CrossShift: Quantifying Interpersonal Differences in Mobile Sensing for Mental Health

Panyu Zhang · panyu@kaist.ac.kr

with M. Park, T. Ismatzoda, A. Mustafakulov, U. Ahmed, O. Najimov, J. Alikhanov, S. Ghosh, U. Lee

KAIST · HumbleBeeAI · Gachon University · BITS Pilani, K K Birla Goa Campus

BACKGROUND

Smartphones and Wearables Enable In-the-Wild Affective and Mental Health Computing

Continuous smartphone and wearable sensing captures physiology, behavior, and context in daily life, enabling scalable modeling of affective and mental health states

- Smartphones and wearables are widely used in everyday life
- They continuously capture physiological, behavioral, and contextual signals
- With self-reports as labels, they enable affective and mental-health computing



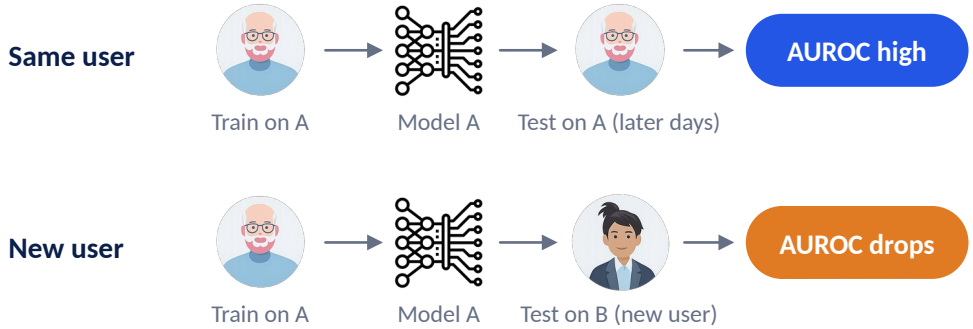
This talk CrossShift — a framework that tests which user-level differences exist, measures them, and attributes the cross-user performance drop to them

RESEARCH GAP

Measuring a Drop Is Not the Same as Explaining It

To fix cross-user failure, we must know what shifts between users: behavior, labels, or the behavior → label mapping

- A model that works well on its own user drops sharply on an unseen user
- Prior studies report the same within-user → cross-user gap
- “Users differ” is an observation, not an explanation
- CrossShift asks which kind of difference actually causes the drop



Reported cross-user (leave-one-subject-out) performance in prior work

Ciman & Wac ’18
F1 0.79–0.85 within-user → 0.70–0.80 LOSO

Toshnazarov et al. ’24
65.8% F1 under LOSO

Zhang et al. ’24
58% AUROC under LOSO

Ciman, M., Wac, K. Individuals’ Stress Assessment Using Human–Smartphone Interaction Analysis. IEEE TAFFC, 2018.
Toshnazarov, K., et al. SOSW: Stress Sensing With Off-the-Shelf Smartwatches in the Wild. IEEE IoT Journal, 2024. · Zhang, P., et al. A Reproducible Stress Prediction Pipeline with Mobile Sensor Data. IMWUT, 2024.

RESEARCH QUESTIONS

Research Questions

We ask what interpersonal differences exist, and which ones explain cross-user generalization failure

RQ1

What interpersonal differences are present across users?

Detect — are covariate, label, conditional, and concept shifts present between users?

RQ2

Which interpersonal differences contribute most to cross-user performance loss?

Attribute — which shift magnitudes best explain the drop when a model moves to a new user?

