iSyncTab: Learning Cross-Modal Feature Sequencing for Image-Tabular Data via Neural Synchrony

*Al Zadid Sultan Bin Habib[1], Md Younus Ahamed[1], Prashnna Kumar Gyawali[1], Gianfranco Doretto[2], and *Donald A. Adjeroh[1]

[1] LCSEE, West Virginia University, Morgantown, WV 26506, USA
{ah00069,ma00087}@mix.wvu.edu,{prashnna.gyawali,donald.adjeroh}@mail.wvu.edu
[2] SCI Institute and DBMI, The University of Utah, Salt Lake City, UT 84112, USA
doretto@utah.edu

Abstract. Multimodal learning of image and tabular data is a challenging problem, especially given the unstructured feature representations involved. We address this by framing feature sequencing as a Column Permutation Problem and showing how reordering columns can reduce dispersion and improve training stability for multimodal learning of image and tabular data. We propose iSyncTab, an approach that enforces these learned sequential constraints through a joint classification and order-consistency objective. We introduce Neural Synchrony-guided Paired Feature Sequencing (NS-PFS), a neuroscience-inspired NeuroAI approach that aligns image and tabular feature clusters through a Hungarian assignment of their synchrony matrix, derived from cluster energy and centroid similarity. The resulting cross-modal feature sequence provides a coherent feature alignment that preserves structural synchrony between modalities. We then introduce an Order-aware Memory-augmented Transformer (OMT) with a custom loss function that enforces these learned feature sequence constraints. Experiments on multimodal benchmarks demonstrate that systematically derived feature sequences consistently boost predictive performance and model robustness. Our findings underscore feature sequencing as an important inductive bias for structured multimodal learning of image and tabular data. Neural synchrony-driven sequencing enhances cross-modal coherence and representation quality for image and tabular data.

Keywords: Neural Synchrony · Image-Tabular Multimodal Learning · Cross-Modal Feature Sequencing · NeuroAI

1 Introduction

Multimodal learning, which integrates heterogeneous data such as images and tables, underpins many AI applications in healthcare [27], autonomous sys-

* Corresponding authors. Code: https://github.com/zadid6pretam/iSyncTab; Project webpage: https://www.zadidhabib.com/isynctab.html.

tems [79], and scientific discovery [61]. Despite the strong performance of modality-specific backbones e.g., convolutional neural networks (CNNs) for vision and Transformers for structured inputs [35, 53], reliably integrating their representations remains challenging. A key but underexplored factor is feature sequencing and structural alignment when tabular and visual features are combined: unlike images or text, tabular columns have no canonical order, yet recent work shows that sequencing can materially affect optimization and representation stability in tabular deep learning [44, 84]. In multimodal settings, this issue is amplified because features are often flattened and concatenated, obscuring cross-modal correspondences; sequence-sensitive operators such as self-attention and feed-forward mixing must then operate over potentially arbitrary permutations, which motivates permutation-aware modeling [108] and structured cross-modal fusion mechanisms [70].

Most recent models for tabular or multimodal learning emphasize new attention designs [82], contextual embeddings [53], ensemble-style regularization [39], and hybrid fusion pipelines. Recent multimodal methods further combine tabular and imaging signals via contrastive pretraining and shared latent transformers, using cross-hierarchical attention, interaction modules, and temporal embeddings to improve robustness and clinical prediction [15, 21, 22, 45, 56, 66]. However, they rarely treat feature sequencing as a first-class design choice. We revisit this neglected aspect by formulating sequencing as a Column Permutation Problem (CPP) [8, 28, 63, 64], aiming to find a permutation that reduces inter-feature dispersion and promotes cross-modal structural coherence. To address this, we propose Neural Synchrony-guided Paired Feature Sequencing (NS-PFS), which aligns tabular and image feature clusters via an energy-phase synchrony matrix and solves the resulting bipartite matching with a Hungarian assignment [59]. The induced synchronized ordering is then fed to an Order-aware Memory-augmented Transformer (OMT) with a Linformer backbone [92] to efficiently model long multimodal feature-token sequences.

Neural synchrony denotes the precise temporal alignment of oscillatory activity across neuronal populations, enabling efficient information transfer via phase and energy matching [30]. Under the Communication Through Coherence (CTC) hypothesis, phase-aligned assemblies communicate more effectively because their excitability windows coincide [31]. Empirically, gamma-band synchrony is linked to fast feedforward feature binding, whereas alpha/beta rhythms support longer-range predictive or modulatory interactions [4, 80]; in multisensory settings, synchrony between sensory cortices (e.g., visual and auditory) reflects crossmodal matching and supports integration [6, 60, 74]. In parallel, feature ordering or sequencing has a long history in pattern recognition, notably in Incremental Attribute Learning (IAL) where features arrive sequentially and must be ranked before training [93]. Although set-based formulations often assume order invariance [108], column order in practice affects redundancy exposure and the ability to capture dependencies; ranking by Fisher score, correlation, or entropy can reduce interference and error relative to unordered baselines [95, 97], motivating task-aware ordering strategies [94, 96]. Recent work also emphasizes that tabular

models can be brittle to column permutations, prompting permutation-invariant designs [9, 26]; for example, TabICL [77] ensembles predictions across multiple column permutations, while COPER [25] fuses images and tables via an end-to-end multi-view clustering objective based on permutation-driven correlation for clustering.

We introduce iSyncTab, a neuro-inspired multimodal image-tabular framework that uses synchrony-guided, metric-generalized feature sequencing to produce coherent, dispersion-minimized cross-modal feature sequence. Across benchmark datasets, iSyncTab consistently improves accuracy, stability, and data efficiency over unordered fusion baselines, demonstrating that feature sequencing is a strong inductive bias for multimodal deep learning and that synchrony-driven sequencing enhances cross-modal alignment and representation quality.

Key Contributions:

- We propose NS-PFS, a neuro-inspired approach which aligns feature clusters across modalities via an energy-similarity synchrony matrix and Hungarian matching.
- We design an OMT backbone that exploits NS-PFS-ordered multimodal tokens through memory-based readout and order-consistency regularization.
- We combine these to develop iSyncTab, which shows consistent performance gains and improved training stability across multimodal image-tabular benchmarks, validating the importance of structured feature sequencing in multimodal representation learning.
- We introduce an efficiency-centric evaluation and composite efficiency score that jointly benchmarks multimodal image-tabular models under accuracy, FLOPs, parameter count, and peak GPU-memory budgets, revealing the favorable accuracy-cost trade-off of iSyncTab.

2 Methodology

We address multimodal learning of tabular and image features by casting feature sequencing as a CPP. Given the fused multimodal feature matrix $X_{\text{mm}} \in \mathbb{R}^{N\times m}$ with N samples and $m = d_{\text{tab}} + d_{\text{vis}}$ features, we seek a permutation $\pi \in S_m$ that minimizes a dispersion cost in Eq. 1.

$$\min_{\pi\in S_m} D(\pi) = \sum_{1\le u<v\le m} w_{uv}\,|\pi(u) - \pi(v)| \tag{1}$$

where u and v index features, w_{uv} quantifies feature dissimilarity, and S_m is the symmetric group over $\{1,\ldots,m\}$. Fig. 1 illustrates the iSyncTab architecture.

Feature Representation and Fusion. For images, a CNN encoder (e.g., ResNet-50 [47]) extracts a visual representation $\mathbf{x}_{\text{vis}}^{(s)} \in \mathbb{R}^{d_{\text{vis}}}$ for sample s. For tabular data, a Transformer encoder f_{tab} extracts $\mathbf{x}_{\text{tab}}^{(s)} \in \mathbb{R}^{d_{\text{tab}}}$. We concatenate these representations as in Eq. 2, yielding $\mathbf{x}_{\text{mm}}^{(s)} \in \mathbb{R}^{d_{\text{tab}}+d_{\text{vis}}}$, and stack them across samples to obtain $X_{\text{mm}} = [\mathbf{x}_{\text{mm}}^{(1)};\ldots;\mathbf{x}_{\text{mm}}^{(N)}] \in \mathbb{R}^{N\times m}$, where $m = d_{\text{tab}}+d_{\text{vis}}$.

$$\mathbf{x}_{\text{mm}}^{(s)} = \text{concat}\left(\mathbf{x}_{\text{tab}}^{(s)}, \mathbf{x}_{\text{vis}}^{(s)}\right) \in \mathbb{R}^{d_{\text{tab}}+d_{\text{vis}}} \tag{2}$$

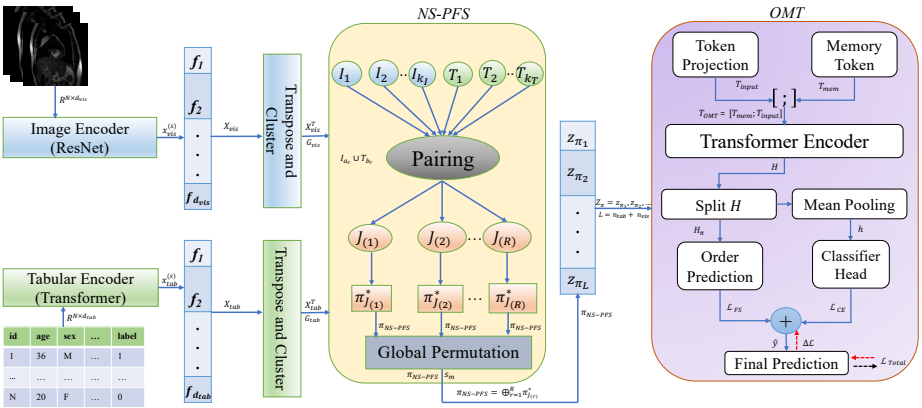


Fig. 1: End-to-end iSyncTab architecture. Image and tabular encoders produce modality-specific feature matrices, which NS-PFS transposes, clusters, and pairs using energy and phase-analogue synchrony. NS-PFS solves local CPPs over matched joint clusters $J_{(r)}$ and concatenates the resulting local orders into the global permutation $\pi_{\text{NS-PFS}} = \bigoplus_{r=1}^{R} \pi^{\star}_{J_{(r)}}$, where $R = \min(k_I, k_T)$. The ordered multimodal token sequence $\mathbf{Z}_\pi$ is processed by OMT and trained with $\mathcal{L}_{\text{CE}}$ and $\mathcal{L}_{\text{FS}}$.

