Cross-Temporal Attention Fusion (CTAF) for Multimodal Physiological Signals in Self-Supervised Learning

Arian Khorasani
Department of Information Technologies, HEC Montréal
Mila-Quebec Artificial Intelligence Institute
Arian.Khorasani@mila.quebec

Théophile Demazure
Department of Informatio n Technologies, HEC Montréal
theophile.demazure@hec.ca

Abstract

We study multimodal affect modeling when EEG and peripheral physiology are asynchronous, which most fusion methods ignore or handle with costly warping. We propose Cross-Temporal Attention Fusion (CTAF), a self-supervised module that learns soft bidirectional alignments between modalities and builds a robust clip embedding using time-aware cross attention, a lightweight fusion gate, and alignment-regularized contrastive objectives with optional weak supervision. On the K-EmoCon dataset, under leave-one-out cross-validation evaluation, CTAF yields higher cosine margins for matched pairs and better cross-modal token retrieval within one second, and it is competitive with the baseline on three-bin accuracy and macro-F1 while using few labels. Our contributions are a time-aware fusion mechanism that directly models correspondence, an alignment-driven self-supervised objective tailored to EEG and physiology, and an evaluation protocol that measures alignment quality itself. Our approach accounts for the coupling between the central and autonomic nervous systems in psychophysiological time series. These results indicate that CTAF is a strong step toward label-efficient, generalizable EEG-peripheral fusion under temporal asynchrony.

1 Introduction

Predicting latent psychophysiological constructs from multimodal biosignals is hard because modalities operate on different time scales and are not perfectly synchronous. Electroencephalography (EEG), electrodermal activity (EDA), blood-volume pulse (BVP), and electrocardiography (ECG) offer complementary views of central and autonomic systems, yet their latencies and dynamics differ in ways that matter. Labeled data are also scarce while models must generalize across tasks and subjects [Aristimunha et al., 2025]. Recent surveys recommend handling imperfect synchronization, supporting cross-modal retrieval, and enforcing representation invariance to improve robustness in these settings [Jiang et al., 2025, Baltrušaitis et al., 2019].

Self-supervised learning provides a route to exploit abundant unlabeled sensor data while encouraging invariance to nuisance variability. Contrastive and redundancy-reduction objectives promote stability to sensor noise and subject differences [Jiang et al., 2025, Chen et al., 2020, Grill et al., 2020, Zbontar et al., 2021, Bardes et al., 2022]. Time-series SSL further addresses local stochasticity and temporal warping through context and view consistency [Yue et al., 2022, Sermanet et al., 2018]. In physiology,

Preprint.

cross-lead and cross-timestamp pretraining has improved ECG tasks without labels, underscoring the promise of SSL in this domain [Kiyasseh et al., 2021].

Building on this foundation, we introduce Cross-Temporal Attention Fusion (CTAF), a self-supervised alignment-and-fusion module for synchronized yet imperfectly aligned multimodal physiological time series. CTAF learns soft cross-temporal correspondences between modalities instead of assuming frame-level synchrony, and it couples these alignments with contrastive self-supervision and redundancy-reduction to avoid collapse [Sermanet et al., 2018, Wang et al., 2019, Chen et al., 2020, Radford et al., 2021, Bardes et al., 2022, Zbontar et al., 2021]. The design is mask-aware for missing tokens and label-efficient, and the same encoders can be probed with supervised heads when labels are available without architectural changes [Bardes et al., 2022].

We evaluate CTAF under subject-wise leave-one-out cross-validation and compare against strong hypercomplex baselines [Lopez et al., 2023, 2024a,b]. Empirically, CTAF improves cross-modal retrieval under a temporal tolerance and increases the separation between matched and mismatched clip embeddings, and it is competitive on discretized recognition while using the same preprocessed inputs and splits. Our contributions are a timing-aware self-supervised fusion module that explicitly handles cross-modal lags, an evaluation protocol centered on alignment and cross-subject generalization, and a reproducible baseline suite for multimodal time-series learning [Jiang et al., 2025]

2 Related Works

Self-supervised representation learning for time series. In self-supervised learning (SSL), models are trained on tasks where the supervision signal is derived from the input data itself, without requiring labels. Self-supervised learning for time series has advanced rapidly in recent years. Contrastive predictive coding [van den Oord et al., 2019] proposed SSL for sequences. Subsequent work refined this with temporal neighborhoods coding [Tonekaboni et al., 2021], hierarchical and temporal contrasts [Yue et al., 2022], and temporal/contextual contrasting [Eldele et al., 2021]. CTAF builds on this line of research, but selects positive pairs using learned cross-temporal alignment rather than strictly synchronous counterparts.

