress induces alcohol, nicotine, and cannabis craving and negative affect among sexual minority young adults: An experimental study. *Neurobiology of Stress* 32 (2024), 100668.
[84] Sungwoo Moon, Sun Sook Kim, and Byungjoo Choi. 2024. Comparative feasibility study of physiological signals from wristband-type wearable sensors to assess occupants' thermal comfort. *Energy and Buildings* 308 (2024), 114032.
[85] Vivian Genaro Motti and Kelly Caine. 2015. Users' privacy concerns about wearables: impact of form factor, sensors and type of data collected. In *International conference on financial cryptography and data security*. Springer, 231–244.
[86] Inbal Nahum-Shani, Shawna N Smith, Bonnie J Spring, Linda M Collins, Katie Witkiewitz, Ambuj Tewari, and Susan A Murphy. 2016. Just-in-time adaptive interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support. *Annals of behavioral medicine* (2016), 1–17.
[87] Simon Ollander, Christelle Godin, Aurélie Campagne, and Sylvie Charbonnier. 2016. A comparison of wearable and stationary sensors for stress detection. In *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 004362–004366.
[88] Debora Pascuzzi and Andrea Smorti. 2017. Emotion regulation, autobiographical memories and life narratives. *New Ideas in Psychology* 45 (2017), 28–37.
[89] Bilge Pehlivanoğlu, Nezih Durmazlar, and Dicle Balkancı. 2005. Computer adapted Stroop colour-word conflict test as a laboratory stress model. *Journal of Clinical Practice and Research* 27, 2 (2005), 58.
[90] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. 2018. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.
[91] Olga Perski, Emily T Hébert, Felix Naughton, Eric B Hekler, Jamie Brown, and Michael S Businelle. 2022. Technology-mediated just-in-time adaptive interventions (JITAIs) to reduce harmful substance use: a systematic review. *Addiction* 117, 5 (2022), 1220–1241.
[92] Olga Perski, Emily T. Hébert, Felix Naughton, Eric B. Hekler, Jamie Brown, and Michael S. Businelle. 2022. Technology-mediated just-in-time adaptive interventions (JITAIs) to reduce harmful substance use: a systematic review. *Addiction* 117, 5 (2022), 1220–1241.
[93] Kenzie L Preston and David H Epstein. 2011. Stress in the daily lives of cocaine and heroin users: relationship to mood, craving, relapse triggers, and cocaine use. *Psychopharmacology* 218, 1 (2011), 29–37.
[94] Xin Qin, Jindong Wang, Yiqiang Chen, Wang Lu, and Xinlong Jiang. 2022. Domain generalization for activity recognition via adaptive feature fusion. *ACM Transactions on Intelligent Systems and Technology* 14, 1 (2022), 1–21.
[95] Albert J Qü, Lung-Hao Tai, Christopher D Hall, Emilie M Tu, Maria K Eckstein, Karyna Mishchanchuk, Wan Chen Lin, Juliana B Chase, Andrew F MacAskill, Anne GE Collins, et al. 2025. Nucleus accumbens dopamine release reflects bayesian inference during instrumental learning. *PLOS Computational Biology* 21, 7 (2025), e1013226.
[96] Zahra Rahmati, KOOLAEE ANAHITA KHODABAKHSHI, Mohammd Mehdi Jahangiri, et al. 2021. Investigating the moderating role of self-compassion in the relationship between resilience to stress and drug craving in drug-dependent men. (2021).
[97] Bharath Rathinam and Sinu Ezhumalai. 2021. Resilience among abstinent individuals with substance use disorder. *Indian journal of psychiatric social work* 12, 2 (2021), 96.
[98] Alexandra Rogers and Frances Leslie. 2024. Addiction neurobiologists should study resilience. *Addiction Neuroscience* 11 (2024), 100152.
[99] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. 2019. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731* (2019).
[100] Michael A Sayette. 2016. The role of craving in substance use disorders: theoretical and methodological issues. *Annual review of clinical psychology* 12, 1 (2016), 407–433.
[101] Alexandre Schaefer, Frédéric Nils, Xavier Sanchez, and Pierre Philippot. 2010. Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers. *Cognition and emotion* 24, 7 (2010), 1153–1172.
[102] Christine Dunkel Schetter and Christyn Dolbier. 2011. Resilience in the context of chronic stress and health in adults. *Social and personality psychology compass* 5, 9 (2011), 634–652.
[103] Philip Schmidt, Attila Reiss, Robert Duerichen, and Kristof Van Laerhoven. 2018. Wearable affect and stress recognition: A review. *arXiv preprint arXiv:1811.08854* (2018).

[104] Wolfram Schultz. 2016. Dopamine reward prediction-error signalling: a two-component response. *Nature reviews neuroscience* 17, 3 (2016), 183–195.

