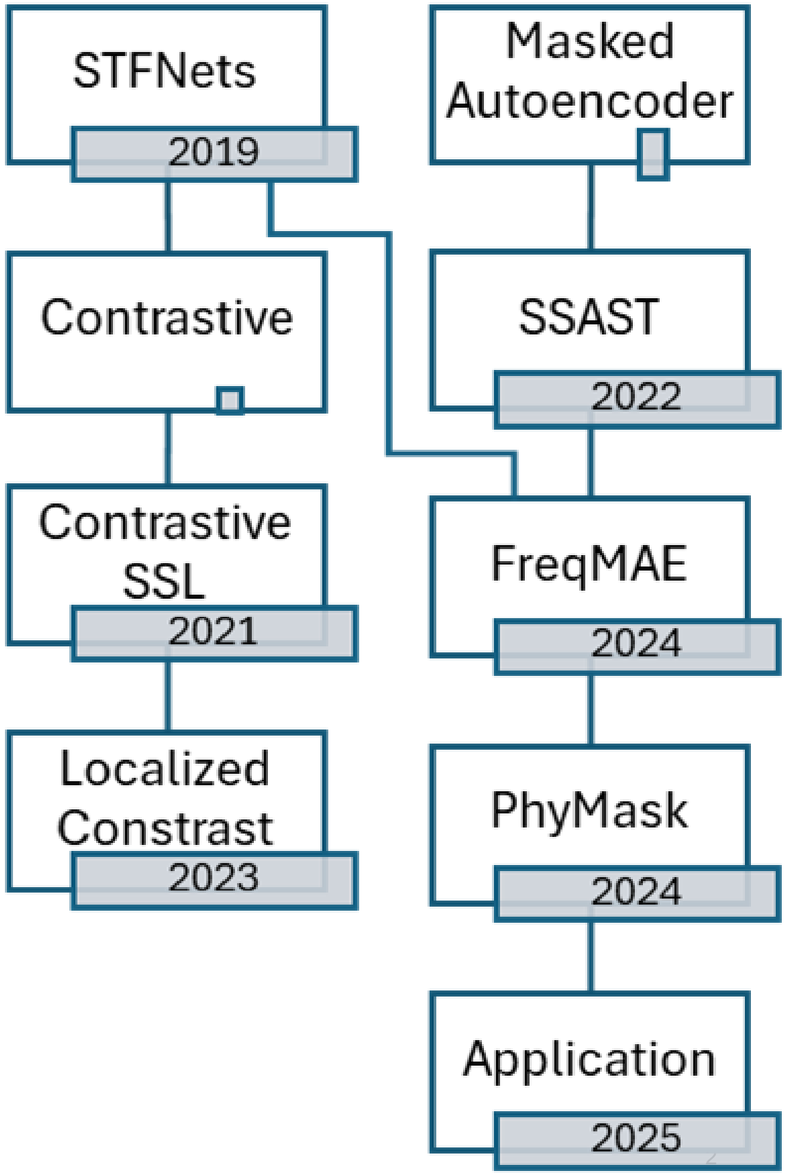
Self-supervised Learning from Frequency Domain Data

Date: 2/24/2026

Presenters: Ebrahim Alfridy, Rafid Murshed, Avery Plote

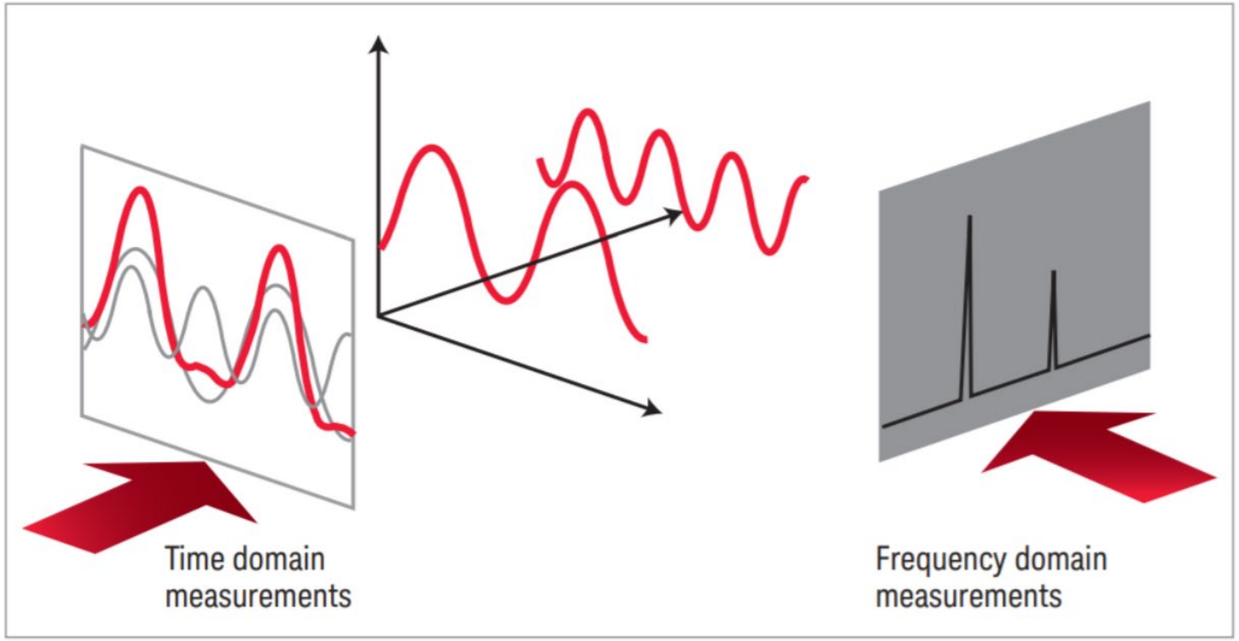
Table of Contents

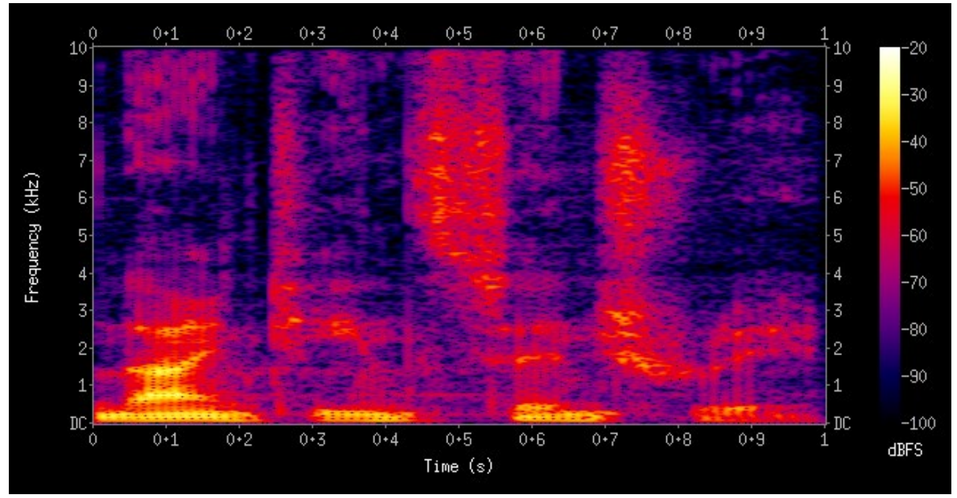


Frequency Domain Recap

When Data Lives in the Frequency Domain

- What is Frequency domain data?
  - Information that tells us the frequency components of a given signal
- What if we still want to keep some time information?
  - Use STFT → Spectrogram
  - Harmonic structure, locality, resolution tradeoffs
- Physical phenomena are often well represented in freq. domain



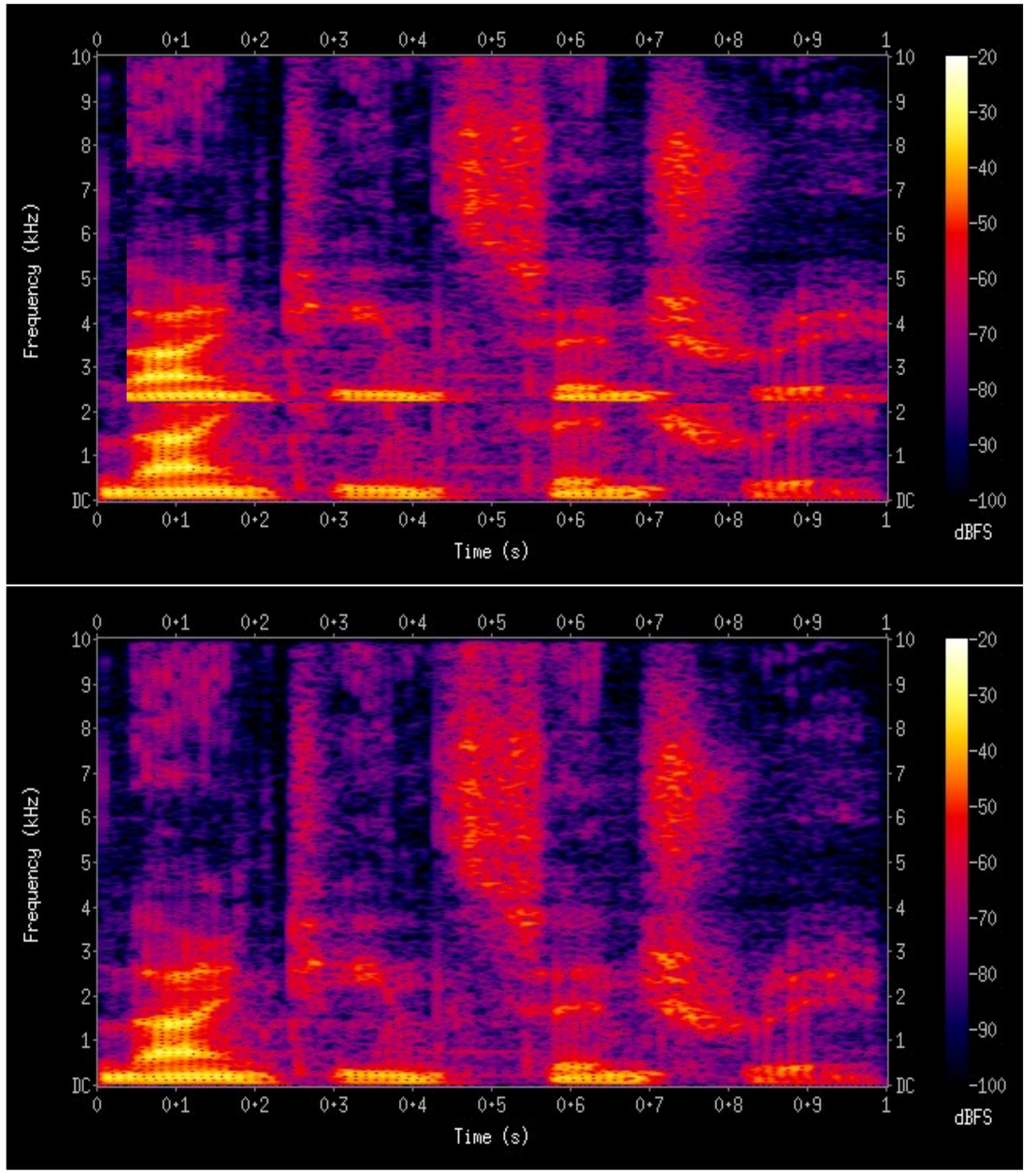


Spectrogram

Process Spectrograms as Images?