Brain and Body neurophysiological responses. The central and the autonomic nervous systems are tightly coupled but operate on different time scales. For example, EEG responses occur within tens of milliseconds, phasic EDA responses time by several seconds, and HRV changes evolve more gradually. Despite these differences, the brain and body function in a coordinated and adaptive manner [Benedek and Kaernbach, 2010, Thayer and Lane, 2000]. Our approach accounts for this brain–body asynchrony through time-aware fusion.

Cross-modal fusion and alignment. Cross-modal fusion and alignment are central challenges in multimodal learning. Classical multi-view correlation approaches, such as Deep Canonical Correlation Analysis [Andrew et al., 2013], and multimodal Transformers have addressed unaligned input streams. For example, MulT applies directional cross-modal attention to handle unaligned sequences [Tsai et al., 2019]. CTAF differs by explicitly learning lag distributions via cross-temporal attention, and leveraging these learned alignments to guide the self-supervised learning objective.

Differentiable alignment and cycle consistency. Soft-DTW offers a differentiable time-warping loss for sequences [Cuturi and Blondel, 2018]; temporal cycle-consistency methods in video have been used to learn correspondences without labels [Dwibedi et al., 2019, Wang et al., 2019]. CTAF borrows the same intuition, imposing bidirectional EEG $\leftrightarrow$ peripheral measures consistency to regularize alignment maps.

Multimodal electrophysiology with contrastive/attention fusion. Recent work has applied SSL and attention mechanisms to EEG in combination with EOG, HR, or EDA signals for sleep staging or emotion recognition, typically assuming synchronicity and fixed windows [Wang et al., 2023, Guo et al., 2023]. CTAF learns soft, time-shifted correspondences to select cross-modal positives, producing generalizable embeddings that reflect heterogeneous, but relevant, temporal dynamics across modalities.

Multimodal affect datasets (EEG + peripheral measures). Community benchmarks pairing EEG with autonomic/peripheral signals (BVP/ECG/EDA), notably DEAP, AMIGOS, WESAD, and K-EmoCon, established realistic cross-modal timing issues and subject variability that motivate

time-aware fusion and evaluation[Koelstra et al., 2012, Park et al., 2020, Miranda-Correa et al., 2021, Schmidt et al., 2018]. In this work, we use K-EmoCon as our primary dataset.

3 Cross-Temporal Attention Fusion (CTAF)

CTAF is a self-supervised fusion module for EEG and peripheral physiology that learns time-aware cross-temporal correspondences and a joint clip representation. For notational consistency we use superscripts (e) for EEG and (p) for peripheral physiology throughout. Let

$$X^{(e)} \in \mathbb{R}^{S_e \times D_e}, \qquad X^{(p)} \in \mathbb{R}^{S_p \times D_p}$$

denote per-bin feature sequences with timestamp vectors

$$t^{(e)} \in \mathbb{R}^{S_e}, \qquad t^{(p)} \in \mathbb{R}^{S_p},$$

and validity masks

$$m^{(e)} \in \{0,1\}^{S_e}, \qquad m^{(p)} \in \{0,1\}^{S_p}.$$

CTAF outputs (i) soft cross-temporal alignment matrices $A_{e\to p} \in \mathbb{R}^{S_e \times S_p}$ and $A_{p\to e} \in \mathbb{R}^{S_p \times S_e}$ that capture token-level correspondences despite unknown lags, and (ii) a robust clip-level embedding $z_f \in \mathbb{R}^d$ (along with modality projections $z^{(e)}, z^{(p)} \in \mathbb{R}^d$) suitable for subject-generalizable downstream use.

3.1 Preliminaries and Problem Setup

We consider paired sequences $\big(X^{(e)}, t^{(e)}, m^{(e)}\big)$ and $\big(X^{(p)}, t^{(p)}, m^{(p)}\big)$ from the same window, sampled at different effective rates with possible dropouts. CTAF has two goals: (1) learn modality-invariant clip embeddings $z^{(e)}$ and $z^{(p)}$ that coincide for the same window, and (2) learn token-level correspondences between $X^{(e)}$ and $X^{(p)}$ that allow variable, sample-dependent lags.

Misalignment is inherent in EEG–autonomic coupling, so CTAF models soft, time-varying correspondences rather than assuming synchrony, learns token- and clip-level links jointly, and forms time-aware positives for contrastive learning instead of enforcing simultaneity. The design is label-efficient and mask-aware, respecting $m^{(e)}$ and $m^{(p)}$ for missing tokens, and it avoids heavy alignment path solvers, which keeps training scalable under realistic cross-modal timing noise.