A: Neural Synchrony-guided Paired Feature Sequencing (NS-PFS). We introduce a biologically inspired, metric-generalized framework that aligns feature clusters across modalities via the principle of neural synchrony. In neuroscience, neural synchrony refers to the temporal coordination of oscillatory activity among distributed neuronal populations, where information transfer is maximized when oscillations are both phase-aligned and energy-balanced [30, 31, 80]. According to the Communication Through Coherence (CTC) hypothesis, assemblies oscillating in compatible phases exhibit efficient bidirectional communication, whereas desynchronized populations exchange only modulatory signals [4].

Metric-Generalized Objective. Given fused multimodal features $X_{\text{mm}} \in \mathbb{R}^{N\times m}$, NS-PFS instantiates the CPP in Eq. 1 by defining feature dissimilarities as Eq. 3.

$$w_{uv} = |M(f_u) - M(f_v)| \tag{3}$$

where f_u and f_v denote feature columns indexed by u and v, and $M(\cdot)$ is a feature-wise metric/divergence (e.g., Kullback-Leibler (KL) or Jensen-Shannon (JS) divergence, variance, or distance-based measures such as Euclidean, Manhattan, cosine, or correlation). This yields a metric-generalized dispersion objective while keeping the CPP form unchanged.

Cross-Modal Clustering. We perform clustering on modality-specific encoder features before constructing the synchronized sequence. Let $X_{\text{tab}} \in \mathbb{R}^{N\times d_{\text{tab}}}$ and $X_{\text{vis}} \in \mathbb{R}^{N\times d_{\text{vis}}}$ denote the tabular and visual feature matrices, respectively. We cluster the transposed feature matrices $G_{\text{tab}} = X_{\text{tab}}^{\top} \in \mathbb{R}^{d_{\text{tab}}\times N}$ and $G_{\text{vis}} = X_{\text{vis}}^{\top} \in \mathbb{R}^{d_{\text{vis}}\times N}$ into tabular clusters $\{T_1, \ldots, T_{k_T}\}$ and visual clusters $\{I_1, \ldots, I_{k_I}\}$, respectively. Each cluster C (either T_b or I_a) is defined by Eq. 4.

$$\xi_C = \sum_{f \in C} M(f), \qquad \mu_C = \frac{1}{|C|} \sum_{f \in C} \phi(f) \tag{4}$$

where ξ_C is the cumulative energy under metric $M(\cdot)$, $\mu_C \in \mathbb{R}^p$ is the cluster centroid, and $\phi(f) \in \mathbb{R}^p$ is a feature embedding, such as a learned projection or normalized feature column. Here, p is the embedding dimension used for clustering and synchrony computation.

Synchrony Score (Energy-Phase Alignment). For each visual cluster I_a and tabular cluster T_b, where $a = 1, \ldots, k_I$ and $b = 1, \ldots, k_T$, we define E_{ab}, S_{ab}, and their product Ψ_{ab} in Eq. 5 and Eq. 6.

$$E_{ab} = 1 - \frac{|\xi_{I_a} - \xi_{T_b}|}{\xi_{I_a} + \xi_{T_b} + \epsilon}, \qquad S_{ab} = \frac{\langle \mu_{I_a}, \mu_{T_b} \rangle}{\|\mu_{I_a}\| \, \|\mu_{T_b}\|} \tag{5}$$

$$\Psi_{ab} = E_{ab} S_{ab} \tag{6}$$

Here, E_{ab} measures energy coherence, S_{ab} captures phase-analogue centroid similarity, and their product Ψ_{ab} quantifies overall synchrony. Ψ_{ab} is maximal when both energy and direction align, akin to coherence or phase-locking value (PLV) in neural systems [4, 60].

Cross-Modal Bipartite Matching. We construct a complete bipartite graph between visual clusters $\{I_1, \ldots, I_{k_I}\}$ and tabular clusters $\{T_1, \ldots, T_{k_T}\}$, with edge weights Ψ_{ab}. The optimal one-to-one cluster pairing $\mathcal{P}^\star$ maximizes total synchrony by Eq. 7.

$$\mathcal{P}^\star = \arg \max_{\mathcal{P} \subseteq \{1,\ldots,k_I\} \times \{1,\ldots,k_T\}} \sum_{(a,b) \in \mathcal{P}} \Psi_{ab} \tag{7}$$

Subject to one-to-one assignment, each visual cluster I_a and each tabular cluster T_b is used at most once. We solve the matching with the Hungarian algorithm [59] in $\mathcal{O}(V^3)$ time, where $V = k_I + k_T$ is the number of clusters, not the number of features. In all our settings, $k_I, k_T \leq 15$, so this cost is negligible relative to encoder and OMT computation. Let $R = |\mathcal{P}^\star| = \min(k_I, k_T)$ denote the number of matched cluster pairs.

For each matched pair $(a, b) \in \mathcal{P}^\star$, we form a joint cluster by Eq. 8.

$$J_{(a,b)} = I_a \cup T_b \tag{8}$$

$$\alpha_{J_{(a,b)}} = \frac{\sum_{f \in J_{(a,b)}} M(f)}{\sum_{(a',b') \in \mathcal{P}^\star} \sum_{f \in J_{(a',b')}} M(f)} \tag{9}$$

Here, $\alpha_{J_{(a,b)}}$ in Eq. 9 is the normalized importance score of the joint cluster. We rank the matched joint clusters by this score and denote the ranked clusters as $J_{(1)}, \ldots, J_{(R)}$.

Regime-aware Synchrony Scaling (HH/HL/LH/LL). Let τ_H and τ_L be synchrony thresholds with $0 < \tau_L < \tau_H < 1$. Each visual-tabular cluster pair (a, b) is assigned to a regime $\rho_{ab} \in \{\mathrm{HH}, \mathrm{HL}, \mathrm{LH}, \mathrm{LL}\}$ based on the ranks of its cluster energies and synchrony score Ψ_{ab}. A regime-specific modulation

$\gamma(\rho_{ab}, \Psi_{ab})$ adjusts within-cluster dispersion weights using Eq. 10.

$$\gamma(\rho_{ab}, \Psi_{ab}) = \begin{cases} g_H(\Psi_{ab}), & \rho_{ab} = \text{HH}, \\ g_L(\Psi_{ab}), & \rho_{ab} = \text{LL}, \\ g_A(\Psi_{ab}), & \rho_{ab} \in \{\text{HL}, \text{LH}\} \end{cases} \tag{10}$$

where $g_H \geq g_A \geq g_L$, $g_H(\Psi) = \Psi$, $g_A(\Psi) = \min\{\Psi, \eta\}$, and $g_L(\Psi) = \min\{\Psi, \zeta\}$ for constants $1 \geq \eta \geq \zeta > 0$. We score each visual-tabular pair by synchrony Ψ_{ab} and energy coherence E_{ab}. Using the two cutoffs (τ_L, τ_H), pairs that are high on both criteria are assigned to the HH regime and receive the strongest coupling g_H; pairs that are low on both criteria are assigned to the LL regime and receive the weaker coupling g_L; mixed pairs are assigned to HL/LH and receive the intermediate coupling g_A. Thus, well-aligned and energy-balanced pairs influence sequencing more strongly, while weak or imbalanced pairs are down-weighted.

Local Joint Sequence. For each ranked joint cluster $J_{(r)}$, where $r = 1, \ldots, R$, let (a_r, b_r) denote its corresponding matched visual-tabular cluster pair. We define local pairwise weights by Eq. 11 and also denote Eq. 12.

$$W^J_{uv} = |M(f_u) - M(f_v)| \tag{11}$$

$$\mathcal{E}_{J_{(r)}} = \{(u, v) \in J^2_{(r)} : \; W^J_{uv} > \delta\} \tag{12}$$

Here, u and v index features inside $J_{(r)}$, and δ is a sparsification threshold applied to local pairwise feature weights. In practice, δ can be set as a fixed quantile of the nonzero weights in $J_{(r)}$. The local sequence minimizes a synchrony-weighted dispersion objective by Eq. 13.

$$\pi^\star_{J_{(r)}} = \arg\min_{\pi_{J_{(r)}}} \sum_{(u,v)\in\mathcal{E}_{J_{(r)}}} \gamma(\rho_{a_r b_r}, \Psi_{a_r b_r})\, W^J_{uv}\, |\pi_{J_{(r)}}(u) - \pi_{J_{(r)}}(v)| \tag{13}$$

Global Permutation. After computing local sequences, we concatenate the ranked joint-cluster sequences to form the global NS-PFS permutation by Eq. 14.

$$\pi_{\text{NS-PFS}} = [\, \pi^\star_{J_{(1)}} \| \pi^\star_{J_{(2)}} \| \cdots \| \pi^\star_{J_{(R)}} \,] \tag{14}$$

where $\|$ denotes ordered concatenation and $R = \min(k_I, k_T)$ is the number of matched visual-tabular cluster pairs.

Equivalently, the synchronized sequence can be viewed as a composite CPP with global weights by Eq. 15 and Eq. 16.

$$\widetilde{W}_{uv} = \begin{cases} \gamma(\rho_{a_r b_r}, \Psi_{a_r b_r})\, |M(f_u) - M(f_v)|, & \text{if } \exists r \in \{1, \ldots, R\} \text{ s.t.} f_u, f_v \in J_{(r)}, \\ |M(f_u) - M(f_v)| & \text{otherwise} \end{cases} \tag{15}$$

$$\pi_{\text{NS-PFS}} = \arg\min_{\pi \in S_m} \sum_{1 \leq u < v \leq m} \widetilde{W}_{uv}\, |\pi(u) - \pi(v)| \tag{16}$$

Here, $J_{(r)}$ denotes the r-th ranked joint cluster, and (a_r, b_r) denotes its corresponding matched visual-tabular cluster pair.