- Similar or dissimilar spectrograms?
  - Dissimilar - frequency shift
  - Images look very similar though!

- We need a way to leverage the *unique properties of frequency data*



Design Constraints in Frequency-Domain SSL

Structural Constraint

How can we train models while using the unique structure of frequency-domain data?

STFNets

Representation Constraint

How can we learn both local and global time-frequency patterns?

Contrastive Learning

Scalability Constraint

Can we train models without the need for large, labeled datasets?

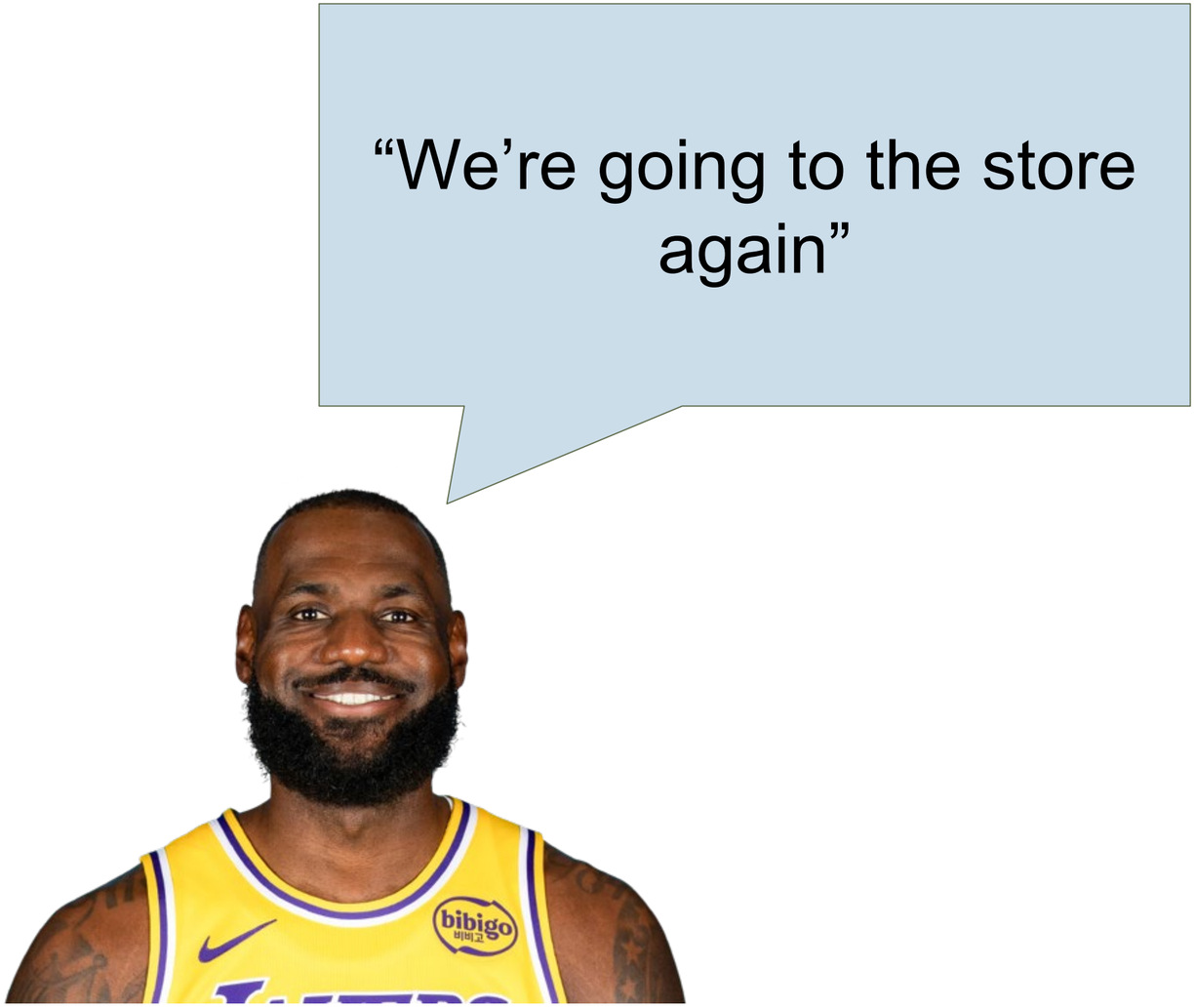
FreqMAE & PhyMask

Teaser / Application

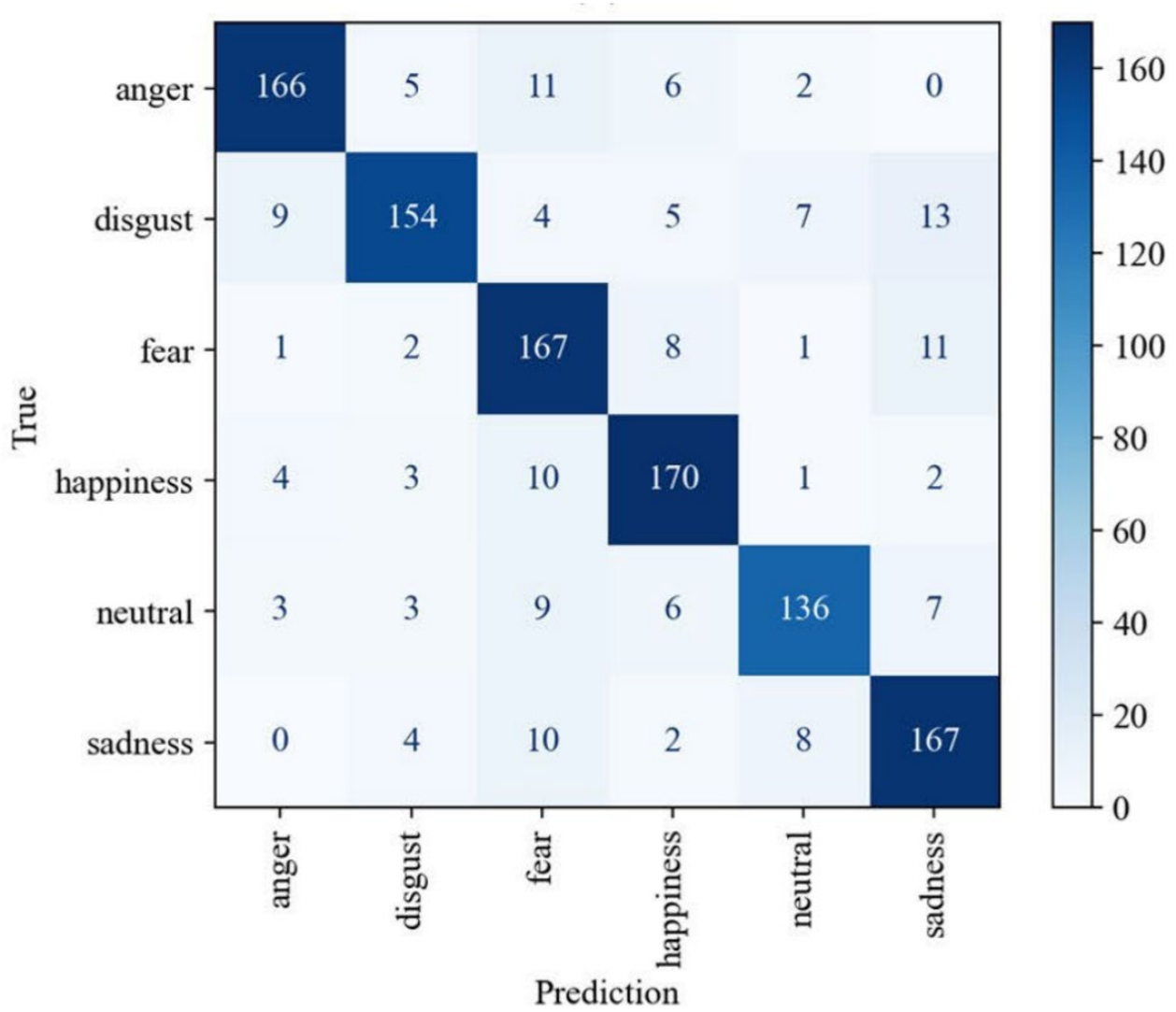
Emotion Recognition from Raw Speech

Self-Supervised models can detect emotion from speed with high accuracy!

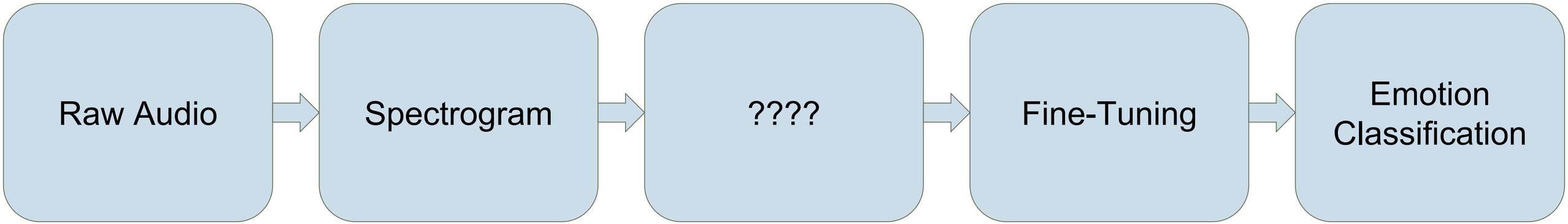
How can this be possible? How did we get here?



Emotion Recognition from Raw Speech



What Enables This?



Structural Modeling (STFNets)

Are ‘Generic’ CNNs Sufficient for IoT Data?

- Generic CNNs are more suited for external perceptual tasks
- Can we treat spectrograms as images and use traditional learning methods?
- **IoT Data** often models real-world phenomena, which is best analyzed in the frequency domain.
  - Can we leverage the unique properties of frequency-domain data?

- **Time-Frequency Hologram:**
  - Beating uncertainty principle
  - Compute multiple STFTs of different length
- **Frequency Filtering Layer:**
  - Learn which frequencies / features best predict network outputs

Designing Networks for Frequency Data

- **Time-Frequency Hologram:**
  - Beating uncertainty principle
  - Compute multiple STFTs of different length
- **Frequency Filtering Layer:**
  - Learn which frequencies / features best predict network outputs