How we answer them measure each shift between users → measure the cross-user performance drop → relate the two

Four Shift Types, Four Different Failure Mechanisms

Users can differ in what they do, how often they are stressed, how they act when stressed, or what their behavior means

1 Covariate shift P(X) differs

Users simply behave differently



User A: exercises a lot



User B: mostly sedentary

2 Label shift P(Y) differs

Users differ in how often they are stressed



User A: mostly relaxed



User B: mostly stressed

3 Conditional shift P(X | Y) differs

In the same mental state, users behave differently



When stressed, A exercises more



When stressed, B exercises less

4 Concept shift P(Y | X) differs

The same behavior means a different mental state



Running relieves A's stress



Running raises B's stress

FRAMEWORK

CrossShift: A New Shift-Attribution Framework

A four-component framework links user-level shifts to cross-user transferability loss



Shift testing

Are the four shifts statistically present between users?

p-values and effect sizes per user group

Shift quantification

How large is each shift for every pair of users?

one distance per user pair

Model evaluation

How much does a model lose on a new user?

Δ = AUROC (self) − AUROC (cross)

Regression attribution

Which shift explains the loss?

β per shift type × behavior category

Step 1 answers RQ1 (detect) · steps 2–4 answer RQ2 (attribute) · each user is treated as one domain

DATASETS

Datasets and Behavior Feature Categories

We analyze user-level shifts across eight mobile mental-health sensing datasets and five behavioral feature categories

- 6 public datasets: EmoPhone, K-EmoPhone, StudentLife, CrossCheck, GLOBEM, College Experience Study
- Features are grouped into shared behavioral categories for cross-dataset analysis
- Static demographics are excluded from user-level shift analysis

Feature category	Feature Streams
Social Behavior	call count/duration; message sent/received count
Phone Usage	app usage; screen events; typing/key events
Physical Status	activity types; steps; distance; calories
Mobility	places visited; time at clustered locations
Sleep	previous-night sleep duration; sleep onset time

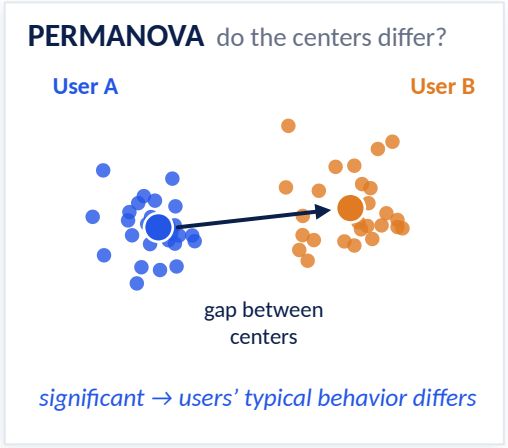
Zhang, P., et al. EmoPhone: A Multi-Wave Dataset for In-the-Wild Mobile and Wearable Affect Sensing. NeurIPS Evaluations and Datasets Track, 2026.
Kang, S., et al. K-EmoPhone: A Mobile and Wearable Dataset with In-Situ Emotion, Stress, and Attention Labels. Scientific Data, 2023.
Wang, R., et al. StudentLife: Assessing Mental Health, Academic Performance and Behavioral Risk Factors for College Students. Ubicomp, 2014.
Ben-Zeev, D., et al. CrossCheck Study: Digital Indicators of Relapse for Schizophrenia Spectrum Disorders. Cornell PBH, 2015.
Xu, X., et al. GLOBEM: Cross-Dataset Generalization of Longitudinal Human Behavior Modeling. IMWUT, 2023.
Nepal, S., et al. Capturing the College Experience: A Four-Year Mobile Sensing Study of Mental Health, Resilience and Behavior of College Students during the Pandemic. IMWUT, 2024.

METHOD · RQ1

RQ1: Are Between-User Shifts Statistically Detectable?

Two permutation tests answer two intuitive questions about any two users' behavior data

- Each user is one group; tests run per behavior category × shift type
- Covariate X / conditional shift X|Y: PERMANOVA and PERMDISP on the multivariate behavior features
- Label Y / concept shift Y|X: Pearson χ^2 — does the label (or behavior → label) table depend on who the user is?