Biological Interpretation. High-High (HH) pairs correspond to "binding synchrony" (gamma-band coupling and feedforward integration); Low-Low (LL) pairs represent "contextual coherence"; High-Low and Low-High (HL/LH) pairs emulate "predictive or modulatory" dynamics akin to top-down beta/alpha versus bottom-up gamma routing [6, 62, 74]. Thus, Ψ_{ab} functions as a coherence-like synchrony term that modulates cross-modal fusion strength, linking metric-based alignment with neural communication theory. See [12, 37, 81, 88] for HH (binding/gamma), [7, 86] for HL/LH (feedforward-feedback routing), and [7, 86] for LL (long-range coherence); see also Sec. A in the supplement. The NS-PFS permutation is first derived over the synchronized multimodal feature sequence and is then applied to the corresponding multimodal token sequence used by OMT. Under the feature-as-token encoding used in iSyncTab, each ordered feature corresponds to one data token; hence the OMT sequence length satisfies $L = n_{\text{tab}} + n_{\text{vis}}$, with $L = m$ when $n_{\text{tab}} = d_{\text{tab}}$ and $n_{\text{vis}} = d_{\text{vis}}$.

B: Order-aware Memory-augmented Transformer Backbone (OMT). For downstream tasks, we extend the vanilla Transformer to a linear-complexity variant using Linformer [92] for scalability. Given tabular tokens $\mathbf{Z}_{\text{tab}} \in \mathbb{R}^{B\times n_{\text{tab}}\times d_h}$ and visual tokens $\mathbf{Z}_{\text{vis}} \in \mathbb{R}^{B\times n_{\text{vis}}\times d_h}$, we first concatenate them along the token axis as $\mathbf{Z}_{\text{mm}} = [\mathbf{Z}_{\text{tab}}; \mathbf{Z}_{\text{vis}}] \in \mathbb{R}^{B\times L\times d_h}$, where $L = n_{\text{tab}} + n_{\text{vis}}$ is the number of multimodal tokens. NS-PFS returns a permutation $\pi_{\text{NS-PFS}}$ over these L tokens. Let $\mathbf{Z}_\pi = [\mathbf{z}_{\pi_1}, \ldots, \mathbf{z}_{\pi_L}]$ denote the resulting ordered multimodal token sequence.

$$\mathbf{T}_{\text{input}} = \mathbf{Z}_\pi \in \mathbb{R}^{B\times L\times d_h} \tag{17}$$

$$\begin{aligned} \mathbf{C}_{\text{mem}} &= [\mathbf{c}_1, \ldots, \mathbf{c}_{n_{\text{mem}}}] \in \mathbb{R}^{n_{\text{mem}}\times d_h}, \\ \mathbf{T}_{\text{mem}} &= \text{Repeat}_B(\mathbf{C}_{\text{mem}}) \in \mathbb{R}^{B\times n_{\text{mem}}\times d_h}, \\ \mathbf{T}_{\text{OMT}} &= [\mathbf{T}_{\text{mem}}; \mathbf{T}_{\text{input}}] \in \mathbb{R}^{B\times (n_{\text{mem}}+L)\times d_h} \end{aligned} \tag{18}$$

Here, $\mathbf{C}_{\text{mem}}$ contains n_{mem} learnable memory tokens, and $\text{Repeat}_B(\cdot)$ broadcasts them across the batch. These memory tokens act as global readout tokens over the NS-PFS-ordered tabular and visual tokens. The Linformer encoder [92] replaces the standard Transformer to reduce attention complexity from $\mathcal{O}(L^2)$ to $\mathcal{O}(L)$ for a fixed projection dimension. We encode the OMT input as Eq. 19.

$$\mathbf{H}_{\text{OMT}} = \text{Linformer}(\mathbf{T}_{\text{OMT}}) \in \mathbb{R}^{B\times (n_{\text{mem}}+L)\times d_h} \tag{19}$$

After encoding, we split the output into memory-token states $\mathbf{H}_{\text{mem}} \in \mathbb{R}^{B\times n_{\text{mem}}\times d_h}$ and ordered data-token states $\mathbf{H}_\pi \in \mathbb{R}^{B\times L\times d_h}$. The aggregated memory representation $\mathbf{h} = \frac{1}{n_{\text{mem}}} \sum_{\ell=1}^{n_{\text{mem}}} \mathbf{H}_{\text{mem},\ell}$ feeds the classification head, while the order-prediction head is applied only to the ordered data-token states as in Eq. 20.

$$\hat{\mathbf{y}} = \text{Head}_{\text{cls}}(\mathbf{h}), \qquad \mathbf{q} = \text{Head}_{\text{ord}}(\mathbf{H}_\pi) \in \mathbb{R}^{B\times L} \tag{20}$$

The target $\boldsymbol{\beta}_\pi \in \mathbb{R}^L$ is induced by $\pi_{\text{NS–PFS}}$ via $\beta_{\pi,\ell} = \frac{\ell-1}{L-1}$ for $\ell = 1, \ldots, L$. Equivalently, for an original token identity u, $\beta_u = \frac{\text{rank}_\pi(u)-1}{L-1}$. During training, $\boldsymbol{\beta}_\pi$ is broadcast across the batch to match $\mathbf{q} \in \mathbb{R}^{B\times L}$. The end-to-end training loop that integrates NS-PFS with OMT and the proposed loss function in Eq. 21

Algorithm 1 Neural Synchrony-guided Paired Feature Sequencing (NS-PFS)

Require: Visual feature clusters $\{I_a\}_{a=1}^{k_I}$, tabular feature clusters $\{T_b\}_{b=1}^{k_T}$, metric $M(\cdot)$

for each $C \in \{I_a\}_{a=1}^{k_I} \cup \{T_b\}_{b=1}^{k_T}$ **do**
  $\xi_C = \sum_{f \in C} M(f), \quad \mu_C = \frac{1}{|C|} \sum_{f \in C} \phi(f)$
end for
for each (I_a, T_b) **do**
  $E_{ab} = 1 - \frac{|\xi_{I_a} - \xi_{T_b}|}{\xi_{I_a} + \xi_{T_b} + \epsilon}$
  $S_{ab} = \frac{\langle \mu_{I_a}, \mu_{T_b} \rangle}{\|\mu_{I_a}\| \, \|\mu_{T_b}\|}; \quad \Psi_{ab} = E_{ab} S_{ab}$
end for
Find $\mathcal{P}^\star$ (Hungarian) maximizing $\sum_{(a,b) \in \mathcal{P}} \Psi_{ab}$
Set $R = |\mathcal{P}^\star|$
for each $(a, b) \in \mathcal{P}^\star$ **do**
  $J_{(a,b)} = I_a \cup T_b; \quad$ apply $\gamma(\rho_{ab}, \Psi_{ab}); \quad$ solve $\pi^\star_{J_{(a,b)}}$
end for
$\pi_{\text{NS–PFS}} = [\, \pi^\star_{J_{(1)}} \| \ldots \| \pi^\star_{J_{(R)}} \,]$

Ensure: Global permutation $\pi_{\text{NS–PFS}}$

is given in Algorithm O1 in the Supplementary Material (Sec. O).

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}}(y, \hat{\mathbf{y}}) + \lambda_{\text{FS}} \frac{1}{BL} \sum_{\iota=1}^{B} \sum_{\ell=1}^{L} (q_{\iota\ell} - \beta_{\pi,\ell})^2 \tag{21}$$

NS-PFS aligns multimodal token clusters based on an energy-phase synchrony principle, integrating image and tabular signals in a biologically interpretable manner. NS-PFS coupled with the OMT results in an architecture that combines biological interpretability, explicit multimodal integration for image-tabular data, and scalable linear-complexity attention.

3 Experiments and Results

A: Datasets and Evaluation. We evaluate iSyncTab on six image-tabular datasets: DVM [52] with 176,414 car images, 17 attributes, and 283 classes, imputing missing tabular values via CTGAN [103]; Deep Lesion (DLes) [67, 104] with 32,725 CT slices and metadata, retaining 1,327 valid image-tabular pairs after 35-feature preprocessing and matching; HAM10000 (HAM) [17, 68, 85] with 10,015 dermoscopic images and four metadata attributes; CheXpert (CheX) [54], using a Kaggle subset [3] of 9,999 X-ray/tabular pairs with 17 features for Cardiomegaly classification; Pokémon (Pok) [19] with 1,025 samples and 15 non-text attributes; and Pet Finder (Pet) [51] with 14,652 images and 19 features. We use a stratified 64/16/20 train/val/test split (similar as [75]) and report test accuracy, Avg. Rank/Regret (Table 1), noise robustness (Table 3), and efficiency (Fig. 2). Significance is tested with Friedman [29], Nemenyi [71], and Wilcoxon-Holm [18], with CD diagrams [18] in the supplement (Sec. E).

B: Implementation Details. Image features are extracted with ResNet50 [47],

following STiL [21] and TIP [22]; tabular features are encoded by a transformer encoder and fused through NS-PFS. iSyncTab is tuned with Optuna [1] for 20 trials per dataset, following tabular tuning practice [32–34, 36], while baselines use recommended settings. Experiments run in PyTorch on 8× RTX A6000 GPUs and 2× Xeon Gold 5320 CPUs; values appear in the supplement (Sec. B). Missing tabular values use median numerical imputation and categorical/text encoding, except DVM, where missingness is imputed with CTGAN [103].

C: Computational Complexity Analysis. iSyncTab is dominated by NS-PFS sequencing and OMT encoding. NS-PFS performs visual-tabular cluster matching with the Hungarian algorithm [59] in $\mathcal{O}(K^3)$, where $K = \max(k_I, k_T) \ll m$. For each matched joint cluster $J_{(r)}$, the local CPP scales as $\mathcal{O}(|J_{(r)}|^2)$ over sparsified edges, giving total sequencing cost $\mathcal{O}\left(K^3 + \sum_{r=1}^{R} |J_{(r)}|^2\right)$, where $R = \min(k_I, k_T)$. The OMT backbone uses Linformer attention [92], reducing attention cost from $\mathcal{O}(L^2)$ to $\mathcal{O}(Lr_{\mathrm{LF}})$, i.e., linear in the ordered token length L for fixed projection rank r_{LF}. Thus, iSyncTab avoids full quadratic feature-token interaction while retaining transformer-style fusion. Empirically, it yields ∼30-40% lower runtime and memory than full self-attention multimodal baselines (Fig. 2); further analyses are in the supplement (Sec. D).