[105] Meisam Jamshidi Seikavandi, Fabricio Batista Narcizo, Ted Vucurevich, Andrew Burke Dittberner, and Paolo Burelli. 2025. MuMTAffect: A Multimodal Multitask Affective Framework for Personality and Emotion Recognition from Physiological Signals. In *Proceedings of the 3rd International Workshop on Multimodal and Responsible Affective Computing*. 100–108.

[106] Ovishake Sen, Raghav Soni, Darpan Virmani, Akshar Parekh, Patrick Lehman, Sarthak Jena, Adithi Katikhaneni, Adam Khalifa, and Baibhab Chatterjee. 2025. Low-Latency Neural Inference on an Edge Device for Real-Time Handwriting Recognition from EEG Signals. *arXiv preprint arXiv:2510.19832* (2025).

[107] Harshit Sharma, Yi Xiao, Victoria Tumanova, and Asif Salekin. 2022. Psychophysiological Arousal in Young Children Who Stutter: An Interpretable AI Approach. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 6, 3 (2022), 1–32.

[108] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538* (2017).

[109] Alakhsimar Singh, Nischay Verma, Kanav Goyal, Amritpal Singh, Puneet Kumar, and Xiaobai Li. 2024. VisioPhysioENet: Multimodal Engagement Detection using Visual and Physiological Signals. *arXiv preprint arXiv:2409.16126* (2024).

[110] Rajita Sinha. 2009. Modeling stress and drug craving in the laboratory: implications for addiction treatment development. *Addiction biology* 14, 1 (2009), 84–98.

[111] Rajita Sinha et al. 2024. Stress and substance use disorders: risk, relapse, and treatment outcomes. *The Journal of clinical investigation* 134, 16 (2024).

[112] Rajita Sinha, Dana Catapano, and Stephanie O'Malley. 1999. Stress-induced craving and stress response in cocaine dependent individuals. *Psychopharmacology* 142, 4 (1999), 343–351.

[113] Rajita Sinha, Tiffany Fuse, L-R Aubin, and S Stephanie O'Malley. 2000. Psychological stress, drug-related cues and cocaine craving. *Psychopharmacology* 152, 2 (2000), 140–148.

[114] Congzheng Song and Ananth Raghunathan. 2020. Information leakage in embedding models. In *Proceedings of the 2020 ACM SIGSAC conference on computer and communications security*. 377–390.

[115] John Strang, Nora D Volkow, Louisa Degenhardt, Matthew Hickman, Kimberly Johnson, George F Koob, Brandon DL Marshall, Mark Tyndall, and Sharon L Walsh. 2020. Opioid use disorder. *Nature reviews Disease primers* 6, 1 (2020), 3.

[116] Alan PL Tai, Mei-Kei Leung, Xiujuan Geng, and Way KW Lau. 2023. Conceptualizing psychological resilience through resting-state functional MRI in a mentally healthy population: a systematic review. *Frontiers in behavioral neuroscience* 17 (2023), 1175064.

[117] Allison S Troy and Iris B Mauss. 2011. Resilience in the face of stress: Emotion regulation as a protective factor. *Resilience and mental health: Challenges across the lifespan* 1, 2 (2011), 30–44.

[118] Allison S Troy, Emily C Willroth, Amanda J Shallcross, Nicole R Giuliani, James J Gross, and Iris B Mauss. 2023. Psychological resilience: an affect-regulation framework. *Annual review of psychology* 74, 1 (2023), 547–576.

[119] Michele M Tugade and Barbara L Fredrickson. 2004. Resilient individuals use positive emotions to bounce back from negative emotional experiences. *Journal of personality and social psychology* 86, 2 (2004), 320.

[120] JHM Tulen, P Moleman, HG Van Steenis, and F Boomsma. 1989. Characterization of stress reactions to the Stroop Color Word Test. *Pharmacology Biochemistry and Behavior* 32, 1 (1989), 9–15.

[121] Victoria Tumanova and Nicole Backes. 2019. Autonomic nervous system response to speech production in stuttering and normally fluent preschool-age children. *Journal of Speech, Language, and Hearing Research* 62, 11 (2019), 4030–4044.

[122] Michael Ungar and Linda Theron. 2020. Resilience and mental health: How multisystemic processes contribute to positive outcomes. *The Lancet Psychiatry* 7, 5 (2020), 441–448.

[123] Nilofar Vafaie and Hedy Kober. 2022. Association of drug cues and craving with drug use and relapse: a systematic review and meta-analysis. *JAMA psychiatry* 79, 7 (2022), 641–650.

[124] Steven JA van der Werff, Susan M van den Berg, Justine Nienke Pannekoek, Bernet M Elzinga, and Nic JA Van Der Wee. 2013. Neuroimaging resilience to stress: a review. *Frontiers in behavioral neuroscience* 7 (2013), 39.

[125] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[126] Gideon Vos, Kelly Trinh, Zoltan Sarnyai, and Mostafa Rahimi Azghadi. 2023. Generalizable machine learning for stress monitoring from wearable devices: A systematic literature review. *International Journal of Medical Informatics* 173 (2023), 105026.