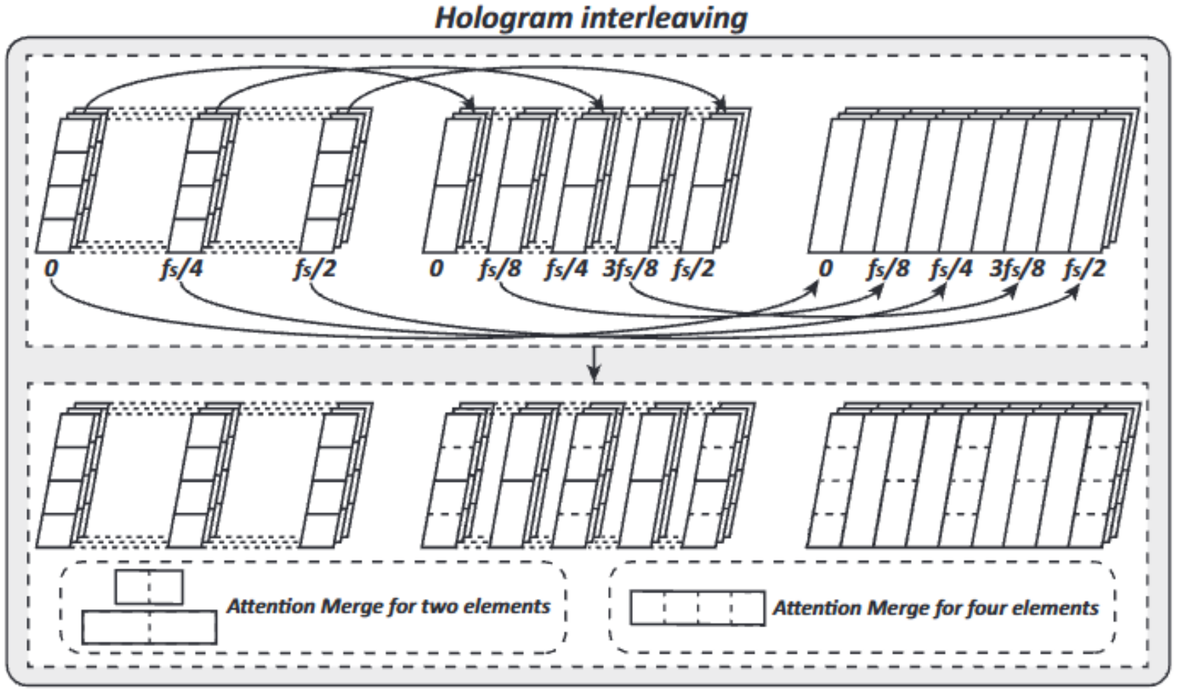
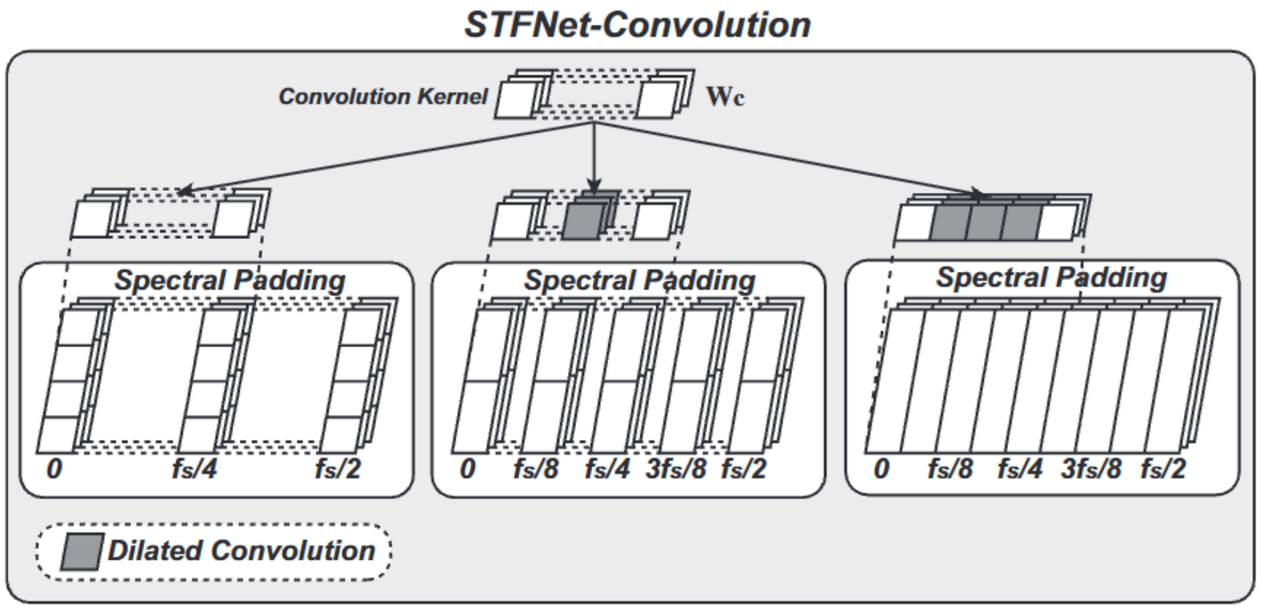


Figure 3: The design of hologram interleaving.

Designing Networks for Frequency Data

- **Frequency-Aware Convolution:**
  - Spectral padding does not distort signal as padding with 0s does
- **Pooling:**
  - Remove unwanted frequency components without zeroing



STFNets: Results

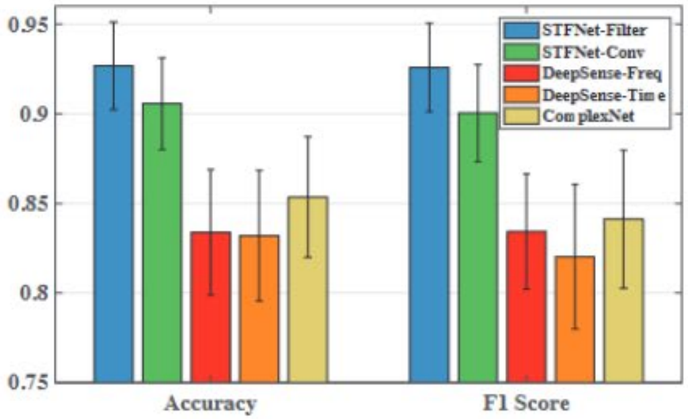


Figure 7: The accuracy and F1 score with 95% confidence interval for motion sensors.

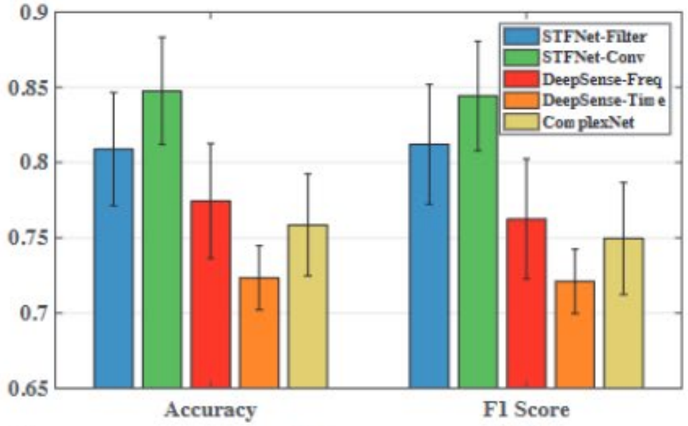


Figure 8: The accuracy and F1 score with 95% confidence interval for WiFi.

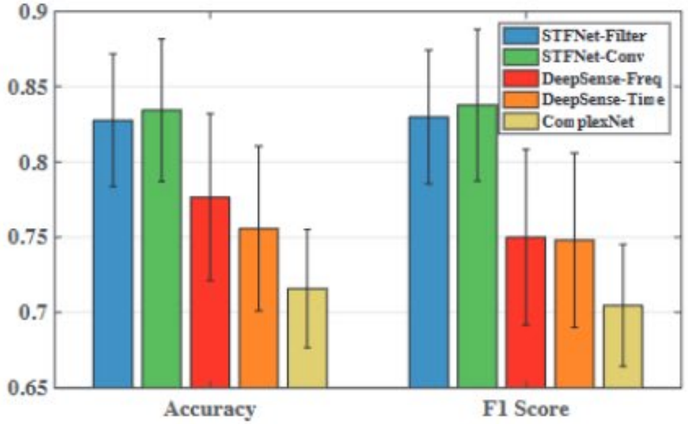


Figure 9: The accuracy and F1 score with 95% confidence interval for Ultrasound.

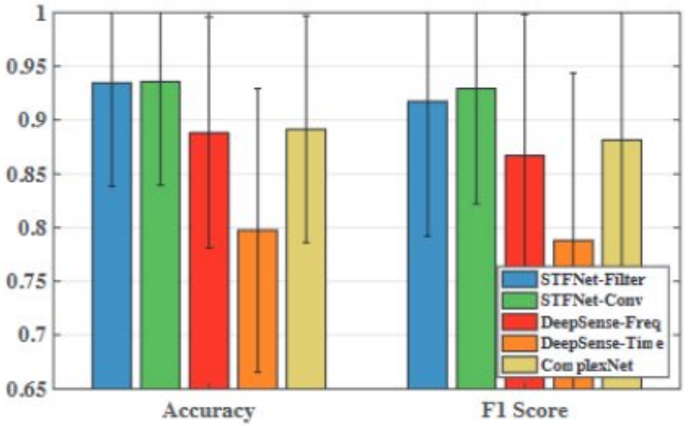
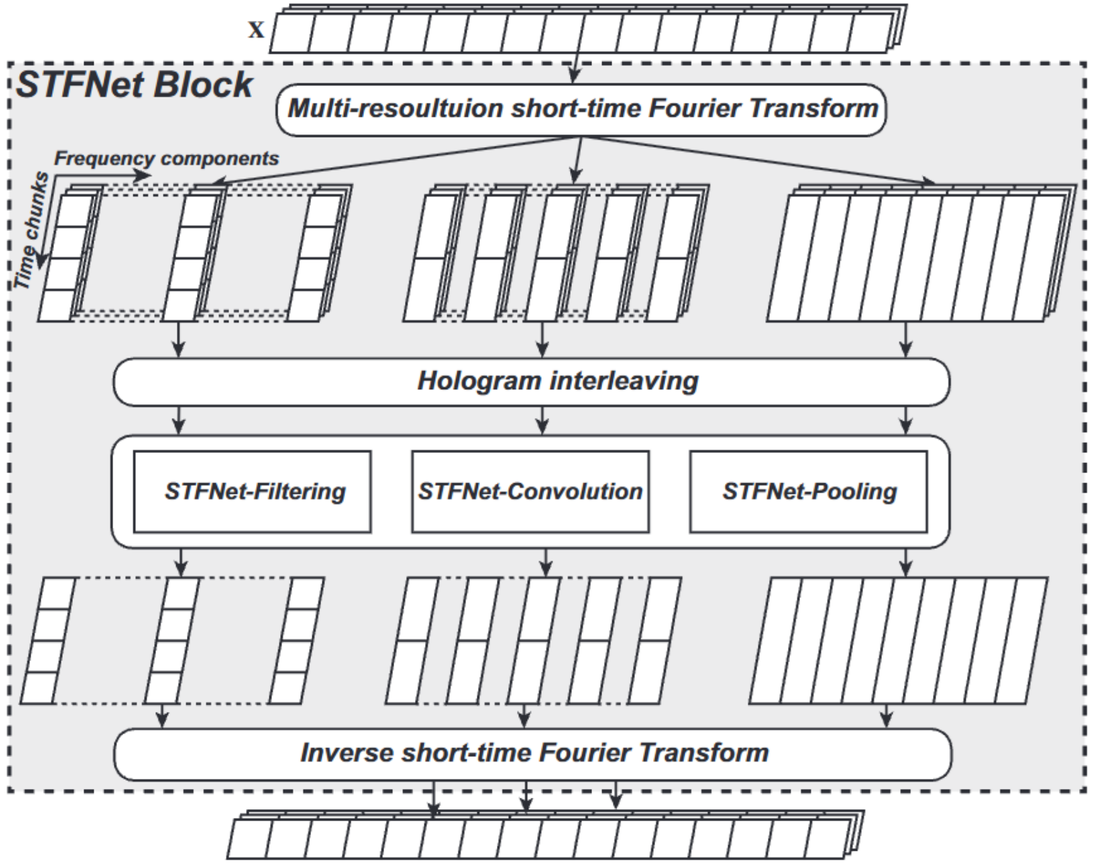


Figure 10: The accuracy and F1 score with 95% confidence interval for Visible light.

Impact of Frequency-Aware Architecture

- **Pros**
  - Improved sensing / model output performance
  - Demonstrated benefit of signal-aware design
  - Shifted architecture philosophy
- **Limitations**
  - Still using supervised learning approach
  - Still dependent on labeled data

Can we overcome these limitations?