D: Comparative Results. We consider three aspects:

D1: Comparing w/ Multimodal & Standard Baselines. Across all six benchmarks, iSyncTab is the top performer in Table 1, achieving the best Avg. Rank (1.08±0.19) and zero Avg. Regret. Gains are strongest on multimodal-friendly datasets such as Pok and Pet, where synchronizing visual features with tabular cues outperforms image-only (ResNet-50, ViT, SimCLR, BYOL), tabular-only (LGBM, CatBoost, TabM, TabSeq, SCARF, SAINT), and strong fusion baselines such as STiL and TIP. ViT narrows the gap on image-dominant DVM, while tabular learners occasionally help when images are less informative, but both show higher regret overall. Results therefore show a consistent trend: multimodal fusion > image-only > tabular-only, with iSyncTab's alignment-aware pairing most reliably converting complementary modalities into accuracy gains. Compared with our prior iStructTab [41], iSyncTab improves on the five directly comparable datasets by $\Delta = +0.31$, $+1.59$, $+1.88$, $+12.45$, and $+0.27$ on DVM, HAM, DLes, CheX, and Pet, respectively.

D2: Comparing in Noisy Scenario. iSyncTab shows strong robustness to label noise (Table 3). It matches or surpasses prior bests across all noise rates, edging CUFIT [107] on average (78.41% vs. 78.3%). Notably, iSyncTab attains top accuracy at 10% and 20% noise and remains competitive at heavier corruption (40-60%), where CUFIT slightly leads at the highest rates. The results suggest iSyncTab's image-tabular synchrony retains discriminative signals under noisy supervision, yielding the best mean accuracy across tested regimes.

D3: Comparing in Computational Efficiency. We compare image-tabular models [15, 21–23, 45, 56, 66, 100] on DVM using an efficiency-centric protocol inspired by EfficientNet-style accuracy-cost analysis [48, 69, 83] and broader efficiency studies over FLOPs, parameters, memory, training cost, and hardware performance [10, 11, 73]. Since image-tabular pipelines often fuse

Table 1: Test accuracy (%) across six datasets; higher is better. Baselines follow the STiL benchmark [21], with additional tabular-only baselines and an image-only ViT baseline. Avg. Rank ranks all 16 methods per dataset using average ties and reports mean ± std across datasets; lower is better. Avg. Regret is the mean gap to the per-dataset best score. †denotes results from [21], and * denotes results from [15]. All datasets provide one-to-one image-tabular pairing via shared image IDs.

			Dataset statistics							
Images			176414	10015	1327	1025	9999	14652		
Tabular Features			17	4	35	15	17	19		
Classes			283	7	3	18	3	5		
Model	Modality		DVM	HAM	DLes	Pok	CheX	Pet	Avg. Rank	Avg. Regret ↓
	I	T								
Tabular Classifiers										
LGBM [57]		✓	26.89	71.80	76.71	36.10	52.25	39.70	8.67 ± 3.14	36.66 ± 21.12
CatBoost [76]		✓	27.30	72.05	77.18	36.59	53.19	39.40	7.67 ± 2.69	35.35 ± 21.20
TabSeq [44]		✓	26.66	70.14	83.46	15.61	51.10	37.33	11.08 ± 4.25	41.46 ± 25.81
SCARF [5]		✓	64.47†	69.40	83.08	19.02	51.35	33.03	10.33 ± 3.73	35.45 ± 21.17
SAINT [82]		✓	83.36†	70.14	69.17	13.17	35.50	27.29	13.00 ± 3.50	39.07 ± 23.21
TabM [32]		✓	27.09	70.84	84.59	18.05	51.90	38.01	9.25 ± 3.91	40.43 ± 25.68
Image Classifiers										
ResNet50 [47]	✓		32.07†	82.13	69.17	31.22	48.35	35.52	10.25 ± 3.16	39.10 ± 21.92
ViT [20]	✓		88.00*	82.33	69.17	35.61	48.70	34.94	8.58 ± 3.83	29.05 ± 18.93
SimCLR [14]	✓		51.44†	82.38	64.66	37.56	56.15	31.76	8.42 ± 4.87	30.54 ± 17.51
BYOL [38]	✓		47.49†	80.78	56.39	24.39	51.25	34.77	10.83 ± 2.97	39.66 ± 18.25
Multimodal Classifiers										
DAFT [100]	✓	✓	74.22†	74.88	77.07	18.05	67.75	45.45	7.42 ± 2.71	29.27 ± 19.86
Interact Fuse [23]	✓	✓	78.58†	84.87	70.68	11.71	50.90	52.44	8.33 ± 4.71	30.65 ± 22.37
MMCL [45]	✓	✓	85.79†	70.09	67.34	38.05	49.00	29.72	10.67 ± 5.19	27.75 ± 16.03
TIP [22]	✓	✓	98.27†	70.39	69.17	37.56	87.50	83.86	6.00 ± 4.07	10.08 ± 8.09
STiL [21]	✓	✓	99.27†	78.48	81.35	27.32	**88.60**	87.68	4.42 ± 2.83	11.73 ± 20.49
iSyncTab (ours)	✓	✓	**99.60**	**86.82**	**85.63**	**42.44**	**88.60**	**88.02**	**1.08 ± 0.19**	**0.00 ± 0.00**

2048-D ResNet50 features with tabular inputs, we report GFLOPs, parameters, peak inference GPU memory, and a composite complexity proxy. For each method, we compute Eff_i by aggregating min-max normalized accuracy-to-cost ratios over GFLOPs, parameters, memory, and FLOPs × Params, using weights $\mathbf{w} = (0.40, 0.25, 0.20, 0.15)$. Accuracy values are taken from original papers when available; TIME [66] reports no DVM accuracy and is marked `NaN`. Fig. 2 summarizes the weighted efficiency ranking, Pareto frontier, accuracy-complexity trend, radar profile, and factorized resource views. Across panels, iSyncTab achieves the best accuracy while using FLOPs close to light baselines such as DAFT and far below MIITA [15], TIP, or STiL; its parameter budget is closer to DAFT/InterFuse than heavier STiL/TIP models, and it attains the lowest peak memory. Thus, iSyncTab's gains are not due to simply scaling model size. Full Eff_i derivations and normalization details are in the supplement (Sec. D).

E: Ablation on Model Components. Table 4 ablates key iSyncTab components on DVM and HAM, with and without hyperparameter search. The full model, ResNet-50 [47]+NS-PFS+OMT, achieves the best overall performance, confirming the synergy between synchrony-guided sequencing and order-aware memory modeling. Removing memory tokens causes major drops, especially on DVM (99.80→92.68 w/o tuning; 99.60→87.60 w/ tuning) and also hurts HAM (78.72→77.90; 86.82→80.12). Disabling NS-PFS reduces accuracy relative to the full model (DVM: 99.80→96.80; HAM: 78.72→78.16), while random column shuffling causes the largest degradation (DVM: 99.80→86.72; HAM:

Table 2: HAM ablation (inference-only). Accuracy (%) under tabular perturbations (↓ better), image blur level σ, and kNN agreement@5 (%). ECE = 0.259.

Perturb frac	Tab Shuffle ↓	Tab Drop ↓	σ	kNN-Agree@5
0.00	48.88	48.88	0.00	82.33
0.10	48.88	48.88	0.50	82.33
0.25	48.98	48.88	1.00	82.33
0.50	49.08	48.78	1.50	82.33
0.75	48.73	48.38	2.00	82.33
1.00	49.08	48.48		82.33

Table 3: Test accuracy (%) under label noise on HAM. Baselines from CUFIT [107]; we add iSyncTab (Ours) as an extra row.

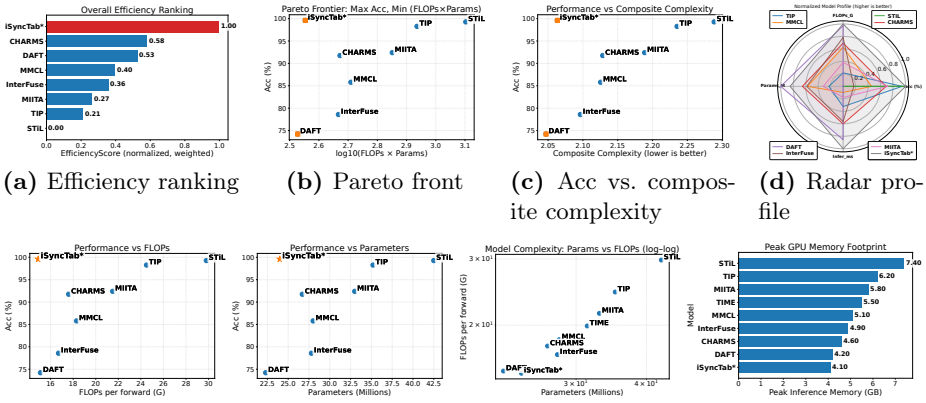
Method/ noise level	**0.1**	**0.2**	**0.4**	**0.6**	**Mean**
Full-training [107]	66.5	62.6	56.1	59.9	61.3
Linear probing [107]	75.6	75.3	71.0	61.9	71.0
Rein [99]	78.6	72.1	54.9	37.8	60.8
Co-teaching [46]	81.5	79.1	74.3	67.3	75.5
JoCor [98]	81.1	79.4	73.9	67.1	75.4
CoDis [101]	81.9	80.1	74.1	66.1	75.5
CUFIT [107]	82.6	81.5	**79.1**	**70.1**	78.3
iSyncTab (Ours)	**82.63**	**82.23**	78.83	69.96	**78.41**

Table 4: Ablation on iSyncTab components on the DVM and HAM datasets under different configurations. Here, w/ tuning means with hyperparameter search and w/o tuning means without hyperparameter search.