3.2 CTAF Architecture

Modality encoders. Each stream is encoded with a lightweight Conv-Transformer that is both time-aware and mask-aware. With inputs $X^{(e)} \in \mathbb{R}^{S_e \times D_e}$, $X^{(p)} \in \mathbb{R}^{S_p \times D_p}$, timestamps $t^{(e)}, t^{(p)}$ and validity masks $m^{(e)}, m^{(p)}$,

$$H^{(e)} = \mathrm{Enc}_e\big(X^{(e)},\, \phi(t^{(e)});\ \mathrm{kpm} = \neg m^{(e)}\big), \qquad H^{(p)} = \mathrm{Enc}_p\big(X^{(p)},\, \phi(t^{(p)});\ \mathrm{kpm} = \neg m^{(p)}\big),$$

where $\phi(\cdot)$ are sinusoidal time features (added or concatenated), and "kpm" is the key-padding mask used to prevent attention to invalid tokens. Encoders output $H^{(e)} \in \mathbb{R}^{S_e \times d}$, $H^{(p)} \in \mathbb{R}^{S_p \times d}$.

Bidirectional cross-temporal attention. To expose cross-modal lags, we apply multi-head attention in both directions:

$$\begin{aligned} \tilde{H}^{(e)} &= \mathrm{MHA}\Big(Q{=}H^{(e)},\, K{=}H^{(p)},\, V{=}H^{(p)};\ \mathrm{kpm} = \neg m^{(p)}\Big), \\ \tilde{H}^{(p)} &= \mathrm{MHA}\Big(Q{=}H^{(p)},\, K{=}H^{(e)},\, V{=}H^{(e)};\ \mathrm{kpm} = \neg m^{(e)}\Big). \end{aligned} \tag{1}$$

The resulting attention maps are the soft cross-temporal alignments $A_{e\to p} = \mathrm{softmax}(Q_e K_p^\top/\sqrt{d})$ and $A_{p\to e}$, with invalid keys masked to $-\infty$.

Global summaries, fusion gate, and token pooling. We use two explicit pooling operators:

(1) Masked mean pooling for global summaries:

$$\mathrm{Mean}_m(H) = \frac{\sum_i m_i H_i}{\sum_i m_i + \varepsilon}, \qquad z^{(e)} = \mathrm{Mean}_{m^{(e)}}(\tilde{H}^{(e)}), \;\; z^{(p)} = \mathrm{Mean}_{m^{(p)}}(\tilde{H}^{(p)}).$$

(2) Masked attention pooling for fused tokens: we first form $T = \frac{1}{2}(\tilde{H}^{(e)} + \tilde{H}^{(p)})$ with union mask $m^{(\cup)} = m^{(e)} \vee m^{(p)}$. With a learned query $q \in \mathbb{R}^d$,

$$\alpha_i \propto \exp\Big(\tfrac{q^\top T_i}{\sqrt{d}}\Big) \text{ for valid } i, \quad \alpha_i{=}0 \text{ if } m_i^{(\cup)}{=}0, \quad z_{\text{tok}} = \sum_i \alpha_i T_i.$$

A lightweight fusion gate mixes the modality summaries in a data-dependent way:

$$g = \sigma\big(\mathrm{MLP}([z^{(e)} \,\|\, z^{(p)}])\big), \qquad z_{\text{gate}} = g \odot z^{(e)} + (1-g) \odot z^{(p)}.$$

The final clip embedding is

$$z_f = \tfrac{1}{2}\big(z_{\text{gate}} + z_{\text{tok}}\big),$$

i.e., the fusion gate combines global summaries, and an attention layer pools over fused tokens to yield the clip-level representation.

Projection heads. Three MLP heads map to contrastive space (with ℓ_2 normalization):

$$p^{(e)} = \mathrm{Proj}_e(z^{(e)}), \quad p^{(p)} = \mathrm{Proj}_p(z^{(p)}), \quad p^{(f)} = \mathrm{Proj}_f(z_f).$$

Token-level projections are used when computing alignment/retrieval losses.

Masks are enforced at every sequence-consuming step: (i) in encoders and cross-attention via key-padding masks; (ii) in global summaries via the masked mean; (iii) in attention pooling by zeroing logits of invalid positions before softmax; and (iv) in token-level losses, which operate only on valid indices (or on the union mask for fused tokens). This makes CTAF permutation-invariant over valid tokens, explicitly time-aware via $\phi(t)$, and robust to missing data.