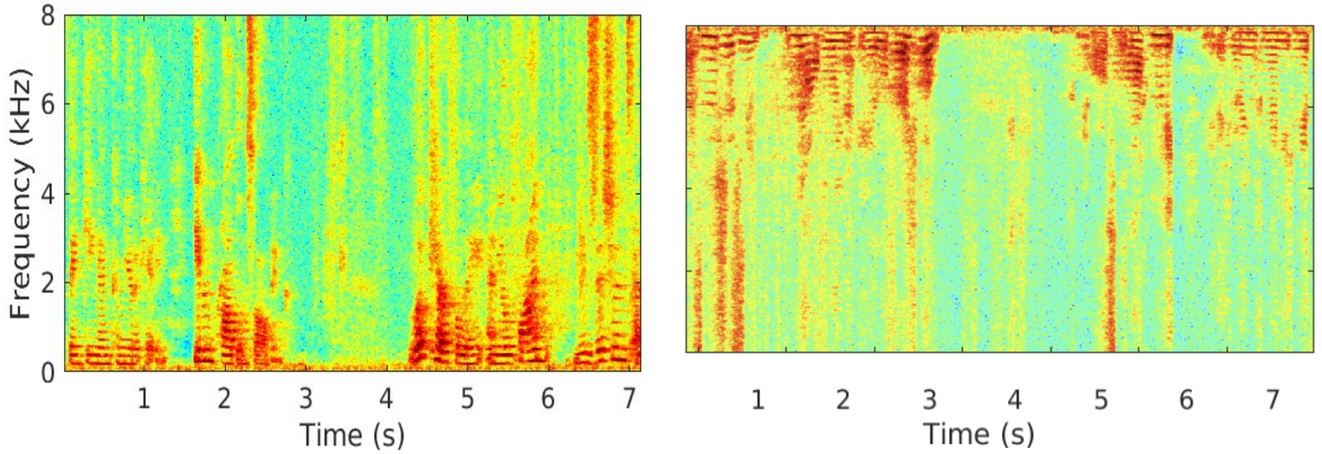
Learning Without Labels (Contrastive)

Contrastive SSRL for Sensing Signals from the T-F Perspective

The Augmentation Dilemma

Similar?

YES

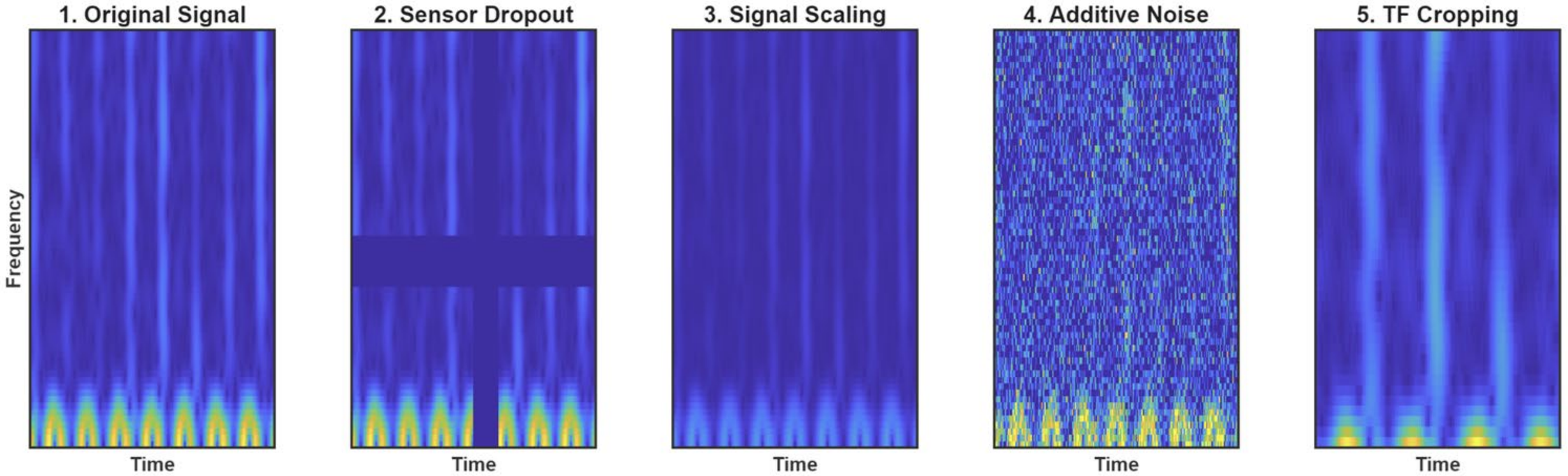


Similar?

NO

The Vision Gap: Image augmentations (rotation, color jitter) destroy signal physics