Variant	**DVM**		**HAM**	
	w/o tuning	**w/ tuning**	**w/o tuning**	**w/ tuning**
Concat Fuse (no sequencing, no memory)	96.50	96.20	76.58	78.42
iSyncTab (ResNet-50[47] + NS-PFS + OMT)	**99.80**	**99.60**	**78.72**	**86.82**
iSyncTab (ResNeXt-50[102] + NS-PFS + OMT)	99.10	97.90	78.14	84.28
iSyncTab w/o memory tokens	92.68	87.60	77.90	80.12
iSyncTab w/o sequencing-loss term	97.80	98.70	78.06	84.36
iSyncTab w/o feature sequencing	96.80	96.50	78.16	78.60
iSyncTab w/ random column shuffle	86.72	82.31	62.06	62.16

78.72→62.06), showing that gains depend on meaningful feature ordering rather than transformer capacity alone. Removing the auxiliary sequencing loss also degrades performance, particularly on tuned HAM (86.82→84.36), indicating that order supervision stabilizes training. Replacing ResNet-50 with ResNeXt-50 [102] yields lower accuracies, suggesting gains come mainly from NS-PFS+OMT rather than backbone scaling. Overall, the results validate synchrony-aware cross-modal sequencing and memory-based sequence modeling for robust image-tabular fusion. Additional analyses are in the supplement. (Sec. F-N).

F: Inference Level Ablation on Calibration and Robustness. To assess inference-time reliability, we follow the STiL [21] ablation protocol on HAM and evaluate calibration, neighborhood consistency, robustness to visual/tabular perturbations, and efficiency. Table 2 summarizes inference-only behavior: accuracy remains nearly flat (48.5-49.1%) across perturbation fractions and Gaussian blur levels, while fused-embedding k-NN agreement@5 stays constant at 82.33%. The expected calibration error is moderate (ECE= 0.259), suggesting room for post-hoc calibration but no severe miscalibration. Fig. 3 provides further breakdowns: the reliability diagram shows confidence broadly tracking empirical accuracy (Fig. 3a), neighborhood consistency remains strong with $> 80\%$ k-NN agreement across tested k (Fig. 3b), and confidence thresholding yields a smooth coverage-reliability trade-off (Fig. 3c). Robustness probes show stable accuracy under mild Gaussian blur with gradual degradation at stronger blur levels (Fig. 3d), and only small fluctuations under tabular shuffling/dropping (Fig. 3e), indicating that the model is not overly dependent on a brittle subset of visual or tabular cues. Finally, the efficiency sweep shows that shortening the kept se-



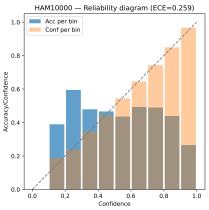
(a) Efficiency ranking **(b)** Pareto front **(c)** Acc vs. composite complexity **(d)** Radar profile

(e) Accuracy vs. FLOPs **(f)** Accuracy vs. parameters **(g)** Params vs. FLOPs (log–log) **(h)** Peak GPU memory footprint

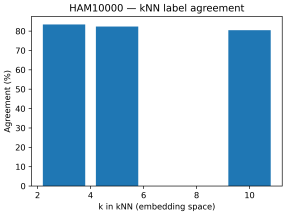
Fig. 2: Comparison of multimodal image-tabular methods on the DVM dataset. Top row: (a) weighted efficiency ranking combining performance and resource usage; (b) Pareto frontier of accuracy vs. complexity (FLOPs × Params); (c) trade-off between accuracy and composite complexity; (d) normalized radar chart summarizing accuracy, computation, and inference latency. Bottom row: (e) test accuracy vs. FLOPs per forward pass; (f) accuracy vs. number of trainable parameters; (g) joint scaling of parameters and FLOPs in log-log space; (h) peak inference-time GPU memory consumption. Accuracies are taken from the respective papers for the DVM dataset; for the TIME model, DVM accuracy is not reported (`NaN`).

quence substantially reduces peak memory with minor latency changes (Fig. 3f), supporting resource-aware deployment. Overall, these ablations indicate that iSyncTab produces reasonably calibrated predictions, preserves coherent fused-embedding neighborhoods, and remains robust to visual/tabular perturbations without retraining, strengthening its suitability for practical clinical inference.

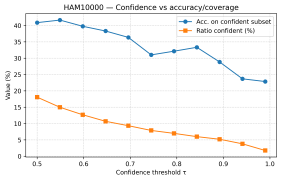
G: Cross-Modal Generality and Sensitivity Analysis. To examine whether the proposed integration mechanism generalizes beyond image-tabular inputs, we extend NS-PFS to audio-video modalities on RAVDESS [65, 72] (also used in [90, 91]). As shown in Table 5, the Linformer-ResNet50 variant reaches the strongest RAVDESS performance, with 97.56% best test accuracy and $95.25 \pm 1.34\%$ five-fold mean accuracy, while Performer [16] provides a slightly lighter memory profile. We also observe that replacing ResNet50 with ResNeXt50 does not consistently improve multimodal fusion, supporting prior findings that stronger pretrained visual backbones do not always transfer better when downstream performance depends on cross-modal feature geometry and alignment [58, 78]. This also motivates our use of ResNet50, which matches prior image-tabular multimodal works such as MMCL [45], TIP [22], STiL [21], and CHARMS [56]. Finally, Fig. 4 evaluates NS-PFS sensitivity under fixed architecture and optimization settings by varying only $M(\cdot)$ and $k_I = k_T$. Performance remains stable across reasonable metric-cluster choices in both HAM image-tabular and RAVDESS audio-video settings, with full sensitivity curves reported in the supplement (Sec. N).



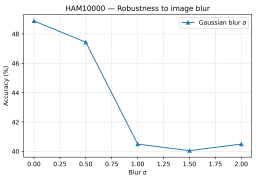
(a) Reliability (ECE=0.259)



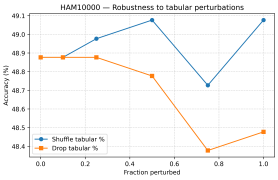
(b) kNN label agreement



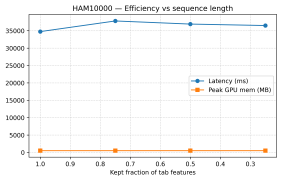
(c) Confidence vs. acc/coverage



(d) Robustness to image blur



(e) Robustness to tabular noise



(f) Efficiency vs. sequence length

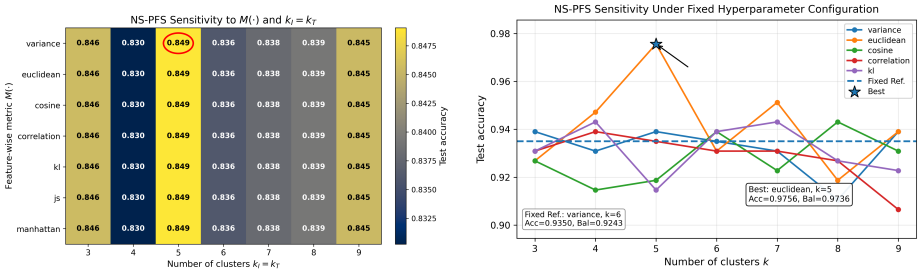
Fig. 3: Ablations of iSyncTab on HAM in the style of STiL [21]: (a) calibration via reliability curve; (b) neighborhood consistency via k-NN label agreement; (c) confidence-accuracy/coverage tradeoff; (d) robustness to image blur; (e) robustness to tabular perturbations; and (f) efficiency as a function of sequence length.

Table 5: RAVDESS audio-video results. Best test accuracy and five-fold mean ± std are reported.

Backbone	Video Enc.	Best Acc. (%)	Mean Acc. (%)	Mean Bal. Acc. (%)	Train Time (min)	Infer. Time (ms)	Peak Mem. Train / Infer. (MB)
Linformer	ResNet50	**97.56**	**95.25 ± 1.34**	**94.72 ± 1.35**	**1.92 ± 0.01**	**229.50 ± 19.10**	392.45 / 302.16
Performer	ResNet50	97.15	94.29 ± 1.59	93.65 ± 1.97	2.23 ± 0.01	272.03 ± 43.48	**367.91** / **268.51**
Linformer	ResNeXt50	94.06	94.31 ± 0.97	93.74 ± 1.32	1.93 ± 0.01	258.17 ± 26.95	394.62 / 305.00
Performer	ResNeXt50	94.72	94.05 ± 0.66	93.51 ± 0.84	2.23 ± 0.02	257.25 ± 12.12	367.97 / 269.67

4 Related Work

A: Multimodal Learning with Tabular and Image Data: Recent work has studied joint representation learning from heterogeneous modalities such as images and tabular data. MMCL [45] uses contrastive pretraining to align image and tabular embeddings and proposes LaaF (Label-as-a-Feature), which appends learnable label tokens during training to strengthen modality interaction. TIP [22] advances this direction with a unified transformer backbone combining masked tabular reconstruction, cross-modal contrastive alignment, and image-tabular matching. DAFT [100] explores dense attentive fusion for scientific pipelines with structured inputs and simulated outputs, while simpler baselines (Concat/Max/Interact Fuse) aggregate modality embeddings via concatenation, pooling, or pairwise interactions [22]. XTab [109] benchmarks fusion strategies using pretrained ResNet features with linear/MLP heads, and subsequent methods introduce specialized cross-modal mechanisms, including hierarchical attention (STiL [21], CHARMS [56]), incremental attribute fusion (MIITA [15]), and temporal embeddings (TIME [66]). Despite architectural dif-



(a) HAM image-tabular **(b)** RAVDESS audio-video

Fig. 4: NS-PFS sensitivity under fixed architecture and optimization settings, varying only $M(\cdot)$ and $k_I = k_T$. Results remain stable across HAM image-tabular and RAVDESS audio-video settings.

ferences, these approaches typically treat fused features as unordered vectors; in contrast, we optimize feature sequencing to reduce dispersion and improve cross-modal alignment, injecting structural priors into fusion. Closest to ours, iStructTab [41] uses graph-enhanced descriptor sequencing to reduce image-tabular feature dispersion, while iSyncTab extends this to synchrony-guided cross-modal CPP with visual-tabular cluster pairing and OMT fusion.