[127] Thomas A Waite and Lesley G Campbell. 2006. Controlling the false discovery rate and increasing statistical power in ecological studies. *Ecoscience* 13, 4 (2006), 439–442.

[128] Shuailei Wang, Yumiao He, and Yuguang Huang. 2025. Global, regional, and national trends and burden of opioid use disorder in individuals aged 15 years and above: 1990 to 2021 and projections to 2040. *Epidemiology and Psychiatric Sciences* 34 (2025), e32.

[129] Xiaoming Wang, Hongliu Yu, Søren Kold, Ole Rahbek, and Shaoping Bai. 2023. Wearable sensors for activity monitoring and motion control: A review. *Biomimetic Intelligence and Robotics* 3, 1 (2023), 100089.

[130] Zhong Wang, Lanfang Lei, and Peibei Shi. 2023. Smoking behavior detection algorithm based on YOLOv8-MNC. *Frontiers in Computational Neuroscience* 17 (2023), 1243779.
[131] Aviv Weinstein, Sue Wilson, Jayne Bailey, Judy Myles, and David Nutt. 1997. Imagery of craving in opiate addicts undergoing detoxification. *Drug and Alcohol Dependence* 48, 1 (1997), 25–31.
[132] Gill Windle. 2011. What is resilience? A review and concept analysis. *Reviews in clinical gerontology* 21, 2 (2011), 152–169.
[133] Dieter Wolke, Yanlin Zhou, Yiwen Liu, Robert Eves, Marina Mendonça, and E Sabrina Twilhaar. 2025. A systematic review of conceptualizations and statistical methods in longitudinal studies of resilience. *Nature Mental Health* 3, 9 (2025), 1088–1099.
[134] You Wu, Qingwei Mi, and Tianhan Gao. 2025. A comprehensive review of multimodal emotion recognition: Techniques, challenges, and future directions. *Biomimetics* 10, 7 (2025), 418.
[135] Yi Xiao, Harshit Sharma, Sawinder Kaur, Dessa Bergen-Cico, and Asif Salekin. 2025. Human Heterogeneity Invariant Stress Sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 9, 3 (2025), 1–42.
[136] Yi Xiao, Harshit Sharma, Zhongyang Zhang, Dessa Bergen-Cico, Tauhidur Rahman, and Asif Salekin. 2024. Reading between the heat: Co-teaching body thermal signatures for non-intrusive stress detection. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 4 (2024), 1–30.
[137] Ayako Yamashita, Shin-ichi Yoshioka, and Yuki Yajima. 2021. Resilience and related factors as predictors of relapse risk in patients with substance use disorder: a cross-sectional study. *Substance abuse treatment, prevention, and policy* 16, 1 (2021), 40.
[138] Chunyu Yang, You Zhou, and Mengfan Xia. 2020. How resilience promotes mental health of patients with DSM-5 substance use disorder? The mediation roles of positive affect, self-esteem, and perceived social support. *Frontiers in Psychiatry* 11 (2020), 588968.
[139] Özgür Yürür, Chi Harold Liu, Zhengguo Sheng, Victor CM Leung, Wilfrido Moreno, and Kin K Leung. 2014. Context-awareness for mobile sensing: A survey and future directions. *IEEE Communications Surveys & Tutorials* 18, 1 (2014), 68–93.
[140] Bonan Zhang, Chao Chen, Ickjai Lee, Kyungmi Lee, and Kok-Leong Ong. 2025. A survey on security and privacy issues in wearable health monitoring devices. *Computers & Security* (2025), 104453.
[141] Yiran Zhao, Yujie Tao, Grace Le, Rui Maki, Alexander Adams, Pedro Lopes, and Tanzeem Choudhury. 2023. Affective touch as immediate and passive wearable intervention. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 4 (2023), 1–23.

Table 11. Architecture details of the best HR-AUC main model.

Component	**Layers**	**Input → Output**	**Notes**
Stress Encoder	4	303 → 384	Frozen, dropout = 0.2
Physiology Branch	2	303 → 384	Dropout = 0.193
Gate Projection	2	32 → 303	Sigmoid output (feature gating)
Fusion Projection	1	32 → 384	Subject-conditioned
Fusion Gate	2	3*384 → 1	Sample-wise scalar weight
Classifier	3	384 → 2	Binary logits
Aux Stress Head	1	384 → 2	Training only

A Methodological Transparency and Reproducibility Appendix

This appendix documents implementation details, data usage, and evaluation procedures for the current version of the proposed framework. Due to IRB restrictions involving a vulnerable population with opioid use disorder (OUD), raw physiological recordings and autobiographical narratives cannot be publicly released. To support methodological transparency, we release the full analysis and modeling code used for statistical analysis, model training, and evaluation in an anonymized repository. Anonymized GitHub link.