3.3 Learning Objectives and Training Strategy

CTAF couples complementary terms that bind the modalities semantically, learn time-aware token matches, and stabilize the representation.

Clip-Level Cross-Modal Contrast (InfoNCE [Chen et al., 2020]). Let $z_e, z_p \in \mathbb{R}^d$ be L^2-normalized clip projections for EEG and physiology. With temperature T and batch index i,

$$\mathcal{L}_{\text{con}} = \tfrac{1}{2}\sum_i \Big[-\log \frac{\exp(\langle z_e^i, z_p^i\rangle/T)}{\sum_j \exp(\langle z_e^i, z_p^j\rangle/T)} - \log \frac{\exp(\langle z_p^i, z_e^i\rangle/T)}{\sum_j \exp(\langle z_p^i, z_e^j\rangle/T)} \Big].$$

Soft Cross-Temporal Alignment. Project token sequences to $\tilde{H}^{\text{eeg}}, \tilde{H}^{\text{phys}}$ and normalize tokens. For sample i, token similarity $S_{su}^i = \langle \tilde{h}_s^{\text{eeg}}, \tilde{h}_u^{\text{phys}}\rangle/T$. Define soft targets by a Gaussian over time deltas,

$$W_{su}^i \propto \exp\Big(-\tfrac{(t_s^{\text{eeg}} - t_u^{\text{phys}})^2}{2\sigma^2}\Big), \quad \sum_u W_{su}^i = 1.$$

We minimize row- and column-wise cross-entropies:

$$\mathcal{L}_{\text{align}} = \tfrac{1}{2}\sum_i \Big[-\sum_s \sum_u W^i_{su}\, \log \text{softmax}_u(S^i_{su}) - \sum_u \sum_s \hat{W}^i_{us}\, \log \text{softmax}_s(S^i_{su}) \Big],$$

with $\hat{W}$ the column-normalized transpose, encouraging near-synchronous matches while tolerating realistic lags.

Fusion. Let z_f be the fused clip projection. We tether it to the modalities with

$$\mathcal{L}_{\text{fuse}} = \big\| z_f - \tfrac{1}{2}(z_e + z_p) \big\|_2^2.$$

VICReg-Style Stabilization. On pre-normalized projections p_e, p_p, we use invariance, variance, and covariance penalties:

$$\mathcal{L}_{\text{inv}} = \|p_e - p_p\|_2^2, \quad \mathcal{L}_{\text{var}} = \tfrac{1}{2} \sum_{m \in \{e,p\}} \text{mean}\big[\max(0, 1 - \text{std}(p_m))\big],$$

$$\mathcal{L}_{\text{cov}} = \tfrac{1}{2} \sum_{m \in \{e,p\}} \text{mean}\big[\text{offdiag}(\text{Cov}(p_m))^2\big].$$

Within-Modality View Contrast. Two stochastic time views per stream (warp/jitter/mask) yield $\mathcal{L}_{\text{view}}$ via symmetric InfoNCE [Chen et al., 2020] within each modality, improving robustness to sampling and segmentation.

Consistency Under Time Jitter. Re-encode the same window with different time jitters; penalize drift of z_f:

$$\mathcal{L}_{\text{cons}} = \|z_f^{(1)} - z_f^{(2)}\|_2^2.$$

Auxiliary Supervision. When window-level labels $y \in \mathbb{R}^2$ (arousal, valence) are available, a lightweight MLP head on the fused clip representation z_f predicts $\hat{y}$. We train this head with a small weight using z-scored targets

$$y_z \;=\; (y - \mu) \oslash \sigma, \qquad \mu, \sigma \in \mathbb{R}^2$$

where μ and σ are the per-dimension mean and standard deviation computed on the training windows of the current fold, and $\oslash$ denotes element-wise division. The loss is

$$\mathcal{L}_{\text{sup}} \;=\; \|\hat{y} - y_z\|_2^2,$$

applied only to windows with valid labels; this term is auxiliary and does not drive the main alignment objective.