B: Deep Learning for Tabular Data: Deep tabular models have progressed from MLP baselines to attention and embedding-centric architectures. TabNet [2] performs attentive feature selection, TabTransformer [53] learns contextual categorical embeddings, and SAINT [82] combines row/column attention with contrastive pretraining. FT-Transformer [35] uses a streamlined Transformer backbone, while VIME [106] applies self-supervised denoising. TANGOS [55] regularizes attention by predictive utility, and SCARF [5] uses masked contrastive learning. TabPFN [49] and TabPFN V2 [50] use meta-learned probabilistic function networks for rapid adaptation on small tabular tasks. Despite these advances, most methods assume a fixed column order and do not optimize feature permutations; TabSeq [44] and DynaTab [42] explicitly model sequencing and show its effect on accuracy. Recent works further highlight complementary directions: GOTabPFN [40] shows that feature ordering can improve compact tokenization for tabular foundation models, while ZAYAN [43] advances tabular learning in remote sensing through disentangled contrastive modeling.

C: Column Permutation and Combinatorial Optimization: CPPs are combinatorial optimization problems that seek an ordering of items minimizing a dispersion-based objective. Lima et al. [64] instantiate this via a Δ-evaluation function that penalizes pairwise variance or dissimilarity. CPPs are closely related to the Traveling Salesman Problem (TSP), a canonical NP-hard benchmark for permutation learning. Pointer Networks [89] generate permutations via attention, and graph-aware variants such as GPN [105] and PGN [87] further exploit relational structure or dynamic routing. These advances motivate our formulation of multimodal feature sequencing as a learnable CPP.

D: Feature Sequencing and Structure-Aware Representation Learning: Feature sequencing can be pivotal for structured representations whenever permutation-sensitive operators are used [9, 26, 77]. TabSeq [44] learns tabular

feature sequences via clustering and feeds them to a Transformer autoencoder, improving downstream performance. Mambular [84] reinforces the point by processing columns sequentially with Mamba layers, showing that feature position affects learning even without attention. Related structure-aware designs encode topology directly, including structure-aware Transformers [13] and graph-based attention models [24]. These approaches are inherently permutation-sensitive because their attention or sequential operations depend on feature order, unlike set/MLP-style treatments that assume order invariance [108].

Positioning. Unlike existing image-tabular classification models such as STiL [21], TIP [22], DAFT [100], MIITA [15], CHARMS [56], MMCL [45], and simple fusion baselines (e.g., Interact Fuse) [22] which mainly improve encoder pretraining, cross-modal attention/interaction, or fusion operators while treating the fused embedding as an unordered vector, iSyncTab explicitly optimizes a feature permutation that reduces column-wise dispersion and enforces a structured inductive bias before sequence-sensitive modeling. A closely related permutation-based approach is COPER [25], but it targets multi-view clustering (reported via clustering accuracy) rather than supervised image-tabular classification; in contrast, iSyncTab is designed end-to-end for multimodal classification accuracy aligned with the above baselines.

5 Conclusion

We introduce iSyncTab, a unified framework that treats multimodal image-tabular fusion as a CPP. Our neuro-inspired NS-PFS aligns visual and tabular feature clusters via energy and phase-analogue synchrony with Hungarian matching, yielding a coherent cross-modal sequence that the OMT backbone exploits through an auxiliary order-consistency loss. Across six image-tabular benchmarks, iSyncTab consistently attains the best accuracy and ranking (Avg. Rank 1.08 ± 0.19; zero Avg. Regret), remains robust to label noise, and improves computational efficiency through sub-quadratic sequencing and Linformer-based $\mathcal{O}(L)$ attention. Ablations confirm that both synchrony-guided sequencing and memory tokens are critical to the gains, while additional audio-video experiments on RAVDESS suggest that NS-PFS can extend beyond the image-tabular setting. These results position neuro-inspired feature sequencing as a strong inductive bias for structured multimodal learning. Future directions include task-adaptive clustering, broader modality combinations such as text and signals, and structured prediction tasks.

Acknowledgements

This work was supported in part by the US National Science Foundation under Awards #1920920, #2125872, and #2223793. We thank the anonymous ECCV reviewers for their valuable feedback and suggestions. We also gratefully acknowledge Zaigham Abbas Randhawa, Pouya Iranmanesh, and Ghazaleh Mirzaee for their assistance with the GPU clusters.

References

[1] Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M.: Optuna: A Next-Generation Hyperparameter Optimization Framework. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 2623–2631 (2019). https://doi.org/10.1145/3292500.3330701

[2] Arik, S.Ö., Pfister, T.: TabNet: Attentive Interpretable Tabular Learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 6679–6687 (2021). https://doi.org/10.1609/aaai.v35i8.16826

[3] Ashery: Chexpert-v1.0-small. https://www.kaggle.com/datasets/ashery/chexpert, accessed: Jun. 12, 2025

[4] Aydore, S., Pantazis, D., Leahy, R.M.: A Note on the Phase Locking Value and Its Properties. Neuroimage **74**, 231–244 (2013). https://doi.org/10.1016/j.neuroimage.2013.02.008

[5] Bahri, D., Jiang, H., Tay, Y., Metzler, D.: SCARF: Self-Supervised Contrastive Learning Using Random Feature Corruption. In: International Conference on Learning Representations (2022)

[6] Bastos, A.M., Lundqvist, M., Waite, A.S., Kopell, N., Miller, E.K.: Layer and Rhythm Specificity for Predictive Routing. Proceedings of the National Academy of Sciences **117**(49), 31459–31469 (2020). https://doi.org/10.1073/pnas.2014868117

[7] Bastos, A.M., Vezoli, J., Bosman, C.A., Schoffelen, J.M., Oostenveld, R., Dowdall, J.R., De Weerd, P., Kennedy, H., Fries, P.: Visual Areas Exert Feedforward and Feedback Influences through Distinct Frequency Channels. Neuron **85**(2), 390–401 (2015). https://doi.org/10.1016/j.neuron.2014.12.018

[8] Behrisch, M., Bach, B., Henry Riche, N., Schreck, T., Fekete, J.D.: Matrix Reordering Methods for Table and Network Visualization. In: Computer Graphics Forum. vol. 35, pp. 693–716. Wiley Online Library (2016). https://doi.org/10.1111/cgf.12935

[9] Brahmavar, S.B., Li, Y., Oliva, J.: Towards Universal Neural Inference. arXiv preprint arXiv:2508.09100 (2025). https://doi.org/10.48550/arXiv.2508.09100

[10] Cai, H., Gan, C., Wang, T., Zhang, Z., Han, S.: Once-for-All: Train One Network and Specialize it for Efficient Deployment. In: International Conference on Learning Representations (2020)

[11] Cai, H., Zhu, L., Han, S.: ProxylessNAS: Direct Neural Architecture Search on Target Task and Hardware. In: International Conference on Learning Representations (2019)

[12] Canolty, R.T., Edwards, E., Dalal, S.S., Soltani, M., Nagarajan, S.S., Kirsch, H.E., Berger, M.S., Barbaro, N.M., Knight, R.T.: High Gamma Power Is Phase-Locked to Theta Oscillations in Human Neocortex. science **313**(5793), 1626–1628 (2006). https://doi.org/10.1126/science.1128115

[13] Chen, D., O'Bray, L., Borgwardt, K.: Structure-Aware Transformer for Graph Representation Learning. In: International Conference on Machine Learning. pp. 3469–3489. PMLR (2022)

[14] Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A Simple Framework for Contrastive Learning of Visual Representations. In: International Conference on Machine Learning. pp. 1597–1607. PMLR (2020)

[15] Chen, X., Xing, Z., Zhang, Z., Tan, W., Yan, B.: Multimodal Inference with Incremental Tabular Attributes. In: Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI-25). IJCAI Organization (2025). https://doi.org/10.24963/ijcai.2025/543

[16] Choromanski, K.M., Likhosherstov, V., Dohan, D., Song, X., Gane, A., Sarlos, T., Hawkins, P., Davis, J.Q., Mohiuddin, A., Kaiser, L., Belanger, D., Colwell, L., Weller, A.: Rethinking Attention with Performers. In: International Conference on Learning Representations (2021)

[17] Codella, N., Rotemberg, V., Tschandl, P., Celebi, M.E., Dusza, S., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M., et al.: Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC). arXiv preprint arXiv:1902.03368 (2019). https://doi.org/10.48550/arXiv.1902.03368

[18] Demšar, J.: Statistical Comparisons of Classifiers over Multiple Data Sets. Journal of Machine Learning Research **7**, 1–30 (2006), https://www.jmlr.org/papers/volume7/demsar06a/demsar06a.pdf

[19] Dhingra, B.: Complete Pokémon Dataset 9th Gen (Image + Tabular). https://www.kaggle.com/datasets/bhavyadhingra00020/complete-pokemon-dataset-9th-gen-img-tabular, accessed: June. 12, 2025

[20] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In: International Conference on Learning Representations (ICLR) (2021)

[21] Du, S., Luo, X., O'Regan, D.P., Qin, C.: STiL: Semi-Supervised Tabular-Image Learning for Comprehensive Task-Relevant Information Exploration in Multimodal Classification. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 15549–15559 (2025)

[22] Du, S., Zheng, S., Wang, Y., Bai, W., O'Regan, D.P., Qin, C.: TIP: Tabular-Image Pre-Training for Multimodal Classification with Incomplete Data. In: European Conference on Computer Vision. pp. 478–496. Springer (2024). https://doi.org/10.1007/978-3-031-72633-0_27

[23] Duanmu, H., Huang, P.B., Brahmavar, S., Lin, S., Ren, T., Kong, J., Wang, F., Duong, T.Q.: Prediction of Pathological Complete Response to Neoadjuvant Chemotherapy in Breast Cancer Using Deep Learning with Integrative Imaging, Molecular and Demographic Data. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part