A.1 Current Main Model Implementation

Overview. The main model operates on a 303-dimensional physiological feature vector together with a subject-level guidance representation. The guidance input can take two forms: (i) HR-dynamics-based embeddings derived from post-task heart-rate dynamics, or (ii) semantic embeddings derived from autobiographical memory narratives.Details on how these two guidance representations are extracted are provided in Appendix D.

In the HR-dynamics configuration, the subject representation is formed by concatenating a 16-dimensional post calm task embedding and a 16-dimensional post stress task embedding into a 32-dimensional vector. In the semantic configuration, the representation is formed by concatenating 1536-dimensional good-memory and bad-memory embeddings into a 3072-dimensional vector.

The model follows a dual-path architecture consisting of a frozen stress branch and a trainable physiology branch. Subject-level guidance is applied through feature-level input gating and representation-level fusion. Table 11 summarizes the architecture of the HR-dynamics configuration.

Architecture Details. Each trainable module is implemented using standard feedforward blocks consisting of a linear transformation followed by GELU activation, layer normalization, and dropout where applicable. The frozen stress encoder shares the same MLP structure but is pretrained separately and remains fixed during craving training.

Subject-Guided Conditioning. The 64-dimensional HR-AUC embedding is first processed through a shared projection block. From this representation, two pathways are derived. The gate pathway produces feature-wise multiplicative gates that are applied element-wise to the 303-dimensional physiological input before it is passed to the trainable physiology branch. The fusion pathway produces a subject-conditioned representation used during branch fusion.

Fusion Mechanism. Fusion is implemented using a sample-dependent gating mechanism. The projected stress representation, the trainable physiology representation, and the subject-conditioned fusion representation are concatenated into a 1152-dimensional vector. This vector is processed by a two-layer gating network to produce a scalar fusion weight, which is used to interpolate between the frozen stress branch and the trainable physiology branch.

Classifier Head. The fused representation is passed through a three-layer MLP classifier, consisting of two hidden layers followed by a final linear layer that outputs logits for binary craving classification.

Training Configuration. The model is trained for 24 epochs with batch size 30 using AdamW. The selected hyperparameters are:

- Learning rate: 0.00131
- Weight decay: 1.12×10^{-6}
- Gradient clipping: 0.424
- Dropout (trainable branch): 0.193

No learning-rate scheduler is used.

Normalization and Loss. Subject-wise z-score normalization is applied to all physiological features. The model is trained with cross-entropy loss using label smoothing of 0.0175.

Regularization. The following regularization terms are applied during training:

- Input-gate sparsity weight: 0.00102
- Auxiliary stress loss weight: 0.000596
- Orthogonality regularization weight: 0.000541

B Physiological Feature Dimensionality and Quality Control

Feature Extraction. For each 30-second physiological window, we extracted a common multimodal feature space from acceleration (ACC), blood volume pulse (BVP), electrodermal activity (EDA), heart rate (HR), heart-rate variability (HRV), and skin temperature (TEMP). The initial shared feature space contained 334 dimensions.

The modality-wise feature dimensionalities were as follows: 99 ACC features, 34 BVP-derived features, 105 EDA features, 31 HR features, 35 HRV features, and 30 TEMP features.

Across modalities, the feature set included three broad families of descriptors:

- **Summary statistics:** mean, median, standard deviation, variance, interquartile range, minimum, maximum, percentile summaries, skewness, and kurtosis,
- **Shape descriptors:** root mean square (RMS), mean absolute value, median absolute deviation, peak amplitude, crest factor, impulse factor, waveform length, and zero-crossing counts,
- **Spectral descriptors:** total spectral power, spectral centroid, peak frequency, bandwidth, spectral entropy, and band-limited power features.

In addition to these shared descriptors, modality-specific features were included. ACC features captured per-axis statistics (x/y/z), magnitude-based features, inter-axis correlations, signal energy, mean absolute temporal differences, and signal magnitude area. BVP features included waveform statistics, peak-based descriptors, and signal-quality indicators derived from PPG signals, from which HR and HRV features were computed. EDA features included tonic and phasic decomposition outputs and skin conductance response (SCR) event descriptors such as onset counts, peak counts, amplitudes, rise times, and recovery times. HR features included temporal difference measures, and TEMP features primarily consisted of statistical, shape, and spectral summaries.

Quality Control Procedure. A global quality-control procedure was applied uniformly across all participants and conditions prior to model training. Features were excluded based on two criteria:

- **Missingness:** features with a missing-value rate greater than 20% across windows,
- **Low variation:** features with fewer than two unique non-null values,

Feature filtering was performed without reference to stress labels, craving labels, subject identity, or downstream model outcomes.

The removed features were concentrated in measurements known to be unreliable under short-window analysis, including low-frequency HRV components (e.g., ultra-low- and very-low-frequency bands), long-horizon HRV statistics (e.g., SDANN and SDNNI), and a small number of sparse or weakly varying spectral and zero-crossing descriptors across ACC, HR, EDA, and TEMP modalities.