Total Objective. With scheduled weights $\beta_t, \alpha_f, \lambda_{\text{inv}}, \lambda_{\text{var}}, \lambda_{\text{cov}}, \lambda_{\text{view}}, \lambda_{\text{cons}}, \lambda_{\text{sup}}$,

$$\mathcal{L}_{\text{CTAF}} = \mathcal{L}_{\text{con}} + \beta_t\, \mathcal{L}_{\text{align}} + \alpha_f\, \mathcal{L}_{\text{fuse}} + \lambda_{\text{inv}} \mathcal{L}_{\text{inv}} + \lambda_{\text{var}} \mathcal{L}_{\text{var}} + \lambda_{\text{cov}} \mathcal{L}_{\text{cov}} + \lambda_{\text{view}} \mathcal{L}_{\text{view}} + \lambda_{\text{cons}} \mathcal{L}_{\text{cons}} + \lambda_{\text{sup}} \mathcal{L}_{\text{sup}}.$$

3.4 Other Design Choices

We use a short curriculum: the time-jitter amplitude and the weight β_t on $\mathcal{L}_{\text{align}}$ increase over training. The model first binds modalities at the clip level, then sharpens token-level lag handling. Robustness to missing sensors is encouraged via modality dropout—randomly zeroing an entire stream—together with mask-aware pooling.

CTAF is trained largely self-supervised to learn a time-aware shared latent. Dual Transformer encoders with sinusoidal time features produce modality sequences; bi-directional cross-attention exchanges information; a lightweight fusion gate mixes global EEG and physiology summaries;

Table 1: Baseline comparison (3-bin classification). Values are reported as mean ± 95% confidence interval across LOOCV participants.

Metric	**CTAF (ours)**	**HyperFuseNet**
Accuracy	**0.62** ± 0.04	0.58 ± 0.05
Macro-F1	**0.61** ± 0.04	0.57 ± 0.05

and attention pooling over fused tokens yields a clip embedding. The objective combines bidirectional InfoNCE, a soft temporal-alignment loss that tolerates realistic offsets, a VICReg regularizer (invariance/variance/covariance), within-modality view contrast via jitter/warp perturbations, and a consistency term for time-augmented encodings. A small supervised head provides a weak regression signal for arousal/valence. Architecturally we rely on standard self-attention blocks [Vaswani et al., 2023] and modern self-supervised components [van den Oord et al., 2019, Chen et al., 2020, Bardes et al., 2022] adapted to cross-modal, temporally misaligned data.

4 Empirical Evaluation

We evaluate CTAF on K-EmoCon, which provides continuous 5-second arousal–valence annotations with peripheral physiology from Empatica E4 and EEG from NeuroSky [Park et al., 2020]. All methods use identical fixed-length windows from our preprocessing, where we bin each stream into equal-width tokens, compute a 10-channel EEG feature vector and a 4-channel physiology vector covering BVP, EDA, skin temperature, and heart rate, attach the window-level self-report, and apply subject-wise per-channel normalization. We further require a minimum joint coverage of valid EEG and physiology tokens to limit label leakage, and we batch windows without altering their content.

We evaluate with subject-level leave-one-out cross-validation (LOOCV). For each participant we train on the remaining $N - 1$ with fixed hyperparameters, choose the CTAF checkpoint that minimize the validation self-supervised objective, and then test on the held-out subject. Our main alignment metrics are clip-level cosine similarity comparing matched to mismatched EEG-physiology pairs and token-level cross-modal retrieval accuracy within a tolerance τ seconds in both EEG→phys and phys→EEG directions, computed both with and without time features to assess learning beyond explicit timing. For interpretability we also report accuracy and macro-F1 on three pre-registered bins for arousal and valence.

For the baseline, we compare our CTAF against HyperFuseNet, a parameter-efficient hypercomplex fusion network that replaces large real-valued layers with parameterized hypercomplex multiplication and quaternionic operators to exploit inter-channel structure [Lopez et al., 2024a, Andrew et al., 2013, Lopez et al., 2023]. We re-implement the HyperFuseNet head and its modality branches, adapt inputs to our windowed layout, and train it in a supervised manner using the official schedule. Following prior work, we report three-bin accuracy and macro-F1 for HyperFuseNet, which are computed on the same windows and LOOCV splits.

5 Results

Baseline comparison. Under identical preprocessing and LOOCV splits, CTAF outperforms HyperFuseNet on three-bin arousal/valence recognition. Table 1 shows higher accuracy and macro-F1 for CTAF, indicating that its time-aware embeddings support stronger discretized recognition.

Cross-modal alignment. Under LOOCV, CTAF consistently learns time-aware EEG–physiology correspondence. In Table 2 Panel A, matched clips cluster well above mismatches (cosine 0.240 $[0.219, 0.259]$ vs 0.051 $[0.033, 0.067]$), and Figure 1 shows this separation across participants. Figure 2 shows most points above the diagonal, indicating consistent gains across subjects. The effect persists when absolute time is removed (Table 2, Figure 2), with a median per-participant gain of 0.182 (IQR $[0.141, 0.224]$), indicating the model captures correspondence beyond explicit timestamps.