II 23. pp. 242–252. Springer (2020). https://doi.org/10.1007/978-3-030-59713-9_24
[24] Dwivedi, V.P., Luu, A.T., Laurent, T., Bengio, Y., Bresson, X.: Graph Neural Networks with Learnable Structural and Positional Representations. In: International Conference on Learning Representations (2022)
[25] Eisenberg, R., Svirsky, J., Lindenbaum, O.: COPER: Correlation-based Permutations for Multi-View Clustering. In: The Thirteenth International Conference on Learning Representations (2025)
[26] Eremeev, D., Bazhenov, G., Platonov, O., Babenko, A., Prokhorenkova, L.: Turning Tabular Foundation Models into Graph Foundation Models. In: NeurIPS 2025 New Perspectives in Graph Machine Learning Workshop (2025)
[27] Esteva, A., Chou, K., Yeung, S., Naik, N., Madani, A., Mottaghi, A., Liu, Y., Topol, E., Dean, J., Socher, R.: Deep Learning-Enabled Medical Computer Vision. NPJ Digital Medicine **4**(1), 5 (2021). https://doi.org/10.1038/s41746-020-00376-2
[28] Fogel, F., Jenatton, R., Bach, F., d'Aspremont, A.: Convex Relaxations for Permutation Problems. Advances in Neural Information Processing Systems **26** (2013)
[29] Friedman, M.: The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance. Journal of the American Statistical Association **32**(200), 675–701 (1937). https://doi.org/10.1080/01621459.1937.10503522
[30] Fries, P.: A Mechanism for Cognitive Dynamics: Neuronal Communication through Neuronal Coherence. Trends in Cognitive Sciences **9**(10), 474–480 (2005). https://doi.org/10.1016/j.tics.2005.08.011
[31] Fries, P.: Rhythms for Cognition: Communication through Coherence. Neuron **88**(1), 220–235 (2015). https://doi.org/10.1016/j.neuron.2015.09.034
[32] Gorishniy, Y., Kotelnikov, A., Babenko, A.: TabM: Advancing Tabular Deep Learning with Parameter-Efficient Ensembling. In: The Thirteenth International Conference on Learning Representations (2025)
[33] Gorishniy, Y., Rubachev, I., Babenko, A.: On Embeddings for Numerical Features in Tabular Deep Learning. Advances in Neural Information Processing Systems **35**, 24991–25004 (2022)
[34] Gorishniy, Y., Rubachev, I., Kartashev, N., Shlenskii, D., Kotelnikov, A., Babenko, A.: TabR: Tabular Deep Learning Meets Nearest Neighbors. In: The Twelfth International Conference on Learning Representations (2024)
[35] Gorishniy, Y., Rubachev, I., Khrulkov, V., Babenko, A.: Revisiting Deep Learning Models for Tabular Data. Advances in Neural Information Processing Systems **34**, 18932–18943 (2021)
[36] Gorishniy, Y., Rubachev, I., Khrulkov, V., Babenko, A.: Revisiting Deep Learning Models for Tabular Data. Advances in Neural Information Processing Systems **34**, 18932–18943 (2021)
[37] Gray, C.M.: The Temporal Correlation Hypothesis of Visual Feature Integration: Still Alive and Well. Neuron **24**(1), 31–47 (1999). https://doi.org/10.1016/S0896-6273(00)80820-X

[38] Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., Piot, B., Kavukçuoğlu, K., Munos, R., Valko, M.: Bootstrap Your Own Latent-A New Approach to Self-Supervised Learning. Advances in Neural Information Processing Systems **33**, 21271–21284 (2020)

[39] Grinsztajn, L., Oyallon, E., Varoquaux, G.: Why Do Tree-Based Models Still Outperform Deep Learning on Typical Tabular Data? Advances in Neural Information Processing Systems **35**, 507–520 (2022)

[40] Habib, A.Z.S.B., Ahamed, M.Y., Gyawali, P.K., Doretto, G., Adjeroh, D.A.: GOTabPFN: From Feature Ordering to Compact Tokenization for Tabular Foundation Models on High-Dimensional Data. International Conference on Machine Learning (2026)

[41] Habib, A.Z.S.B., Ahamed, M.Y., Gyawali, P.K., Doretto, G., Adjeroh, D.A.: iStructTab: Structured Feature Sequencing for Multimodal Learning of Image and Tabular Data. In: Proceedings of the International Conference on Pattern Recognition. Springer (2026), in press

[42] Habib, A.Z.S.B., Doretto, G., Adjeroh, D.A.: DynaTab: Dynamic Feature Ordering as Neural Rewiring for High-Dimensional Tabular Data. In: Proceedings of the First Workshop on NeuroAI Multimodal Intelligence @ AAAI 2026. Proceedings of Machine Learning Research, vol. 308, pp. 27–57. PMLR (2026), https://proceedings.mlr.press/v308/habib26a.html

[43] Habib, A.Z.S.B., Tasnim, T., Islam, M.E., Tabasum, M.: ZAYAN: Disentangled Contrastive Transformer for Tabular Remote Sensing Data. arXiv preprint arXiv:2604.27606 (2026). https://doi.org/10.48550/arXiv.2604.27606

[44] Habib, A.Z.S.B., Wang, K., Hartley, M.A., Doretto, G., A. Adjeroh, D.: TabSeq: A Framework for Deep Learning on Tabular Data via Sequential Ordering. In: International Conference on Pattern Recognition. pp. 418–434. Springer (2024). https://doi.org/10.1007/978-3-031-78128-5_27

[45] Hager, P., Menten, M.J., Rueckert, D.: Best of Both Worlds: Multimodal Contrastive Learning with Tabular and Imaging Data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23924–23935 (2023)

[46] Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I., Sugiyama, M.: Co-teaching: Robust Training of Deep Neural Networks with Extremely Noisy Labels. Advances in Neural Information Processing Systems **31** (2018)

[47] He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)

[48] Hernandez, D., Brown, T.B.: Measuring the Algorithmic Efficiency of Neural Networks. arXiv preprint arXiv:2005.04305 (2020). https://doi.org/10.48550/arXiv.2005.04305

[49] Hollmann, N., Müller, S., Eggensperger, K., Hutter, F.: TabPFN: A Transformer That Solves Small Tabular Classification Problems in a Second. In: The Eleventh International Conference on Learning Representations (2023)

[50] Hollmann, N., Müller, S., Purucker, L., Krishnakumar, A., Körfer, M., Hoo, S.B., Schirrmeister, R.T., Hutter, F.: Accurate Predictions on Small Data with a Tabular Foundation Model. Nature **637**(8045), 319–326 (2025). https://doi.org/10.1038/s41586-024-08328-6

[51] Howard, A., Apers, M., Jedi, M.: PetFinder.my Adoption Prediction. https://kaggle.com/competitions/petfinder-adoption-prediction (2018), kaggle

[52] Huang, J., Chen, B., Luo, L., Yue, S., Ounis, I.: DVM-CAR: A Large-Scale Automotive Dataset for Visual Marketing Research and Applications. In: Proceedings of the IEEE International Conference on Big Data (BigData). pp. 4130–4137. IEEE (2022). https://doi.org/10.1109/BigData55660.2022.10020634

[53] Huang, X., Khetan, A., Cvitkovic, M., Karnin, Z.: TabTransformer: Tabular Data Modeling Using Contextual Embeddings. arXiv preprint arXiv:2012.06678 (2020). https://doi.org/10.48550/arXiv.2012.06678

[54] Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., Seekins, J., Mong, D.A., Halabi, S.S., Sandberg, J.K., Jones, R., Larson, D.B., Langlotz, C.P., Patel, B.N., Lungren, M.P., Ng, A.Y.: CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 590–597 (2019). https://doi.org/10.1609/aaai.v33i01.3301590

[55] Jeffares, A., Liu, T., Crabbé, J., Imrie, F., van der Schaar, M.: TANGOS: Regularizing Tabular Neural Networks through Gradient Orthogonalization and Specialization. In: The Eleventh International Conference on Learning Representations (2023)

[56] Jiang, J.P., Ye, H.J., Wang, L., Yang, Y., Jiang, Y., Zhan, D.C.: Tabular Insights, Visual Impacts: Transferring Expertise from Tables to Images. In: Forty-first International Conference on Machine Learning (2024)

[57] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.Y.: LightGBM: A Highly Efficient Gradient Boosting Decision Tree. Advances in Neural Information Processing Systems **30** (2017)

[58] Kornblith, S., Shlens, J., Le, Q.V.: Do Better ImageNet Models Transfer Better? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2661–2671 (2019)

[59] Kuhn, H.W.: The Hungarian Method for the Assignment Problem. Naval Research Logistics Quarterly **2**(1-2), 83–97 (1955). https://doi.org/10.1002/nav.3800020109

[60] Lachaux, J.P., Rodriguez, E., Martinerie, J., Varela, F.J.: Measuring Phase Synchrony in Brain Signals. Human Brain Mapping **8**(4), 194–208 (1999).

[61] Levin, R., Cherepanova, V., Schwarzschild, A., Bansal, A., Bruss, C.B., Goldstein, T., Wilson, A.G., Goldblum, M.: Transfer Learning with Deep Tabular Models. In: The Eleventh International Conference on Learning Representations (2023)
[62] Lewis, A.G., Schoffelen, J.M., Schriefers, H., Bastiaansen, M.: A Predictive Coding Perspective on Beta Oscillations during Sentence-level Language Comprehension. Frontiers in Human Neuroscience **10**, 85 (2016). https://doi.org/10.3389/fnhum.2016.00085
[63] Liiv, I.: Seriation and Matrix Reordering Methods: An Historical Overview. Statistical Analysis and Data Mining: The ASA Data Science Journal **3**(2), 70–91 (2010). https://doi.org/10.1002/sam.10071
[64] Lima, J.R., Santos, V.G.M., Carvalho, M.A.M.: A Δ-Evaluation Function for Column Permutation Problems. arXiv preprint arXiv:2409.04926 (2024). https://doi.org/10.48550/arXiv.2409.04926
[65] Livingstone, S.R., Russo, F.A.: The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A Dynamic, Multimodal Set of Facial and Vocal Expressions in North American English. PloS one **13**(5), e0196391 (2018). https://doi.org/10.1371/journal.pone.0196391
[66] Luo, J., Yuan, Y., Xu, S.: TIME: TabPFN-Integrated Multimodal Engine for Robust Tabular-Image Learning. arXiv preprint arXiv:2506.00813 (2025). https://doi.org/10.48550/arXiv.2506.00813
[67] Mader, K.S.: NIH DeepLesion Subset. https://www.kaggle.com/datasets/kmader/nih-deeplesion-subset, accessed: Jan. 22, 2025
[68] Mader, K.S.: Skin Cancer MNIST: HAM10000. https://www.kaggle.com/datasets/kmader/skin-cancer-mnist-ham10000, accessed: Jan. 25, 2025
[69] Menghani, G.: Efficient Deep Learning: A Survey on Making Deep Learning Models Smaller, Faster, and Better. ACM Computing Surveys **55**(12), 1–37 (2023). https://doi.org/10.1145/3578938
[70] Nagrani, A., Yang, S., Arnab, A., Jansen, A., Schmid, C., Sun, C.: Attention Bottlenecks for Multimodal Fusion. Advances in Neural Information Processing Systems **34**, 14200–14213 (2021)
[71] Nemenyi, P.B.: Distribution-Free Multiple Comparisons. Princeton University (1963)
[72] Orvile: Audio-Visual Database (RAVDESS). https://www.kaggle.com/datasets/orvile/ravdess-dataset, online; accessed 2026-05-06
[73] Pan, Z., Cai, J., Zhuang, B.: Fast Vision Transformers with HiLo Attention. Advances in Neural Information Processing Systems **35**, 14541–14554 (2022)
[74] van Pelt, S., Heil, L., Kwisthout, J., Ondobaka, S., van Rooij, I., Bekkering, H.: Beta-and Gamma-band Activity Reflect Predictive Coding in the Processing of Causal Events. Social Cognitive and Affective Neuroscience **11**(6), 973–980 (2016). https://doi.org/10.1093/scan/nsw017
[75] Popov, P., Mahmood, U., Fu, Z., Yang, C., Calhoun, V., Plis, S.: A Simple but Tough-to-Beat Baseline for fMRI Time-Series Classification. NeuroImage **303**, 120909 (2024). https://doi.org/10.1016/j.neuroimage.2024.120909