Each dot is one day of a user's behavior features; both tests permute user labels over Euclidean distances between days · adapted from Nanchen et al. (AIES '23)

Nanchen, A., et al. Keep Sensors in Check: Disentangling Country-Level Generalization Issues in Mobile Sensor-Based Models with Diversity Scores. AIES '23, 2023.

Interpersonal Differences Are Pervasive

User-level shifts appear across behavior categories and shift types

- All four shift types are statistically detectable across users
- Phone usage, sleep, and social behavior show the most consistent between-user diversity
- User-wise normalization reduces covariate/conditional differences, but behavior–label differences persist

RQ1 Evidence: User-level Shifts Are Detected	
✓**Covariate**	Feature distributions differ
✓**Label**	Outcome prevalence differs
✓**Conditional**	Behavior under same label differs
✓**Concept**	Behavior–label mapping differs

Most Consistent Covariate-shift Sources

Phone usage Sleep Social Behavior

METHOD · RQ2

RQ2a: How Large Is Each Shift?

For every user pair (A, B) and behavior category, each shift becomes one number: larger = more different

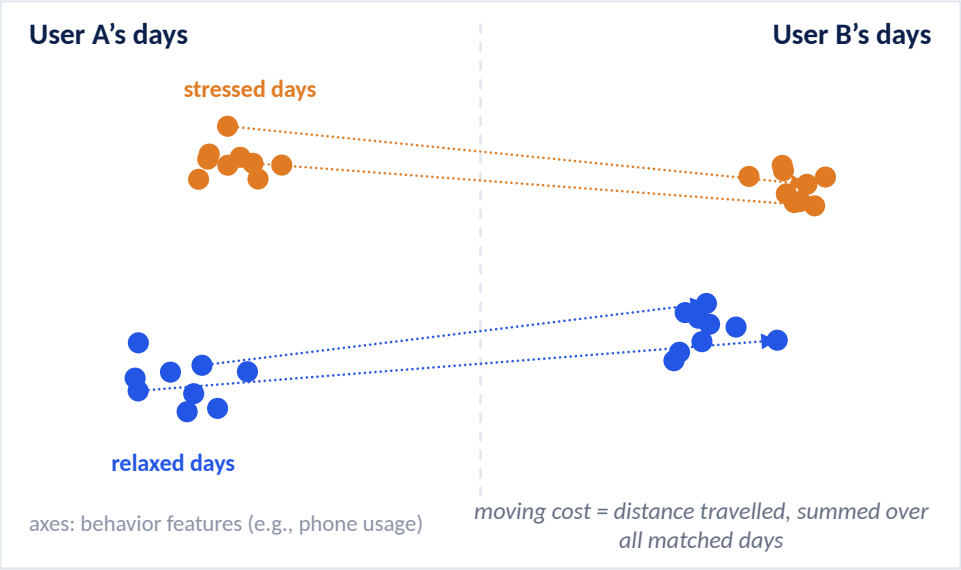
Shift type	What we compare between User A and B	Metric	Intuition
Covariate · P(X)	Distributions of behavior features	Wasserstein distance	*Effort to reshape A's behavior histogram into B's*
Label · P(Y)	How often each user is stressed	L2 distance	*Gap between the two stress rates*
Conditional · P(X | Y)	Behavior within the same mental state	OTDD conditional term, feature-importance normalized	*Effort to move A's stressed days onto B's stressed days (next slide)*
Concept · P(Y | X)	Behavior → mental-state mapping	JS divergence between A's and B's models	*How much the two users' models disagree on the same days*

Alvarez-Melis, D., et al. Geometric Dataset Distances via Optimal Transport. NeurIPS, 2020.
Gardner, J., et al. Benchmarking Distribution Shift in Tabular Data with TableShift. NeurIPS, 2023.
Polo, F. M., et al. A Unified Framework for Dataset Shift Diagnostics. Information Sciences, 2023.
Naumann, P., et al. Wasserstein Distances Made Explainable: Insights into Dataset Shifts and Transport Phenomena. arXiv, 2025.