Clip-level cosine. Figure 1 summarizes per-participant cosine similarity for matched EEG-physiology clips versus mismatched pairs. With time encodings enabled, matched pairs trend higher; Figure 2

Table 2: Cross-modal alignment under LOOCV on K-EmoCon (τ=1 s). Macro means and 95% CIs are reported across held-out participants. Metrics: "cos_pos" (matched EEG–physiology clips), "cos_neg" (mismatched clips), and token retrieval rates "retr@ τ e2p" (EEG→physiology) and "retr@ τ p2e" (physiology→EEG). The lower panel shows absolute improvements when enabling time features (WITH time minus NO time).

Macro means [95% CI]		
Metric	WITH time	NO time
cos_pos	0.240 [0.219, 0.259]	0.051 [0.033, 0.067]
cos_neg	0.004 [-0.007, 0.015]	0.003 [-0.008, 0.014]
retr@τ e2p	0.350 [0.334, 0.366]	0.212 [0.207, 0.220]
retr@τ p2e	0.265 [0.247, 0.284]	0.208 [0.206, 0.210]

WITH time − NO time (absolute gain)		
Metric	Δ mean	95% CI
cos_pos	+0.189	[0.163, 0.215]
cos_neg	+0.001	[-0.010, 0.012]
retr@τ e2p	+0.138	[0.119, 0.156]
retr@τ p2e	+0.057	[0.038, 0.075]

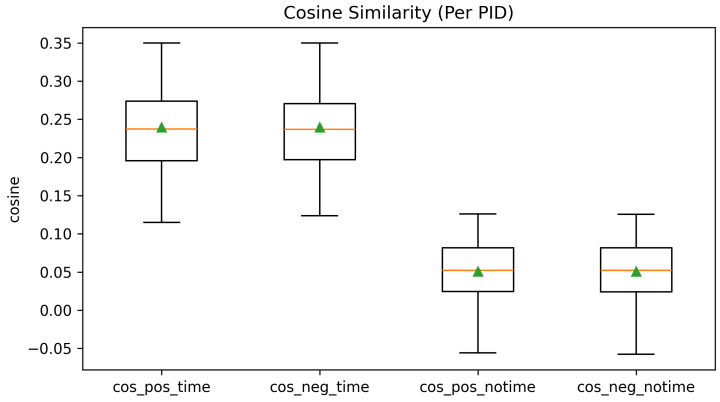

Figure 1: Clip-level cosine similarity distributions for matched (cos_pos) vs. mismatched (cos_neg) EEG–physiology pairs, aggregated over LOOCV folds. Larger separation indicates stronger cross-modal alignment; boxes show median and IQR across participants (PIDs).

shows per-subject with time in the y-axis against no time in the x-axis, and most points lie above the diagonal, indicating that CTAF extracts stronger correspondences when it can exploit temporal features.

Token-level retrieval. Figure 3 reports cross-modal token retrieval within a tolerance τ seconds (EEG→physiology and physiology→EEG). Token retrieval is above chance and asymmetric in the expected direction—EEG→phys 0.350 and phys→EEG 0.265 at τ = 1 s—both exceeding the no-time setting (0.212 and 0.208). These rates increase substantially with time encodings, directly demonstrating learned cross-temporal correspondence. Interestingly, EEG→physiology and physiology→EEG rates are higher with time features than without, as summarized in Table 2. Panel A and visualized in Figure 3. The asymmetry favoring EEG→physiology aligns with expected autonomic lags and further indicates that CTAF captures plausible cross-temporal correspondences.

Together, these results support the core claim that CTAF achieves robust, subject-generalizable cross-modal alignment.

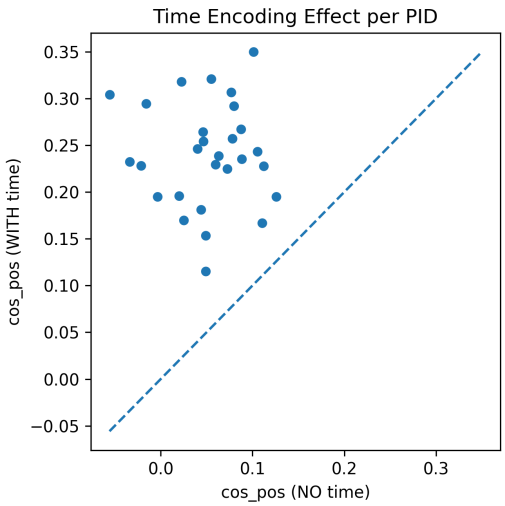

Figure 2: Per-participant (PID) matched cosine (cos_pos) *WITH time (CTAF)* vs. *NO time* ablation. Points above the diagonal indicate improved alignment when using CTAF's time-aware fusion.