Final Feature Space. After quality control and final feature harmonization, the predictive feature space used in all main experiments consisted of 303 physiological dimensions per window. This representation was held fixed across all models and experimental conditions.

C Additional Details for Physiological Modulation Analysis

C.1 Physiological Systems Organization Strategy

For interpretability and higher-level analysis, extracted features in Section 4.1 were further organized into four physiological systems based on their sensing modality and physiological function signals [73, 80, 84, 129]. *Cardiovascular* features capture heart-rate and heart-rate variability dynamics derived from photoplethysmography (PPG); *Electrodermal* features characterize tonic and phasic skin conductance activity reflecting sympathetic arousal; *Movement* features summarize body motion and acceleration patterns from the wrist-worn accelerometer; and *Thermoregulation* features capture changes in peripheral skin temperature. This system-level organization is used to assess whether physiological modulation generalizes beyond individual features and recurs at the level of coordinated physiological subsystems.

C.2 Task-Level Feature Aggregation

Window-level physiological features were first globally standardized and then aggregated to the level of each experimental task block defined in Section 3.3.2. For each participant and task, we averaged all valid windows within the task interval. Task-level aggregation was used because stress and craving annotations were collected at the task level rather than at individual physiological window timestamps.

To isolate task-evoked physiological modulation and reduce between-subject baseline differences, all task-level features were baseline-corrected using the neutral *Calm video viewing* task as a subject-specific reference [20, 55, 56]. This correction emphasizes physiological changes relative to each participant's own relaxed baseline.

C.3 GEE Model Specification

Task-level associations were estimated using population-average generalized estimating equations (GEE) with a Gaussian family and an exchangeable working correlation structure [74]. GEE was used because each participant contributed repeated task observations, and the method estimates population-average effects while providing robust standard errors even when the within-subject correlation structure is imperfectly specified.

Each GEE model included the task condition of interest, group membership (Control vs. OUD), and the number of aggregated windows as covariates. Group membership controlled for between-group differences, while the number of windows controlled for variability in physiological evidence availability across tasks. All GEE models were fitted on the pooled cohort including both Control and OUD participants.

C.4 Feature-Level and System-Level Analyses

Extracted physiological features were organized a priori into four systems based on sensing modality and physiological function: ***Cardiovascular, Electrodermal, Movement, and Thermoregulation*** (Section C.1). Analyses were conducted at both the feature and system levels.

At the feature level, each baseline-corrected task-level feature was tested individually using the GEE framework described above. Multiple comparisons were controlled using the Benjamini–Hochberg false discovery rate (FDR) procedure [13].

At the system level, feature-level effects were summarized within each physiological system using two complementary representations. First, we computed mean-based composites by averaging standardized features within each system. Second, we derived PCA-based representations using the first principal component (PC1) within each system [30]. Both representations were evaluated using the same GEE specification.

System-level effects were considered robust when feature-level changes within the same physiological modality were directionally consistent and supported by permutation-based significance testing. System-level aggregation is used here to assess whether task-related modulation is consistently observed across related features from the same sensing modality, rather than to assign a single physical meaning to the aggregated representation.

D Subject-Level Guidance Embeddings

The model supports two types of subject-level guidance embeddings: (i) HR-dynamics-based embeddings derived from post-task physiological signals, and (ii) semantic embeddings derived from autobiographical narratives. Both representations are computed once per subject and used as fixed inputs during model training and inference.

HR-Dynamics Embedding. The HR-dynamics embedding is constructed from short heart-rate segments collected after calm and stress tasks. For each subject, heart-rate signals are first normalized by subtracting a baseline value computed during a calm reference period (Calm video viewing). Two 15s segments are then extracted: a post-calm segment and a post-stress segment.

Each segment is summarized using 16 statistical and temporal descriptors capturing both magnitude and dynamic properties of heart-rate changes: (1) mean, (2) standard deviation, (3) maximum, (4) minimum, (5) signed area under the curve (AUC), (6) absolute AUC, (7) positive AUC, (8) negative AUC, (9) difference between final and initial values, (10) slope, (11) time to peak, (12) time to trough, (13) peak value, (14) trough value, (15) number of zero crossings, and (16) curvature (mean absolute second difference).

Concatenating the post-calm and post-stress representations yields a 32-dimensional HR-dynamics embedding per subject.

Semantic Embedding. The semantic embedding is derived from autobiographical memory narratives collected during the study. Each subject provides two narratives: one corresponding to a positive (good) memory and one corresponding to a negative (bad) memory.

Each narrative is independently encoded using a pretrained text embedding model. In the current implementation, embeddings are generated using an OpenAI text embedding model (`text-embedding-3-small`), producing a 1536-dimensional vector for each narrative. For robustness analysis, embeddings generated using a higher-capacity model (`text-embedding-3-large`) were also evaluated, but the main results use the smaller model.

The final subject-level semantic representation is constructed by concatenating the good-memory and bad-memory embeddings, resulting in a 3072-dimensional vector.