METHOD · RQ2

How OTDD Measures Conditional Shift

The cheapest way to move one user's days onto the other's — matched within the same mental state



- Each day is a point in behavior-feature space, colored by the self-reported state
- Optimal transport finds the cheapest way to move A's points onto B's; the total moving cost is the distance between the two users
- Conditional term: stressed days may only be matched to stressed days, relaxed to relaxed
- Feature-importance normalization: features that matter for prediction count more; raw scale differences count less

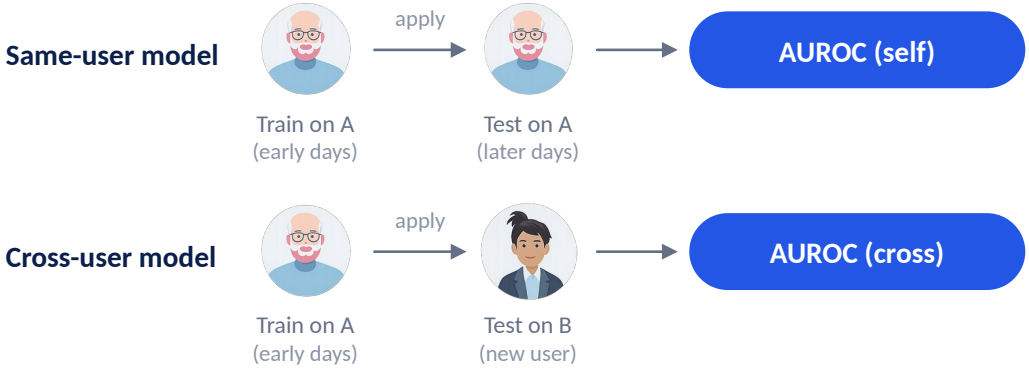
Large cost ⟹ even in the same mental state, A and B behave differently

Alvarez-Melis, D., et al. Geometric Dataset Distances via Optimal Transport. NeurIPS, 2020.
Gardner, J., et al. Benchmarking Distribution Shift in Tabular Data with TableShift. NeurIPS, 2023.

RQ2b: What Transferability Loss Must Be Explained?

For each user pair (A, B): how much does A's model lose when it is applied to B?

- Train one model per source user A on A's early days
- Test it on A's later days (same user) and on every other user B (cross user)
- The drop from same-user to cross-user AUROC is the transferability loss we want to explain
- Only users with enough labels and a reliable same-user baseline are kept



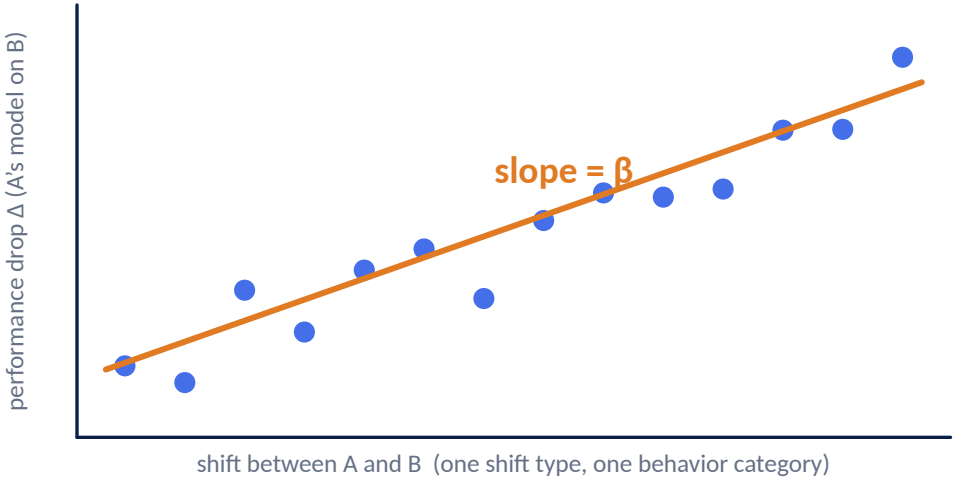
Transferability loss Δ = AUROC (self) − AUROC (cross)

how much A's model loses when it moves to a new user; one Δ for every user pair

RQ2c: Which Shift Explains the Drop?