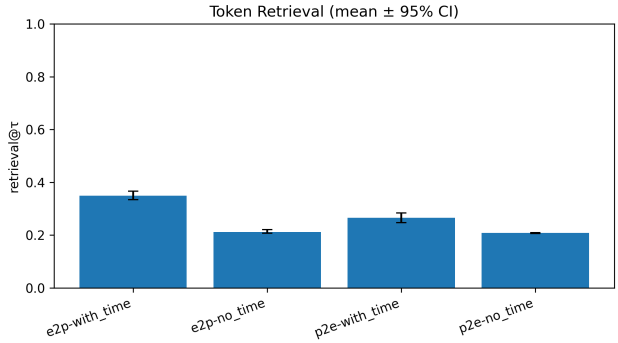

Figure 3: Cross-modal token-retrieval accuracy within tolerance τ seconds for EEG→Phys and Phys→EEG. Bars compare *WITH time (CTAF)* to *NO time* ablation; error bars show 95% bootstrap CIs across participants.

6 Discussion and Conclusion

CTAF learns cross-temporal correspondences rather than assuming synchrony, estimating soft lag distributions between EEG and peripheral signals to construct contrastive positives, stabilize representations, and produce a shared clip embedding. In sensitivity analyses, we observed that the learned attention over lags can be broad for some held-out participants, consistent with consumer-grade EEG and window-level labels. We also found retrieval depends on the tolerance τ and on time encodings by tightening τ from 1.0 s to 0.5 s reduces retrieval via near-misses, and time-aware models consistently outperform no-time variants. These behaviors suggest straightforward extensions, such as sharper priors or temperature schedules to improve identifiability, principled τ selection, and alternative time encodings.

CTAF provides a label-efficient, mask-safe mechanism for time-aware fusion that directly optimizes and evaluates cross-modal alignment. On the K-EmoCon dataset, it yields stronger alignment than time-agnostic variants and compares favorably to a supervised HyperFuseNet baseline while targeting a different objective. We see CTAF as a useful building block for brain–body representation learning, which handles asynchrony, scales to weak labels, and offers diagnostics aligned with the fusion goal. While our evaluation is limited to one dataset, in future work, we will validate generalizability on broader benchmarks and address the computational cost of attention mechanisms.

References

G. Andrew, R. Arora, J. Bilmes, and K. Livescu. Deep canonical correlation analysis. In S. Dasgupta and D. McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1247–1255, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL https://proceedings.mlr.press/v28/andrew13.html.

B. Aristimunha, D. Truong, P. Guetschel, S. Y. Shirazi, I. Guyon, A. R. Franco, M. P. Milham, A. Dotan, S. Makeig, A. Gramfort, J.-R. King, M.-C. Corsi, P. A. Valdés-Sosa, A. Majumdar, A. Evans, T. J. Sejnowski, O. Shriki, S. Chevallier, and A. Delorme. Eeg foundation challenge: From cross-task to cross-subject eeg decoding, 2025. URL https://arxiv.org/abs/2506.19141.

T. Baltrušaitis, C. Ahuja, and L.-P. Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, 2019. doi: 10.1109/TPAMI.2018.2798607.

A. Bardes, J. Ponce, and Y. LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning, 2022. URL https://arxiv.org/abs/2105.04906.

M. Benedek and C. Kaernbach. A continuous measure of phasic electrodermal activity. *Journal of Neuroscience Methods*, 190(1):80–91, 2010. ISSN 0165-0270. doi: https://doi.org/10.1016/j.jneumeth.2010.04.028. URL https://www.sciencedirect.com/science/article/pii/S0165027010002335.

T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations, 2020. URL https://arxiv.org/abs/2002.05709.

M. Cuturi and M. Blondel. Soft-dtw: a differentiable loss function for time-series, 2018. URL https://arxiv.org/abs/1703.01541.

D. Dwibedi, Y. Aytar, J. Tompson, P. Sermanet, and A. Zisserman. Temporal cycle-consistency learning, 2019. URL https://arxiv.org/abs/1904.07846.

E. Eldele, M. Ragab, Z. Chen, M. Wu, C. K. Kwoh, X. Li, and C. Guan. Time-series representation learning via temporal and contextual contrasting, 2021. URL https://arxiv.org/abs/2106.14112.

J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko. Bootstrap your own latent: A new approach to self-supervised learning, 2020. URL https://arxiv.org/abs/2006.07733.