No fine-tuning or task-specific adaptation is applied to the embedding model. The semantic embeddings are computed once per subject and reused across all physiological windows associated with that subject, ensuring strict separation between subject-level context and window-level inputs.

Summary. Both HR-dynamics and semantic embeddings follow the same interface: a fixed subject-level vector that encodes individual differences. They differ only in modality and dimensionality, with HR-dynamics capturing physiological recovery patterns and semantic embeddings capturing narrative-level representations of autobiographical memory.

E Additional Details for Autobiographical Recall as a Resilience Proxy

E.1 Neurophysiological Context: Resilience and Autobiographical Memory Recall

We next introduce a brief neurophysiological context to motivate the use of autobiographical memory recall as a marker of resilience-related individual differences.

Resilience and Positive Autobiographical memories. In opioid use disorder (OUD), chronic drug exposure disrupts both the brain's reward and stress systems, leading to accumulated allostatic load and a pronounced reward deficit [65, 66, 68]. The *reward deficit and stress surfeit* framework proposes that addiction is characterized by two co-occurring processes: reduced sensitivity to non-drug rewards and heightened stress reactivity [67]. Together, these changes fundamentally alter how individuals with OUD respond to emotionally salient cues, including autobiographical memories, particularly the positive memory recalls.

In healthy individuals without substance use disorder, recalling a positive personal memory typically reactivates natural reward circuits, producing a pleasant emotional state similar to re-experiencing the event itself. In this case, the expected reward associated with the memory closely matches the experienced emotional response, resulting in a stable and positive affective experience. In contrast, among individuals with OUD— like the patients in our study who require MOUD—this process can become disrupted. The reward deficit and stress surfeit models established by Koob and colleagues provide a neurobiological framework for understanding this phenomenon. [67, 68]

Because baseline hedonic tone is blunted and stress systems are sensitized, attempts to recall positive, non-drug-related memories may fail to elicit the expected pleasurable response. Instead, positive memory recall can evoke ambivalent affect, anxiety, and physiological stress responses rather than relief or pleasure. This counterintuitive phenomenon reflects altered reward processing following chronic opioid exposure. It has been hypothesized that key reward-related regions, including the nucleus accumbens, function as comparators that evaluate the difference between expected and experienced reward [95, 104]. When recalling a positive memory, individuals with OUD may experience a mismatch between the anticipated pleasure and the diminished emotional response, resulting in a negative prediction error. Rather than reinforcing positive affect, this discrepancy may emphasize loss or inability to feel pleasure, thereby triggering stress and anxiety that manifest as heightened physiological arousal. These contrasting processes can be summarized as follows:

Among healthy individuals without OUD or other SUD:

- Positive memory recall task reactivates the reward circuits and produces a positive hedonic state
- This results in pleasant memory reactivation whereby their expectation matches their experience and results in a matched comparison between the expected and experienced positive memory recall.

Among individuals with OUD in a reward deficit state:

- Positive memory recall task attempt to reactivate reward circuits results in a failure to achieve expected hedonic state
- This results in a discrepancy between the expected positive hedonic state (negative prediction error) and the outcome.
- This negative prediction error may be interpreted as loss or threat and produce anxiety and physiological stress response.

Figure 11 summarizes this pathway. Chronic opioid dependence drives neuroadaptations in the mesolimbic reward circuit (ventral tegmental area, nucleus accumbens), the extended amygdala, and the HPA axis. When a memory cue is encountered, the blunted reward circuit fails to produce the anticipated pleasure, while the sensitized stress system produces exaggerated physiological arousal—creating a feedback loop that links memory-evoked stress to craving and relapse risk.

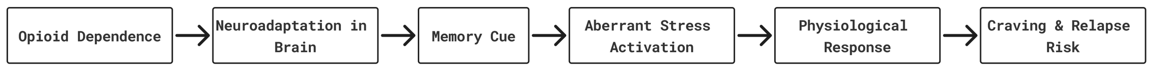

Fig. 11. Neurobiological flow of stress–reward dysregulation in opioid use disorder (OUD). Chronic opioid dependence alters reward and stress circuits, transforming positive memory recall into aberrant stress activation, which increases craving and relapse vulnerability.

Together, this framework provides a mechanistic basis for our hypothesis that the affective and semantic characteristics of positive autobiographical memory recall reflect resilience-related differences in stress reactivity and craving vulnerability.

Resilience and Negative Autobiographical memories. Negative autobiographical memories provide a natural context for examining resilience because they directly engage stress and emotion regulation processes. Prior work shows that emotion regulation plays a central role in how negative autobiographical memories are recalled and experienced [25, 88].

Individuals who use adaptive regulation strategies, such as cognitive reappraisal, tend to recall negative personal events with greater detail and reduced negative bias [25]. In addition, the ability to control or suppress the retrieval of intrusive memories is considered a core component of emotion regulation and has been shown to reduce the emotional impact of unwanted recollections [35, 61].