The Solution: Physics-Aware TF Augmentations



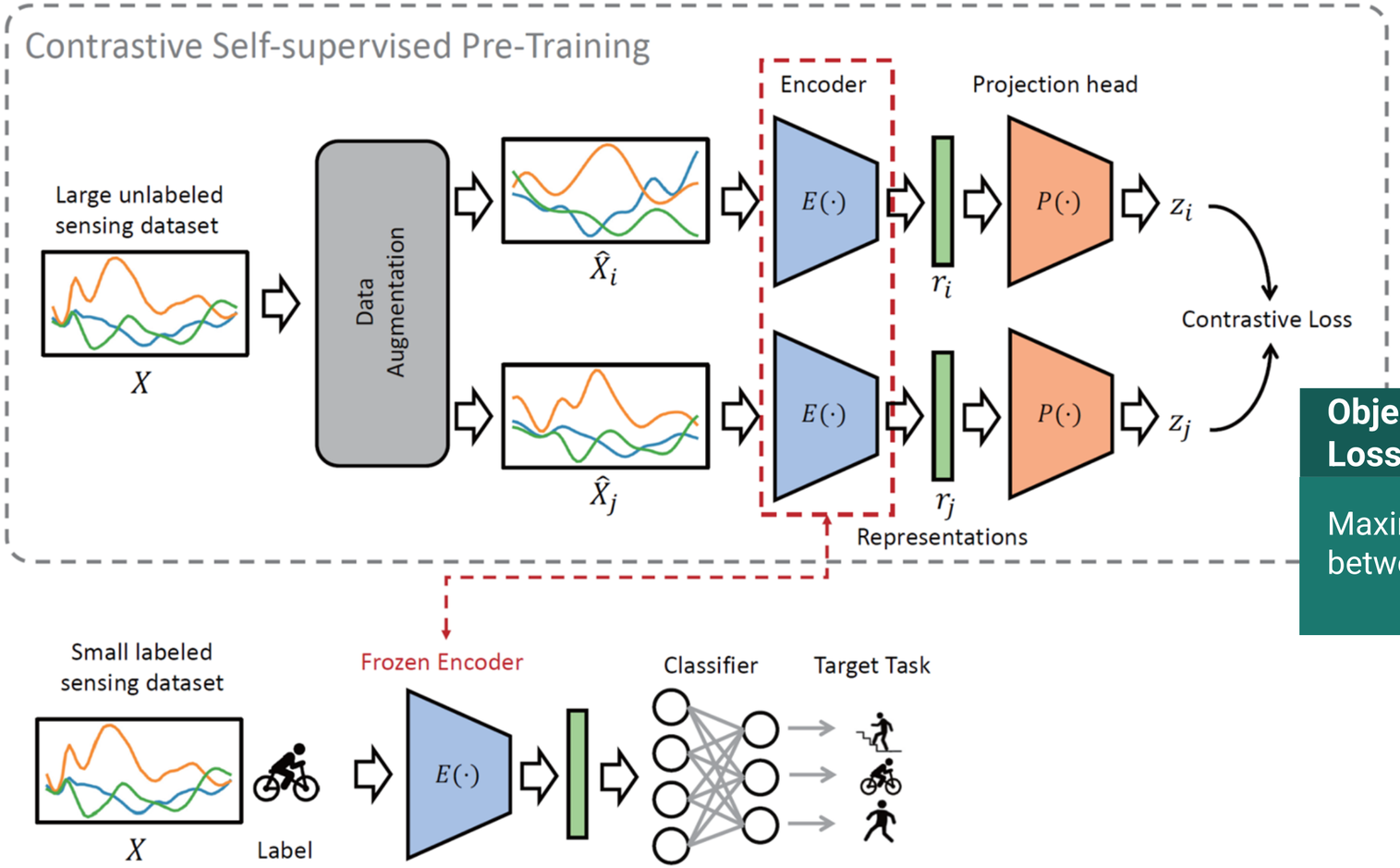
Mimics Hardware Failure

Mimics Varying Sensor Distances

Mimics Ambient Interference

Mimics Localized Events

The Contrastive Architecture



SimCLR-style Pipeline

Two-branch parallel contrastive network

View Generation

Apply two random TF augmentations to one signal

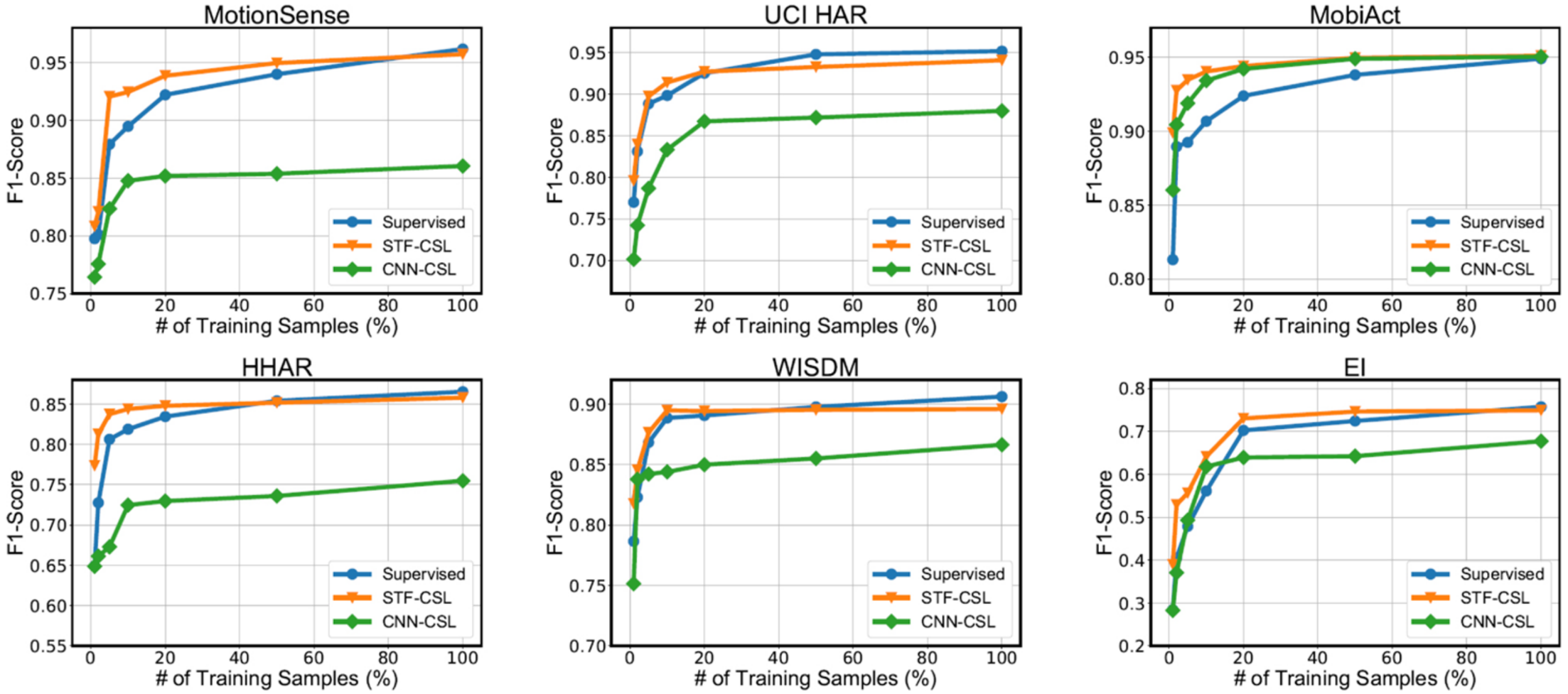
Feature Encoding

Pass views through a CNN backbone

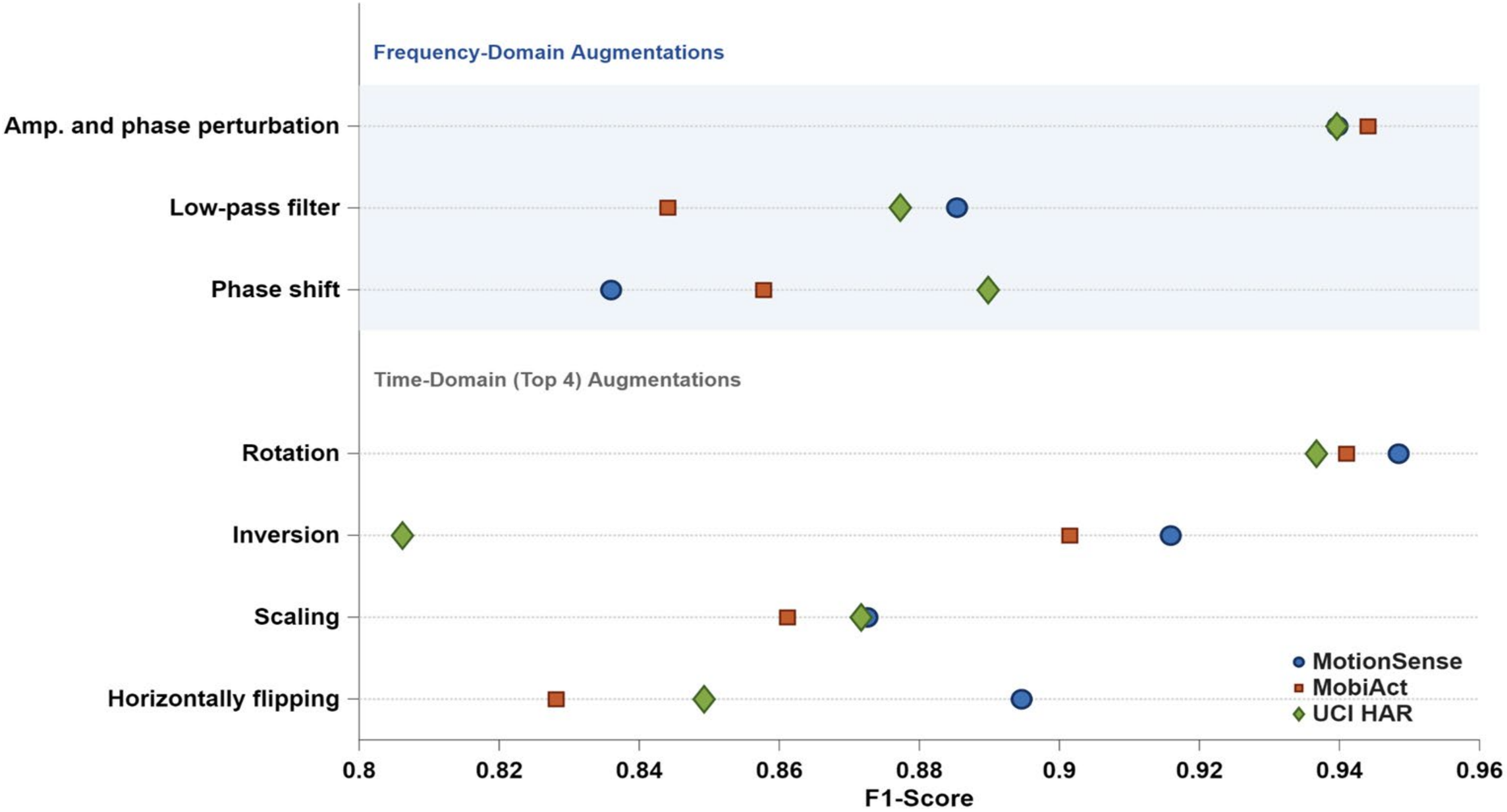
Objective (NTXent Loss)

Maximize agreement between augmented views

Data Efficiency & Cross-Dataset Transfer



Performance under various Augmentation



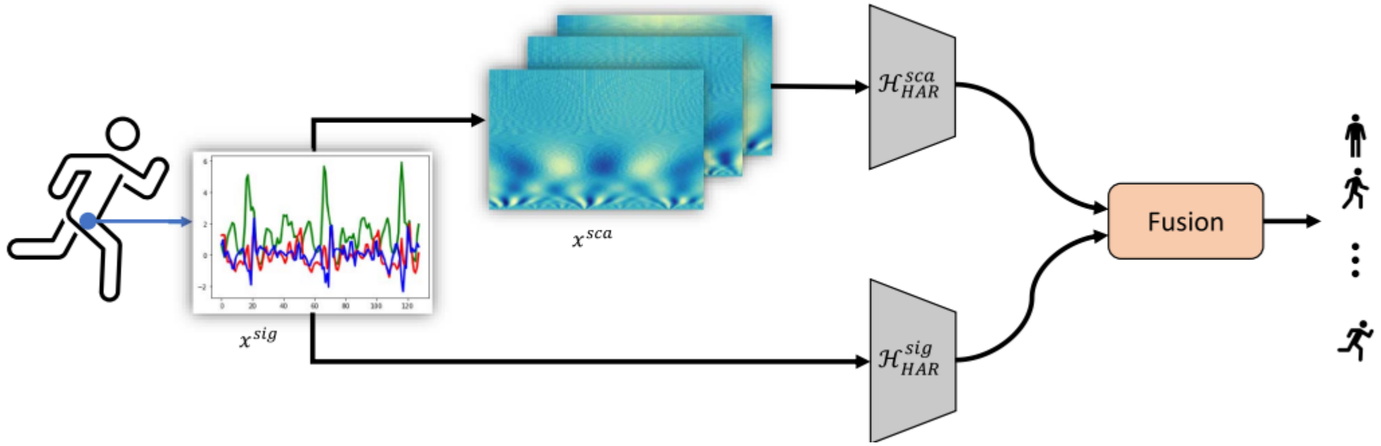
Key Takeaways & Practical Nuances

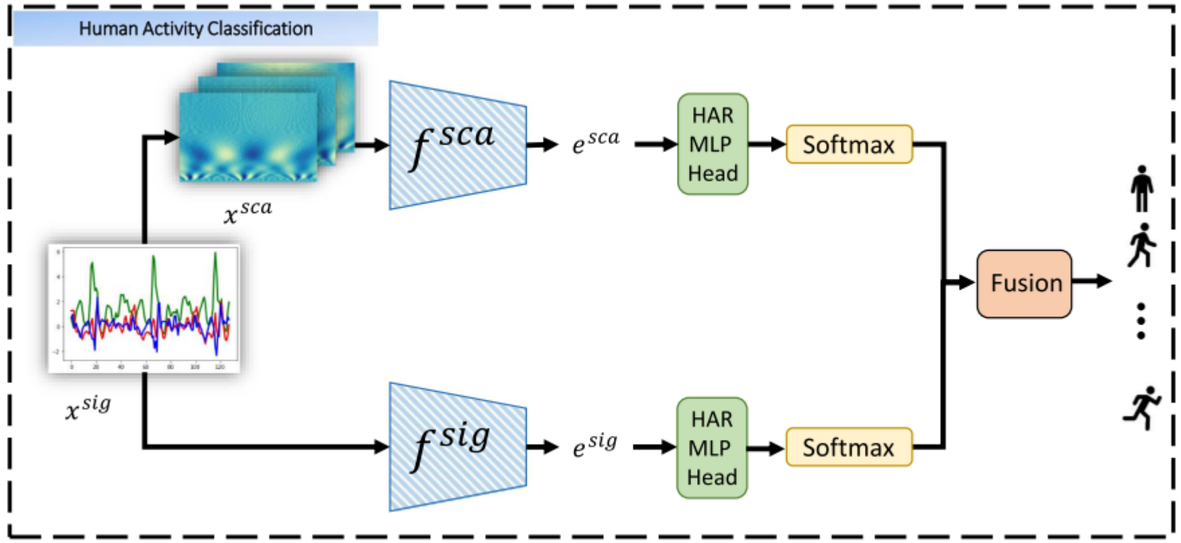
01	The Core Takeaway	• Frequency-domain architectures (STFNet) and augmentations significantly outperform time-domain CNNs for sensing
02	Augmentation Physics	• Amp/Phase perturbations create highly robust representations
03	Physical Invalidity	• Time-domain heuristics (e.g., reversing time, arbitrary rotation) often lack meaningful physical interpretations
04	Transform Rigidity	• The architecture relies strictly on STFT; alternatives like Wavelet transforms remain unexplored

Localized Time-Frequency Contrastive Learning

Global → Localized

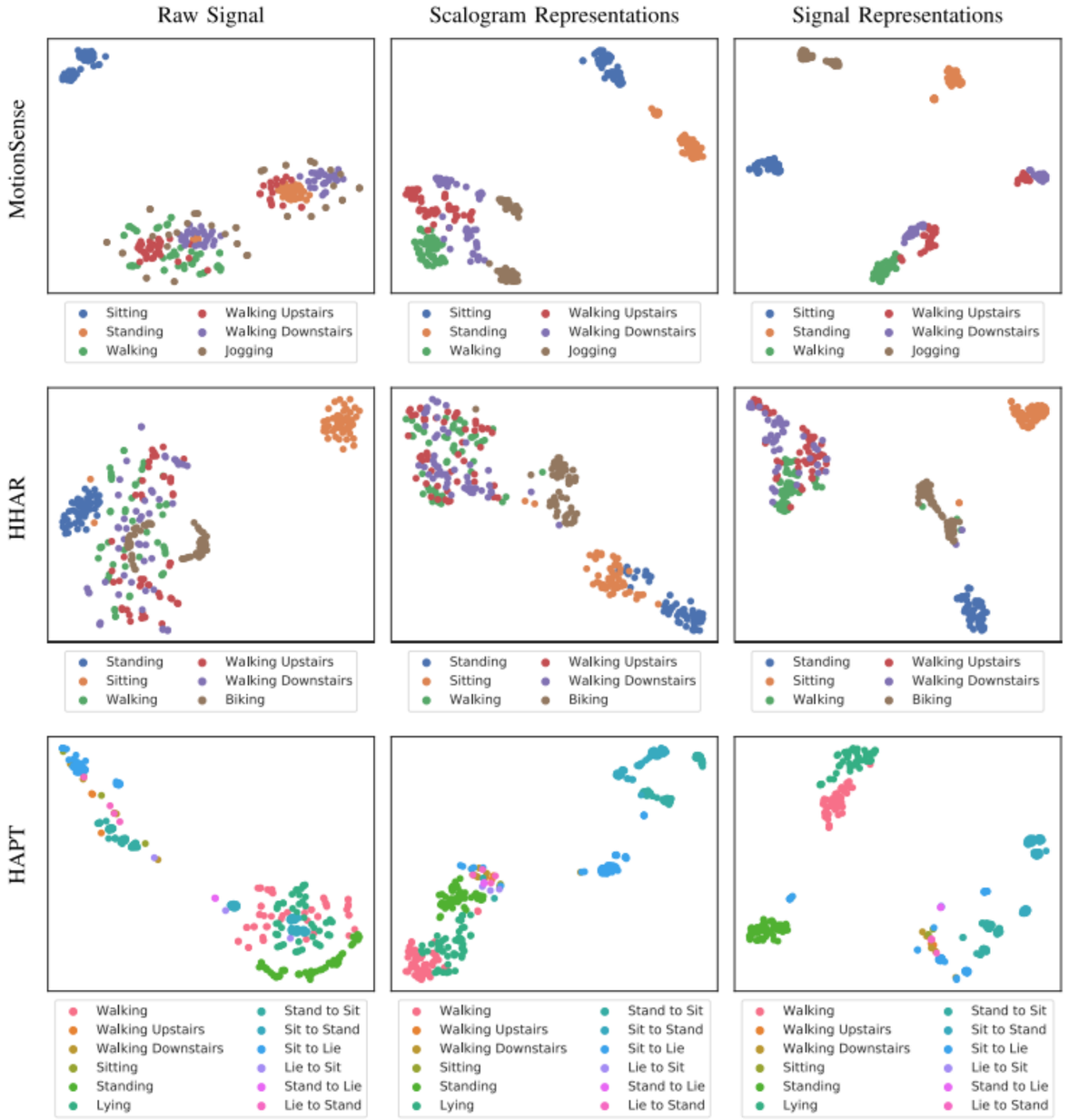
- **Learn from both Time and Frequency Domain**
  - Separate encoders for raw signal and scalogram
  - Use Contrastive Learning for training
- **Fuse the outputs of the separate encoders for final output**