Y. Guo, T. Zhang, and W. Huang. Emotion recognition based on multi-modal electrophysiology multi-head attention contrastive learning, 2023. URL https://arxiv.org/abs/2308.01919.

Y. Jiang, K. Ning, Z. Pan, X. Shen, J. Ni, W. Yu, A. Schneider, H. Chen, Y. Nevmyvaka, and D. Song. Multi-modal time series analysis: A tutorial and survey, 2025. URL https://arxiv.org/abs/2503.13709.

D. Kiyasseh, T. Zhu, and D. A. Clifton. Clocs: Contrastive learning of cardiac signals across space, time, and patients, 2021. URL https://arxiv.org/abs/2005.13249.

S. Koelstra, C. Mühl, M. Soleymani, J. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras. Deap: A database for emotion analysis using physiological signals. *IEEE transactions on affective computing*, 3(1):18–31, 2012. ISSN 1949-3045. doi: 10.1109/T-AFFC.2011.15. eemcs-eprint-21368.

E. Lopez, E. Chiarantano, E. Grassucci, and D. Comminiello. Hypercomplex multimodal emotion recognition from eeg and peripheral physiological signals. In *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pages 1–5, 2023. doi: 10.1109/ICASSPW59220.2023.10193329.

E. Lopez, A. Uncini, and D. Comminiello. Hierarchical hypercomplex network for multimodal emotion recognition, 2024a. URL https://arxiv.org/abs/2409.09194.

E. Lopez, A. Uncini, and D. Comminiello. Phemonet: A multimodal network for physiological signals, 2024b. URL https://arxiv.org/abs/2410.00010.

J. A. Miranda-Correa, M. K. Abadi, N. Sebe, and I. Patras. Amigos: A dataset for affect, personality and mood research on individuals and groups. *IEEE Transactions on Affective Computing*, 12(2): 479–493, 2021. doi: 10.1109/TAFFC.2018.2884461.

C. Y. Park, N. Cha, S. Kang, A. Kim, A. H. Khandoker, L. Hadjileontiadis, A. Oh, Y. Jeong, and U. Lee. K-emocon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations. *Scientific Data*, 7(1), Sept. 2020. ISSN 2052-4463. doi: 10.1038/s41597-020-00630-y. URL http://dx.doi.org/10.1038/s41597-020-00630-y.

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. Van Laerhoven. Introducing wesad, a multimodal dataset for wearable stress and affect detection. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, ICMI '18, page 400–408, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356923. doi: 10.1145/3242969.3242985. URL https://doi.org/10.1145/3242969.3242985.

P. Sermanet, C. Lynch, Y. Chebotar, J. Hsu, E. Jang, S. Schaal, and S. Levine. Time-contrastive networks: Self-supervised learning from video, 2018. URL https://arxiv.org/abs/1704.06888.

J. F. Thayer and R. D. Lane. A model of neurovisceral integration in emotion regulation and dysregulation. *Journal of Affective Disorders*, 61(3):201–216, 2000. ISSN 0165-0327. doi: https://doi.org/10.1016/S0165-0327(00)00338-4. URL https://www.sciencedirect.com/science/article/pii/S0165032700003384. Arousal in Anxiety.

S. Tonekaboni, D. Eytan, and A. Goldenberg. Unsupervised representation learning for time series with temporal neighborhood coding, 2021. URL https://arxiv.org/abs/2106.00750.

Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In A. Korhonen, D. Traum, and L. Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1656. URL https://aclanthology.org/P19-1656/.

A. van den Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding, 2019. URL https://arxiv.org/abs/1807.03748.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need, 2023. URL https://arxiv.org/abs/1706.03762.

X. Wang, A. Jabri, and A. A. Efros. Learning correspondence from the cycle-consistency of time, 2019. URL https://arxiv.org/abs/1903.07593.

Y. Wang, H. Liang, and B. Zhai. Temporal neighborhood based self-supervised pre-training model for sleep stages classification. In *Proceedings of the 2023 15th International Conference on Bioinformatics and Biomedical Technology*, ICBBT '23, page 149–155, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400700385. doi: 10.1145/3608164.3608185. URL https://doi.org/10.1145/3608164.3608185.

Z. Yue, Y. Wang, J. Duan, T. Yang, C. Huang, Y. Tong, and B. Xu. Ts2vec: Towards universal representation of time series, 2022. URL https://arxiv.org/abs/2106.10466.

J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny. Barlow twins: Self-supervised learning via redundancy reduction, 2021. URL https://arxiv.org/abs/2103.03230.