Resilience is commonly conceptualized as a set of adaptive processes that enable individuals to maintain functioning in the face of adversity [118]. Emotion regulation is a key component of these processes [102], as it governs how emotional and physiological responses to stressors and adverse events are modulated [5, 26, 117]. Neurophysiological evidence further suggests that brain systems involved in arousal regulation and emotional reappraisal mediate the relationship between trait resilience and post-stress outcomes [124].

Taken together, these findings suggest that resilience shapes how negative experiences are encoded, recalled, and physiologically regulated. Individual differences in regulatory capacity influence not only the emotional tone and structure of negative autobiographical narratives, but also the magnitude and recovery of stress responses associated with them. Accordingly, negative autobiographical memory recall offers a meaningful lens through which resilience-related differences in stress regulation can be examined.

E.2 Language Embedding Construction

Autobiographical Language was derived from participants' descriptions of recent positive and negative personal experiences. Context-neutral speech was derived from responses to neutral everyday prompts collected during the study protocol. The context-neutral condition was used as a control for general linguistic properties such as speaking style, verbosity, and transcription characteristics without autobiographical or emotional content.

Each transcript was encoded into a high-dimensional semantic embedding using a pretrained language model. For each participant, embeddings from positive and negative autobiographical recall were aggregated to obtain a subject-level autobiographical representation. Context-neutral speech embeddings were processed separately and used only as a comparison condition.

E.3 PLS-Based Representational Alignment

To test whether language representations encode resilience-related recovery information, we used Partial Least Squares (PLS) to align language embeddings with HR-AUC. Unlike PCA, which identifies directions of maximal variance in the embedding space, PLS identifies latent dimensions that maximize covariance between language embeddings and an external variable of interest. Here, the external variable is HR-AUC, our physiology-grounded proxy for post-stress recovery.

For each language condition, embeddings were projected onto a single PLS component to obtain a compact latent representation. We then assessed the monotonic association between the PLS-derived language representation and HR-AUC using Spearman rank correlation. Statistical significance was evaluated using permutation testing with 5,000 permutations, where HR-AUC values were randomly permuted across participants to construct a nonparametric null distribution.

E.4 Interpretation and Limitations

A strong association between autobiographical language and HR-AUC suggests that autobiographical recall captures subject-level information related to physiological recovery. However, this analysis does not imply that language directly measures psychological resilience. Instead, autobiographical language is treated as a resilience-related descriptor that aligns with a physiology-grounded recovery proxy.

Finally, autobiographical narratives may contain sensitive personal information. In deployment, the raw transcript should not be required at inference time; instead, a compact subject-level embedding can be computed once, stored securely, and reused as guidance for wearable craving inference.

F Partial Least Squares (PLS) Implementation Details

Partial Least Squares (PLS) is a multivariate dimensionality reduction technique designed to relate two sets of variables measured on the same observations by extracting latent components that maximize their shared covariance. Given a predictor matrix $\mathbf{X} \in \mathbb{R}^{N \times D}$ (language embeddings) and a target vector $\mathbf{y} \in \mathbb{R}^{N}$ (heart-rate recovery AUC), PLS identifies latent variables defined as linear projections of $\mathbf{X}$ that are optimally aligned with $\mathbf{y}$ in terms of covariance. In this work, we employ a single-component PLS model, yielding a one-dimensional latent representation. Formally, PLS learns a weight vector $\mathbf{w} \in \mathbb{R}^{D}$ such that the projected scores $\mathbf{z} = \mathbf{X}\mathbf{w}$ maximize $\mathrm{Cov}(\mathbf{z}, \mathbf{y})$ under standard normalization constraints. This projection corresponds to a weighted combination of all embedding dimensions rather than the selection of individual dimensions and differs from variance-based methods such as principal component analysis (PCA), which optimize $\mathrm{Var}(\mathbf{X}\mathbf{w})$ independently of the outcome variable. Prior to PLS estimation, language embeddings were standardized to zero mean and unit variance across participants, and heart-rate recovery AUC values were used as provided by the physiological preprocessing pipeline described in Section 5.2. PLS was implemented using a standard Partial Least Squares regression formulation, with the number of components fixed to one to obtain a compact shared latent axis. Associations between the resulting PLS-derived latent scores and physiological recovery were evaluated using Spearman rank correlation, with statistical significance assessed via permutation testing (5,000 permutations), in which HR AUC values were randomly permuted across participants to construct a nonparametric null distribution.

G Qualitative Visualization: Narrative Organization and Appraisal

To aid interpretation of the alignment between autobiographical language and physiological resilience, we present qualitative visualizations of representative autobiographical narratives (Figure 12). Each narrative is rendered using phrase-level appraisal scores generated by a large language model (GPT-4o), prompted to assess resilience-related psychological appraisal within narrative context [12, 58]. The model outputs relative phrase-level scores reflecting positive (red), negative (blue), or neutral (gray) appraisal, with color intensity indicating relative affective strength within each narrative; the full prompt specification is provided in the Appendix I. These visualizations are intended as interpretive tools rather than quantitative measurements and do not represent direct estimates of psychological resilience or momentary affect.