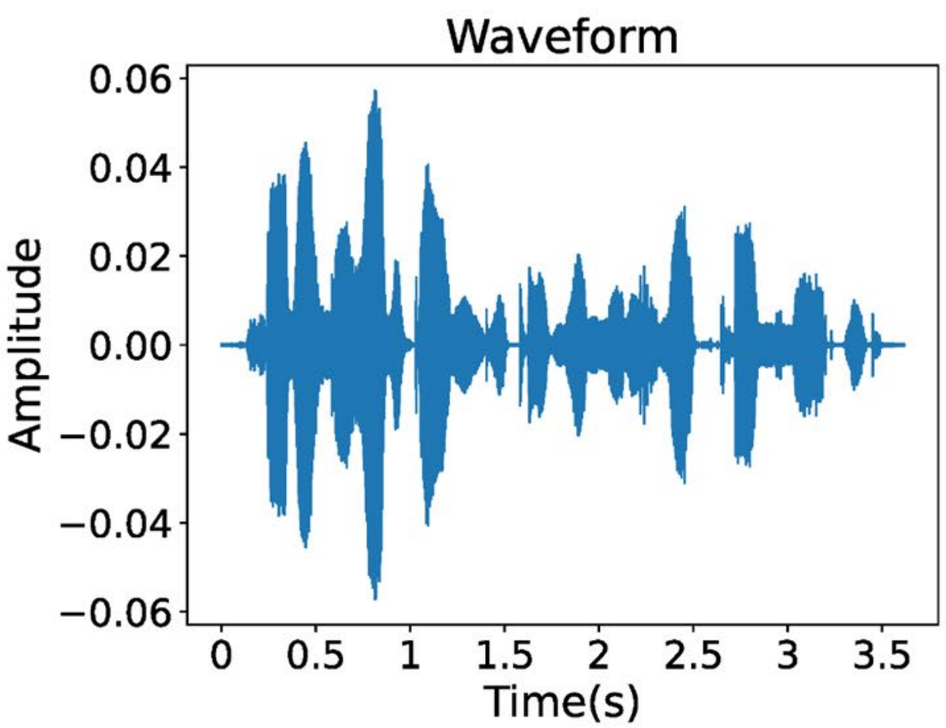
Improvements

- Maintains information about when a frequency occurs
- Uses complementary information from both streams
- Better generalization

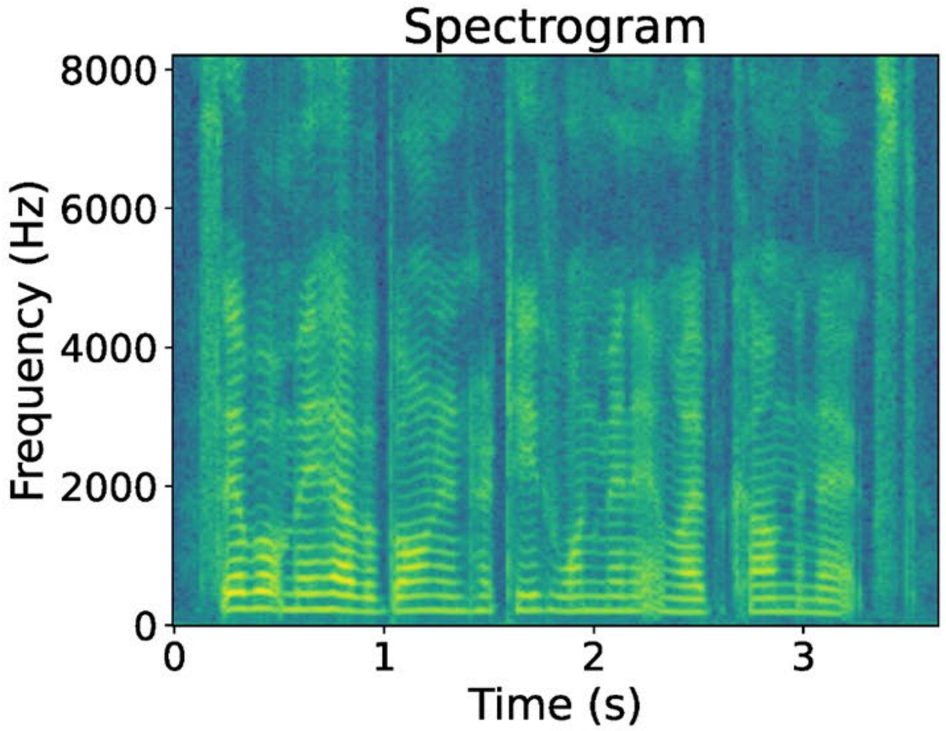


Masked Reconstruction Paradigm (MAE)

1D Raw Audio to 2D Spectrogram



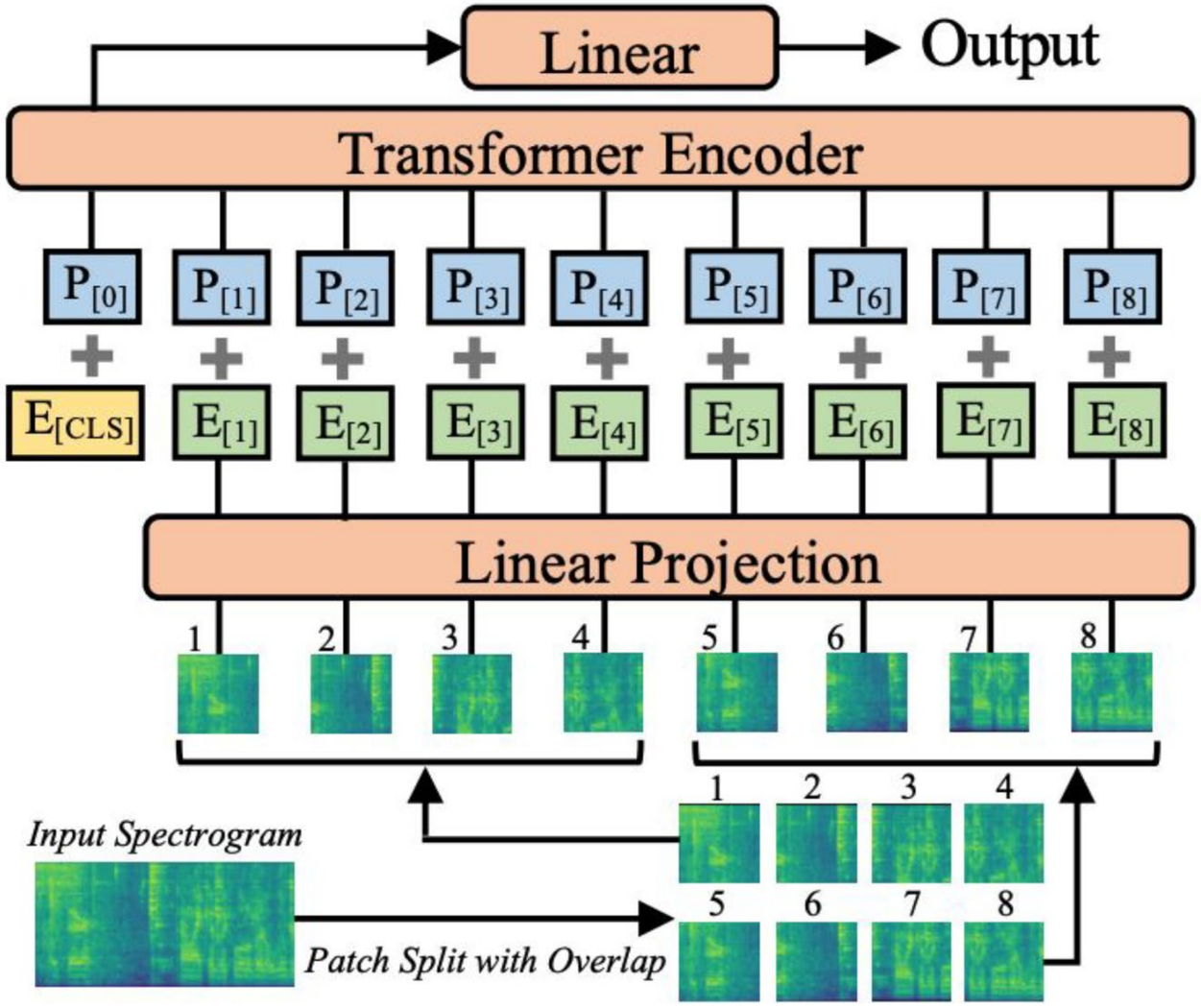
(a)



(b)

The Goal: Learning rich, global time-frequency dynamics without requiring human-annotated labels

AST: Enter the Transformer



- Pure Transformer
- Image-Like Processing
- Global Context
- Variable Length Input
- Fast Convergence

The Bottleneck: Heavily relies on massive supervised datasets

The Data Bottleneck

1 No Inductive Bias

Pure transformers lack the translation invariance of traditional CNNs

2 Data-Hungry

Consequently, the architecture requires massive datasets to learn basic structural patterns