A regression slope β tells how strongly each shift predicts the cross-user drop

- One dot = one user pair (A → B)
- x: how different A and B are on one shift type; y: how much A's model drops on B
- Fit a line for every shift type × behavior category; its slope β is the attribution score
- Compare β across shift types: the largest β points to the main culprit



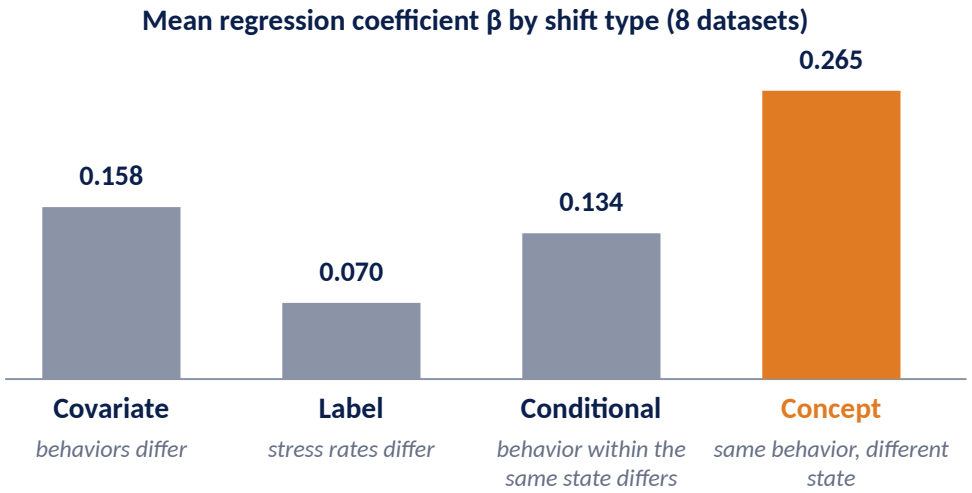
β = extra performance drop per unit of shift · larger β ⟹ that shift explains more of the failure

RESULTS • RQ2

Concept Shift Best Explains Cross-User Failure Overall

The strongest average predictor of transferability loss is concept shift: changes in $P(Y|X)$

- Concept shift — the same behavior meaning a different state — has the largest β
- Covariate and conditional shift still matter, and lead in a few datasets
- So the main cause is a changed behavior → state mapping, not just different behavior