[76] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., Gulin, A.: CatBoost: Unbiased Boosting with Categorical Features. Advances in Neural Information Processing Systems **31** (2018)
[77] Qu, J., HolzmÃžller, D., Varoquaux, G., Morvan, M.L.: TabICL: A Tabular Foundation Model for In-Context Learning on Large Data. In: International Conference on Machine Learning. PMLR (2025)
[78] Raghu, M., Zhang, C., Kleinberg, J., Bengio, S.: Transfusion: Understanding Transfer Learning for Medical Imaging. Advances in Neural Information Processing Systems **32** (2019)
[79] Ramachandram, D., Taylor, G.W.: Deep Multimodal Learning: A Survey on Recent Advances and Trends. IEEE Signal Processing Magazine **34**(6), 96–108 (2017). https://doi.org/10.1109/MSP.2017.2738401
[80] Senkowski, D., Schneider, T.R., Foxe, J.J., Engel, A.K.: Crossmodal Binding through Neural Coherence: Implications for Multisensory Processing. Trends in Neurosciences **31**(8), 401–409 (2008). https://doi.org/10.1016/j.tins.2008.05.002
[81] Singer, W.: Neuronal Synchrony: A Versatile Code for the Definition of Relations? Neuron **24**(1), 49–65 (1999). https://doi.org/10.1016/S0896-6273(00)80821-1
[82] Somepalli, G., Goldblum, M., Schwarzschild, A., Bruss, C.B., Goldstein, T.: SAINT: Improved Neural Networks for Tabular Data via Row Attention and Contrastive Pre-Training. In: NeurIPS 2022 First Table Representation Workshop (2022)
[83] Tan, M., Le, Q.: EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In: International Conference on Machine Learning. pp. 6105–6114. PMLR (2019)
[84] Thielmann, A.F., Kumar, M., Weisser, C., Reuter, A., Säfken, B., Samiee, S.: Mambular: A Sequential Model for Tabular Deep Learning. arXiv preprint arXiv:2408.06291 (2024). https://doi.org/10.48550/arXiv.2408.06291
[85] Tschandl, P., Rosendahl, C., Kittler, H.: The HAM10000 Dataset, A Large Collection of Multi-Source Dermatoscopic Images of Common Pigmented Skin Lesions. Scientific Data **5**(1), 1–9 (2018). https://doi.org/10.1038/sdata.2018.161
[86] Varela, F., Lachaux, J.P., Rodriguez, E., Martinerie, J.: The Brainweb: Phase Synchronization and Large-Scale Integration. Nature Reviews Neuroscience **2**(4), 229–239 (2001). https://doi.org/10.1038/35067550
[87] Veličković, P., Buesing, L., Overlan, M., Pascanu, R., Vinyals, O., Blundell, C.: Pointer Graph Networks. Advances in Neural Information Processing Systems **33**, 2232–2244 (2020)
[88] Vinck, M., Womelsdorf, T., Buffalo, E.A., Desimone, R., Fries, P.: Attentional Modulation of Cell-Class-Specific Gamma-band Synchronization in Awake Monkey Area V4. Neuron **80**(4), 1077–1089 (2013). https://doi.org/10.1016/j.neuron.2013.08.019
[89] Vinyals, O., Fortunato, M., Jaitly, N.: Pointer Networks. Advances in Neural Information Processing Systems **28** (2015)

[90] Wang, H., Xin, X., Qin, X., Jin, M., Ma, J., Xu, D., Jia, J.: Emotion-Conditioned Motion Sub-Spaces with Flow Matching for Real-Time Audio-Driven Talking Heads. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 17769–17777 (2026). https://doi.org/10.1609/aaai.v40i21.38834

[91] Wang, Q., Fan, C., Jia, T., Han, Y., Wu, X.: ND-MRM: Neuronal Diversity Inspired Multisensory Recognition Model. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 15589–15597 (2024). https://doi.org/10.1609/aaai.v38i14.29486

[92] Wang, S., Li, B.Z., Khabsa, M., Fang, H., Ma, H.: Linformer: Self-Attention with Linear Complexity. arXiv preprint arXiv:2006.04768 (2020). https://doi.org/10.48550/arXiv.2006.04768

[93] Wang, T., Guan, S.U.: Feature Ordering for Neural Incremental Attribute Learning Based on Fisher's Linear Discriminant. In: 2013 5th International Conference on Intelligent Human-Machine Systems and Cybernetics. vol. 2, pp. 507–510. IEEE (2013). https://doi.org/10.1109/IHMSC.2013.268

[94] Wang, T., Guan, S.U., Ma, J., Liu, F.: Linear Feature Sensibility for Output Partitioning in Ordered Neural Incremental Attribute Learning. In: International Conference on Intelligent Science and Big Data Engineering. pp. 373–383. Springer (2015). https://doi.org/10.1007/978-3-319-23862-3_37

[95] Wang, T., Guan, S.U., Man, K.L., Park, J.H., Hsu, H.H.: Output Effect Evaluation Based on Input Features in Neural Incremental Attribute Learning for Better Classification Performance. Symmetry **7**(1), 53–66 (2015). https://doi.org/10.3390/sym7010053

[96] Wang, T., Guan, S.U., Man, K.L., Ting, T.: EEG Eye State Identification Using Incremental Attribute Learning with Time-Series Classification. Mathematical Problems in Engineering **2014**(1), 365101 (2014). https://doi.org/10.1155/2014/365101

[97] Wang, T., Zhu, X., Guan, S.U., Man, K.L., Ting, T.: Regression Based on Neural Incremental Attribute Learning with Correlation-based Feature Ordering. In: 2015 IEEE 7th International Conference on Cybernetics and Intelligent Systems (CIS) and IEEE Conference on Robotics, Automation and Mechatronics (RAM). pp. 109–113. IEEE (2015). https://doi.org/10.1109/ICCIS.2015.7274557

[98] Wei, H., Feng, L., Chen, X., An, B.: Combating Noisy Labels by Agreement: A Joint Training Method with Co-Regularization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13726–13735 (2020)

[99] Wei, Z., Chen, L., Jin, Y., Ma, X., Liu, T., Ling, P., Wang, B., Chen, H., Zheng, J.: Stronger Fewer & Superior: Harnessing Vision Foundation Models for Domain Generalized Semantic Segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 28619–28630 (2024)

[100] Wolf, T.N., Pölsterl, S., Wachinger, C., the Alzheimer's Disease Neuroimaging Initiative, the Australian Imaging Biomarkers and Lifestyle Flagship Study of Ageing: DAFT: A Universal Module to Interweave Tabular Data and 3D Images in CNNs. NeuroImage **260**, 119505 (2022). https://doi.org/10.1016/j.neuroimage.2022.119505

[101] Xia, X., Han, B., Zhan, Y., Yu, J., Gong, M., Gong, C., Liu, T.: Combating Noisy Labels with Sample Selection by Mining High-Discrepancy Examples. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1833–1843 (2023)

[102] Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated Residual Transformations for Deep Neural Networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1492–1500 (2017)

[103] Xu, L., Skoularidou, M., Cuesta-Infante, A., Veeramachaneni, K.: Modeling Tabular Data Using Conditional GAN. Advances in Neural Information Processing Systems **32** (2019)

[104] Yan, K., Wang, X., Lu, L., Zhang, L., Harrison, A.P., Bagheri, M., Summers, R.M.: Deep Lesion Graphs in the Wild: Relationship Learning and Organization of Significant Radiology Image Findings in a Diverse Large-Scale Lesion Database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 9261–9270 (2018)

[105] Yang, T., Wang, Y., Yue, Z., Yang, Y., Tong, Y., Bai, J.: Graph Pointer Neural Networks. In: Proceedings of the AAAI conference on Artificial Intelligence. vol. 36, pp. 8832–8839 (2022). https://doi.org/10.1609/aaai.v36i8.20864

[106] Yoon, J., Zhang, Y., Jordon, J., Van der Schaar, M.: VIME: Extending the Success of Self-and Semi-Supervised Learning to Tabular Domain. Advances in Neural Information Processing Systems **33**, 11033–11043 (2020)

[107] Yu, Y., Ko, M., Shin, S., Kim, K., Lee, K.: Curriculum Fine-tuning of Vision Foundation Model for Medical Image Classification Under Label Noise. Advances in Neural Information Processing Systems **37**, 18205–18224 (2024)

[108] Zaheer, M., Kottur, S., Ravanbakhsh, S., Poczos, B., Salakhutdinov, R.R., Smola, A.J.: Deep Sets. Advances in Neural Information Processing Systems **30** (2017)

[109] Zhu, B., Shi, X., Erickson, N., Li, M., Karypis, G., Shoaran, M.: XTab: Cross-Table Pretraining for Tabular Transformers. In: International Conference on Machine Learning. pp. 43181–43204. PMLR (2023)