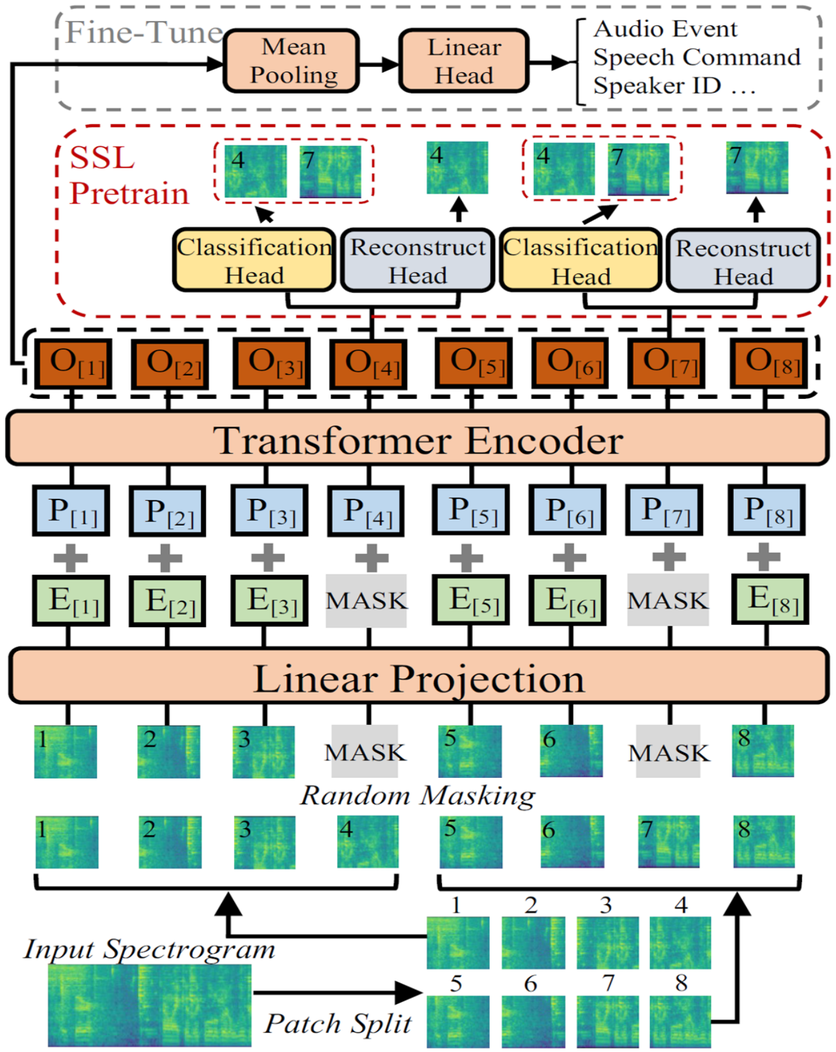
3 The Supervision Trap

Original AST relied heavily on massive, labeled datasets (ImageNet, AudioSet)

4 The Core Challenge

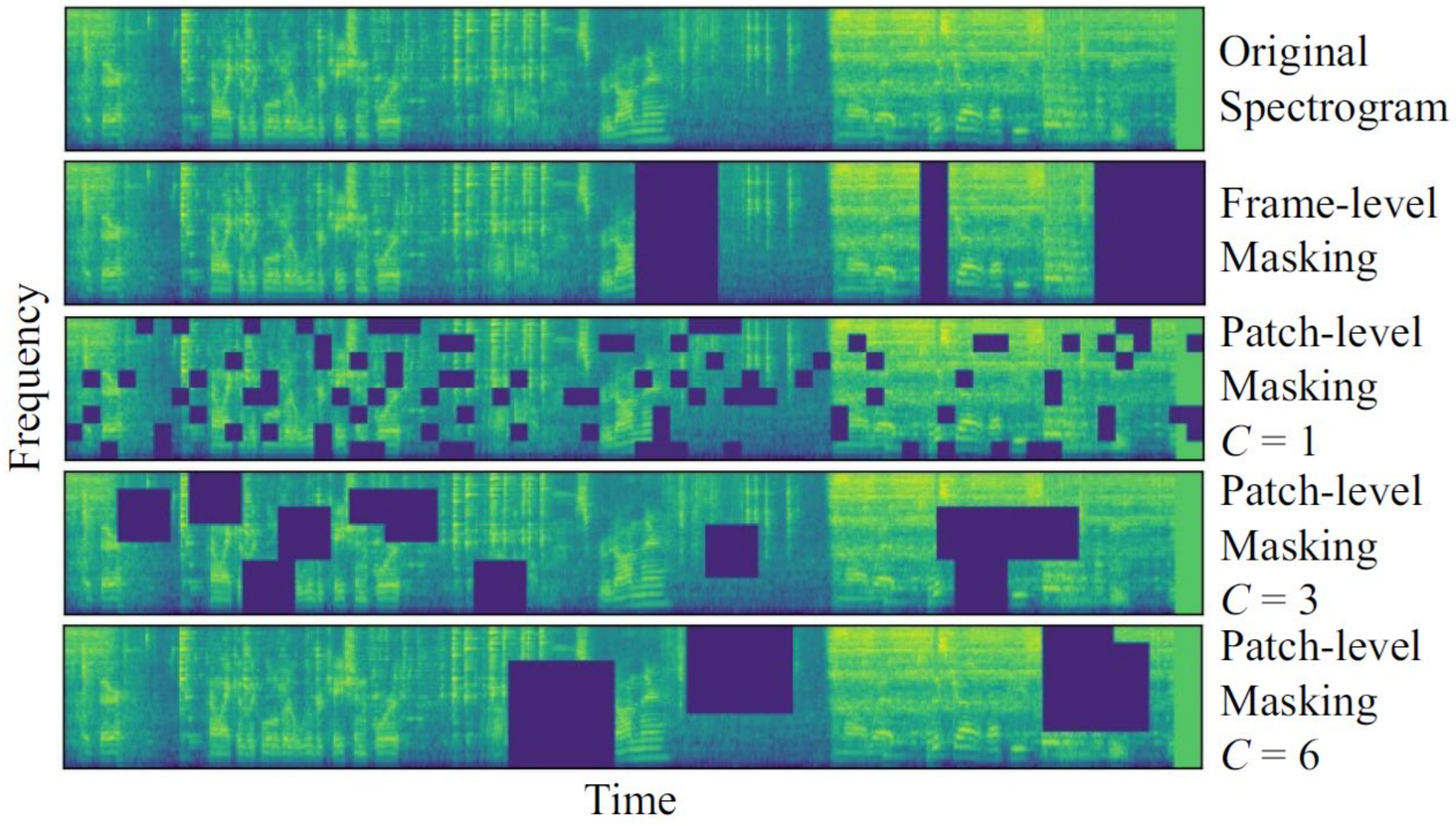
Real-world frequency data is abundant, but human annotations are scarce and expensive

The Solution – SSAST



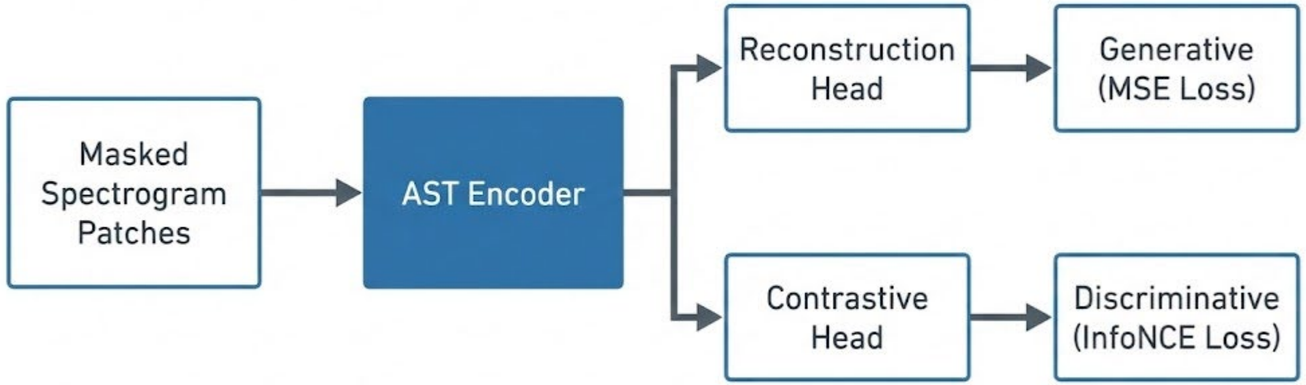
- No ImageNet Pretraining
- Works for both Audio & Speech
- Any Transformer Architecture
- Better Results than Supervised Pre-training

MSPM: Masked Spectrogram Patch Modeling



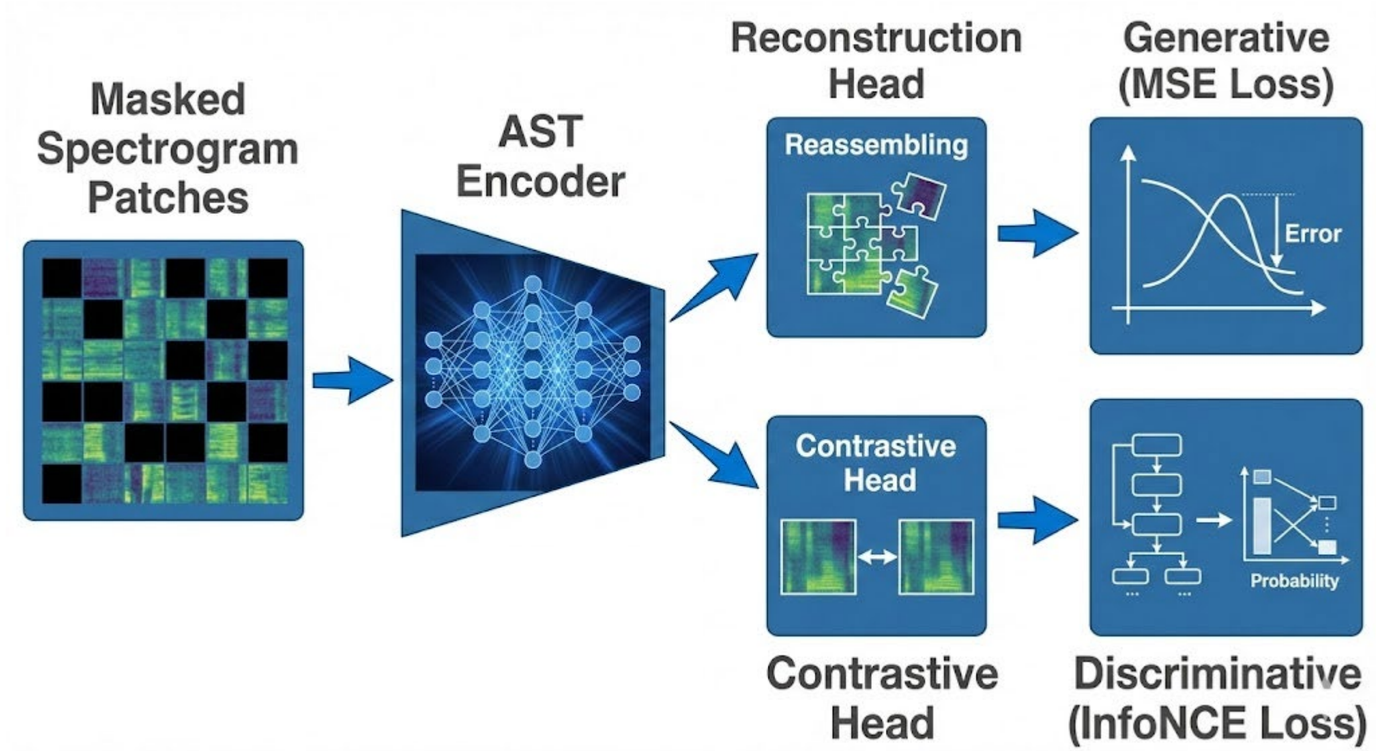
- **Pretext Task:** Predict missing spectrogram patches
- **Unstructured Masking:** Randomly drop individual patches
- **Structured Masking:** Drop whole time/frequency bands
- **Reconstruction:** Teaches global signal semantics