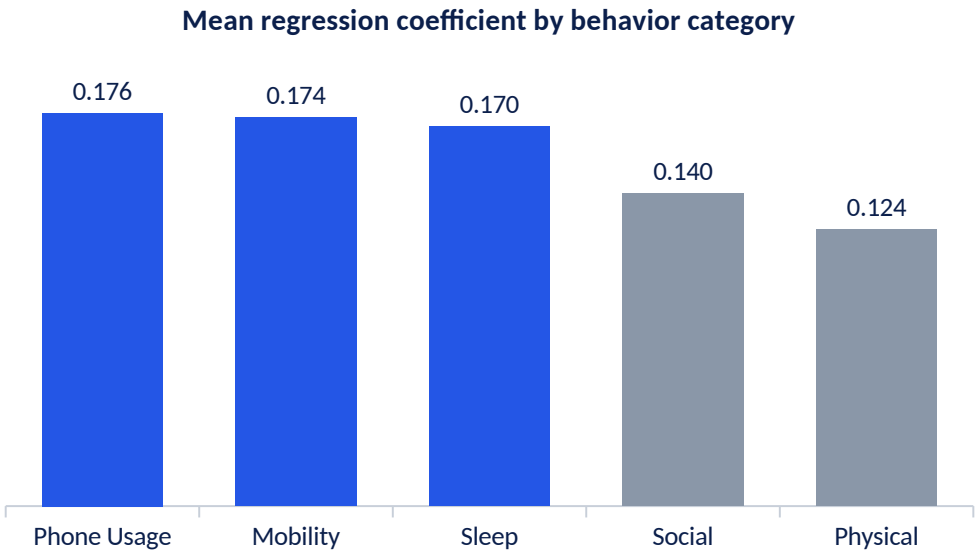


RESULTS · RQ2

Phone Usage, Mobility, and Sleep Drive the Drop

By behavior category, phone usage, mobility, and sleep contribute most on average, but no single category dominates everywhere

- Phone usage shows the largest average attribution coefficient overall
- Mobility and sleep contribute nearly as much; physical status contributes least
- Category importance varies across datasets, suggesting dataset-specific shift patterns



TECHNICAL PROBING

Preliminary Technical Probing: Existing Methods Are Helpful but Insufficient

CrossShift diagnosis motivates technical directions, but existing generalization methods only provide partial relief

- Domain generalization (DG) improves over the XGBoost baseline, but the gain is modest
- TabPFN-style foundation model is competitive, but not clearly stronger
- Partial personalization helps when sparse target-user labels are available
- Out-of-Distribution (OOD)-guided gating suggests CrossShift can help decide when to adapt

Direction	Key result (AUROC)	Takeaway
DG	HHISS 0.5934 vs. XGBoost 0.5693	Modest gain
TabPFN	TabPFN 0.5698	Not clearly stronger
Partial personalization	0.5778 → 0.5988	Target labels help
OOD-guided gating	Adapt 35/152 users; +0.0121	Selective adaptation

Xiao et al., Human Heterogeneity Invariant Stress Sensing, IMWUT 2025.
Cheng et al., Realistic Evaluation of TabPFN v2 in Open Environments. arXiv:2505.16226, 2025.
Meegahapola et al., Generalization and Personalization of Mobile Sensing-Based Mood Inference Models, IMWUT 2023.

IMPLICATIONS

Feature Alignment Alone Is Not Enough

If similar behaviors map to different labels across users, input alignment cannot fully solve cross-user failure

01 **Input alignment is insufficient**

Normalizing or aligning features removes covariate shift, but the behavior → state mapping still differs between users

02 **Personalization is principled**

A few target-user labels address concept shift directly — and CrossShift shows when they are needed

03 **Shift-aware deployment**

Monitor shift scores per user: recalibrate when covariate shift rises, personalize selectively when concept shift persists

Beyond this work we assume a stable behavior → state mapping within each user over the study window; extending CrossShift to within-user drift over time, and reusing shift scores for cohort discovery and group personalization, is next

CONTRIBUTION

CrossShift Turns Cross-User Drops into Shift Diagnoses

A reusable framework, a clear diagnosis, and guidance for what to do about it

Framework

test → quantify → evaluate → attribute

A reusable, open-source framework for explaining cross-user generalization failure, applied to 8 datasets and 696 users

Diagnosis

concept shift is the primary driver

The same behavior maps to different mental states across users — and this explains the drop better than different behavior itself

Guidance

adapt when, and only when, needed

Informs domain adaptation and generalization, OOD user detection, and selective personalization

THANK YOU

Thank You — Questions?

Panyu Zhang · KAIST · panyu@kaist.ac.kr

Paper

Code

EmoPhone Dataset

UbiComp ISWC 2026

KAIST 1971

ICLab INTERACTIVE COMPUTING LAB

TAKE-HOME MESSAGE

Cross-user failure is driven mainly by concept shift: the same behavior maps to different mental states across users

Feature alignment alone is not enough — diagnose the shift, then adapt or personalize selectively