(a) Participant with OUD recalling a positive memory

(b) Participant with OUD recalling a negative memory

(c) Control participant recalling a positive memory

(d) Control participant recalling a negative memory

Fig. 12. Qualitative visualizations of autobiographical narratives based on LLM-derived, phrase-level resilience-related appraisal. Red indicates relatively positive resilience-related appraisal, blue indicates relatively negative appraisal, and gray indicates neutral language; color intensity reflects the relative strength of appraisal within each narrative.

Across narratives, systematic differences in affective distribution and organization are evident. Positive autobiographical memories from control participants tend to exhibit localized and thematically coherent clusters of positively appraised language, whereas negative memories show denser and more sustained negatively appraised phrasing organized around salient events. In contrast, autobiographical narratives from participants with opioid use disorder (OUD) frequently display mixed or negatively appraised segments even during positive memory recall, accompanied by abrupt affective shifts and reduced thematic integration. These patterns suggest difficulties in maintaining coherent positive appraisal despite ostensibly positive content, consistent with prior narrative and clinical accounts of disrupted emotion regulation and meaning integration in addiction-related populations [3, 38].

Taken together, these qualitative observations indicate that autobiographical narratives encode structured information about how individuals appraise, organize, and regulate personal experience. Differences in affective coherence, mixed valence, and narrative organization reflect systematic resilience-related appraisal processes rather than simple differences in memory valence, providing qualitative support for the modeling approach introduced in RQ3.

H Resilience Grouping

To analyze subject-specific differences in physiological interpretation, participants are grouped into high- and low-resilience groups based on their post-stress heart-rate recovery. Specifically, resilience is quantified using the HR-AUC measure derived from post-stress heart-rate dynamics, computed as the deviation from baseline following stress exposure. Lower HR-AUC values indicate faster recovery and higher resilience, while higher HR-AUC values indicate slower recovery and lower resilience. Participants are split into two groups using a median split over subject-level HR-AUC values. Subjects with HR-AUC values below the median are assigned to the high-resilience group, and those above the median are assigned to the low-resilience group.

I LLM Prompt for Resilience-Related Narrative Appraisal

To generate the qualitative narrative visualizations presented in Appendix G, we employ a large language model (GPT-4o) as an interpretive annotation tool to assess *resilience-related psychological appraisal* within autobiographical narratives. This procedure is designed to support qualitative inspection of narrative organization, mixed valence, and appraisal coherence, rather than to produce quantitative estimates of psychological resilience or momentary affective states.

The language model is prompted to analyze how individuals appraise and organize personal experiences in narrative language. Specifically, the model is first asked to form an overall impression of the narrative's psychological tone along a resilience–vulnerability dimension (e.g., resilient versus vulnerable). This overall assessment serves only as contextual background to anchor relative phrase-level appraisal within the narrative and is not treated as a measurement of resilience.

The model then segments each narrative into short, continuous phrases (1–5 words each), ensuring full coverage of the text in sequential order. For each phrase, the model assigns a relative appraisal score reflecting how positively, negatively, or neutrally the phrase contributes to the narrative's overall psychological appraisal. Importantly, the prompt explicitly allows local deviations and mixed appraisal, such that negatively appraised segments may appear within an overall positive narrative and vice versa. All phrase-level scores are interpreted relative to the narrative context and are not intended to be comparable across individuals or narratives.

The full prompt used in our experiments is reproduced below for transparency and reproducibility.

Prompt:

You are a psychologist and linguistics expert who analyzes how individuals appraise and organize personal experiences in narrative language.

Your task is to assess resilience-related psychological appraisal within an autobiographical narrative. This appraisal reflects how positively, negatively, or neutrally experiences are interpreted and integrated, rather than momentary emotional intensity.

First, read the entire text carefully to form an overall impression of the narrative's psychological tone along a resilience–vulnerability dimension, ranging from −1 (highly vulnerable, hopeless, anxious) to +1 (highly resilient, hopeful, calm, confident). This overall score is used only as contextual background.

Next, segment the text into short, continuous phrases (1–5 words each), ensuring that every word in the text is covered exactly once and in order.

For each phrase:

- Consider the phrase's psychological meaning within the broader narrative context.
- Assign a relative appraisal score that reflects how positively, negatively, or neutrally the phrase contributes to the narrative's overall psychological appraisal.
- Allow local deviations and mixed appraisal, even when the overall narrative tone is positive or negative.
- Output a score between −1 and +1, where 0 represents neutral or descriptive language.

Return ONLY valid JSON in the following format:

```
{
  "overall_score": <float between -1 and 1>,
  "phrases": [
    {"phrase": "<text>", "score": <float between -1 and 1>}
  ]
}
```

These annotations are used exclusively for qualitative visualization and interpretive analysis and are not used as predictive features or targets in any modeling stage.