Joint Learning Objective



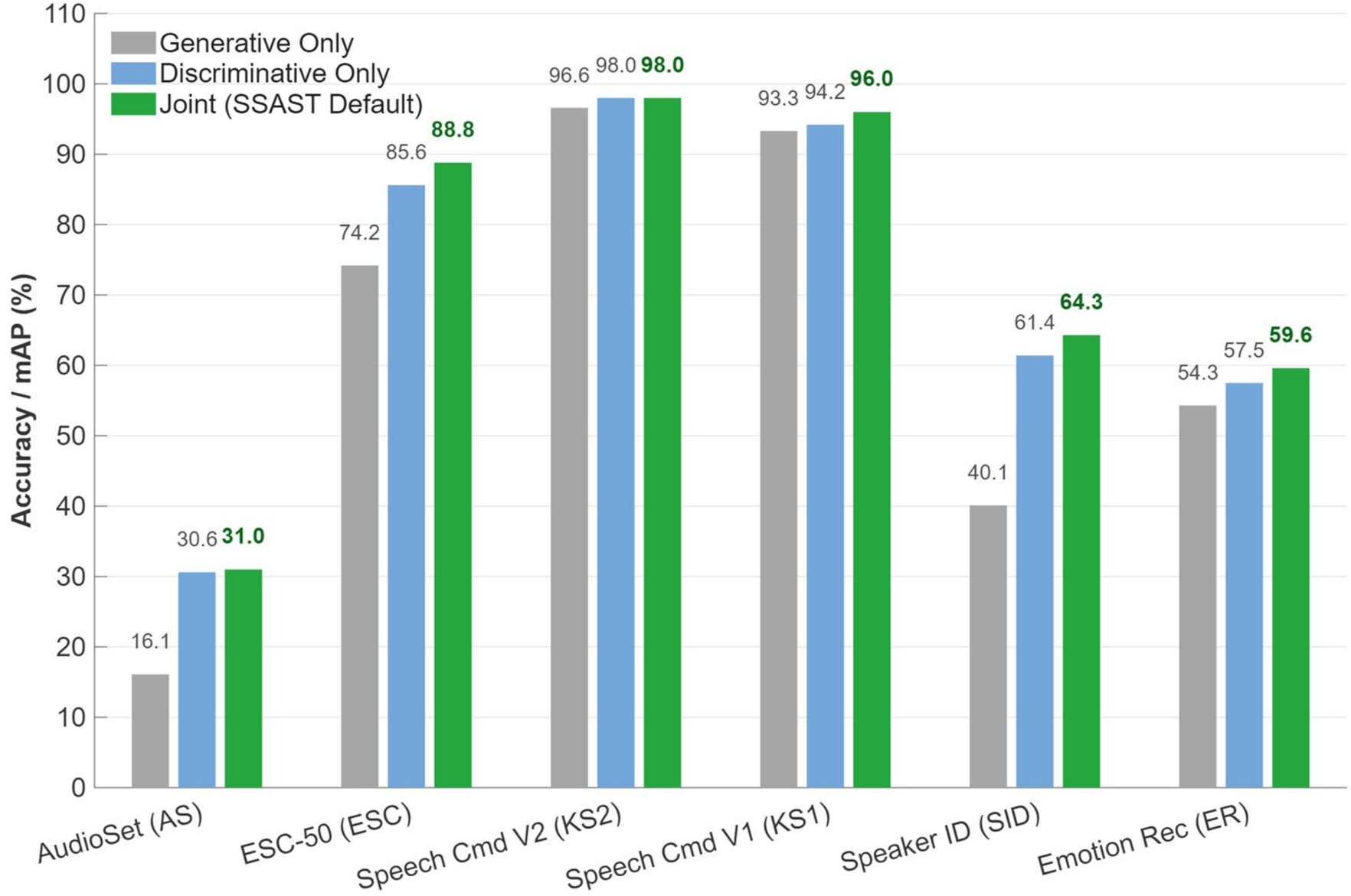
- **Dual Objective:** Combine generative and discriminative tasks
- **Generative (MSE):** Reconstruct exact masked patch values
- **Discriminative (InfoNCE):** Contrastive learning for global context
- **The Synergy:** Richer representations without human labels

Joint Learning Objective

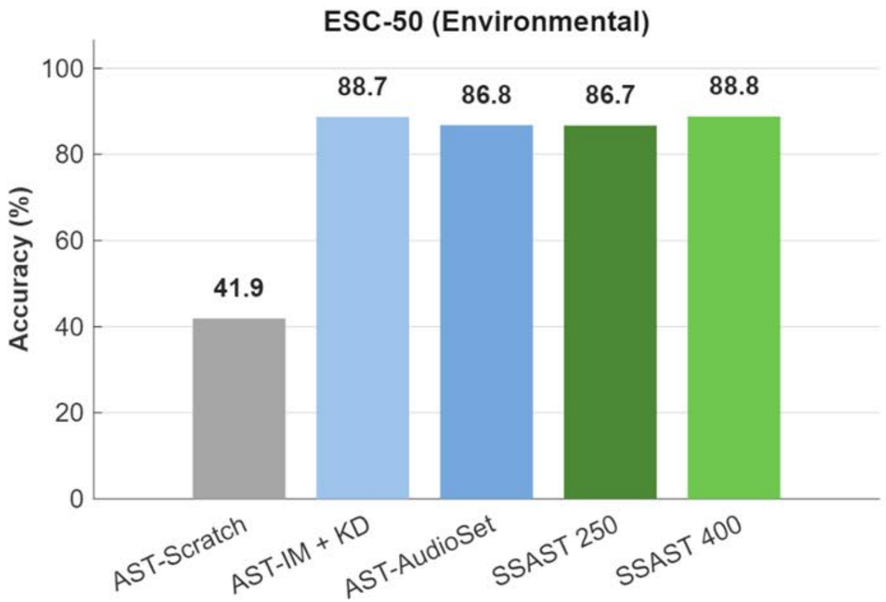


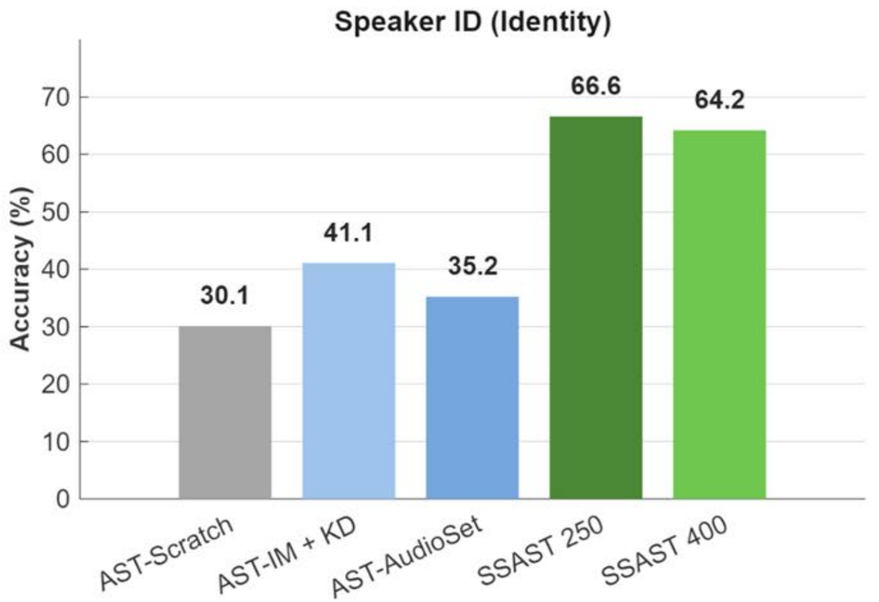
- **Dual Objective:** Combine generative and discriminative tasks
- **Generative (MSE):** Reconstruct exact masked patch values
- **Discriminative (InfoNCE):** Contrastive learning for global context
- **The Synergy:** Richer representations without human labels

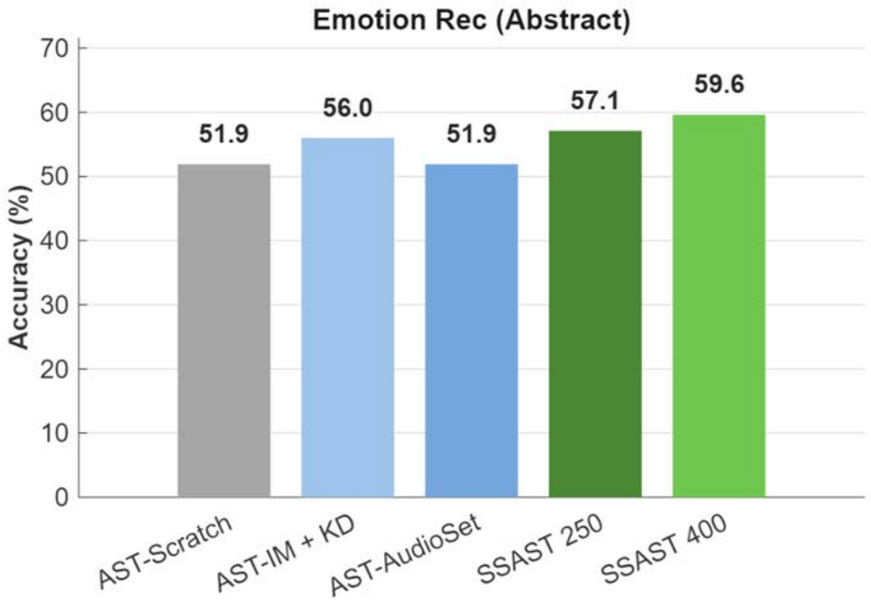
Ablation Study

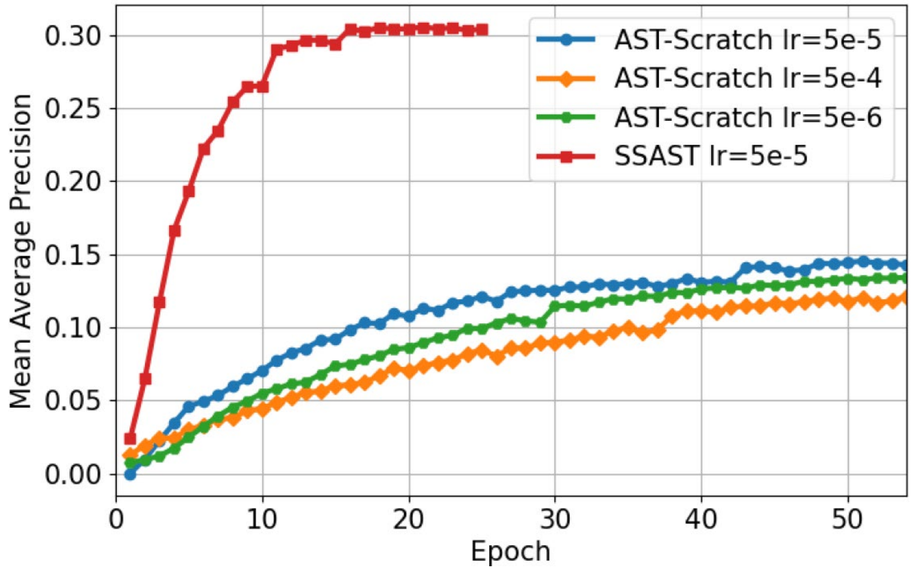


SSAST vs AST: Performance Comparison









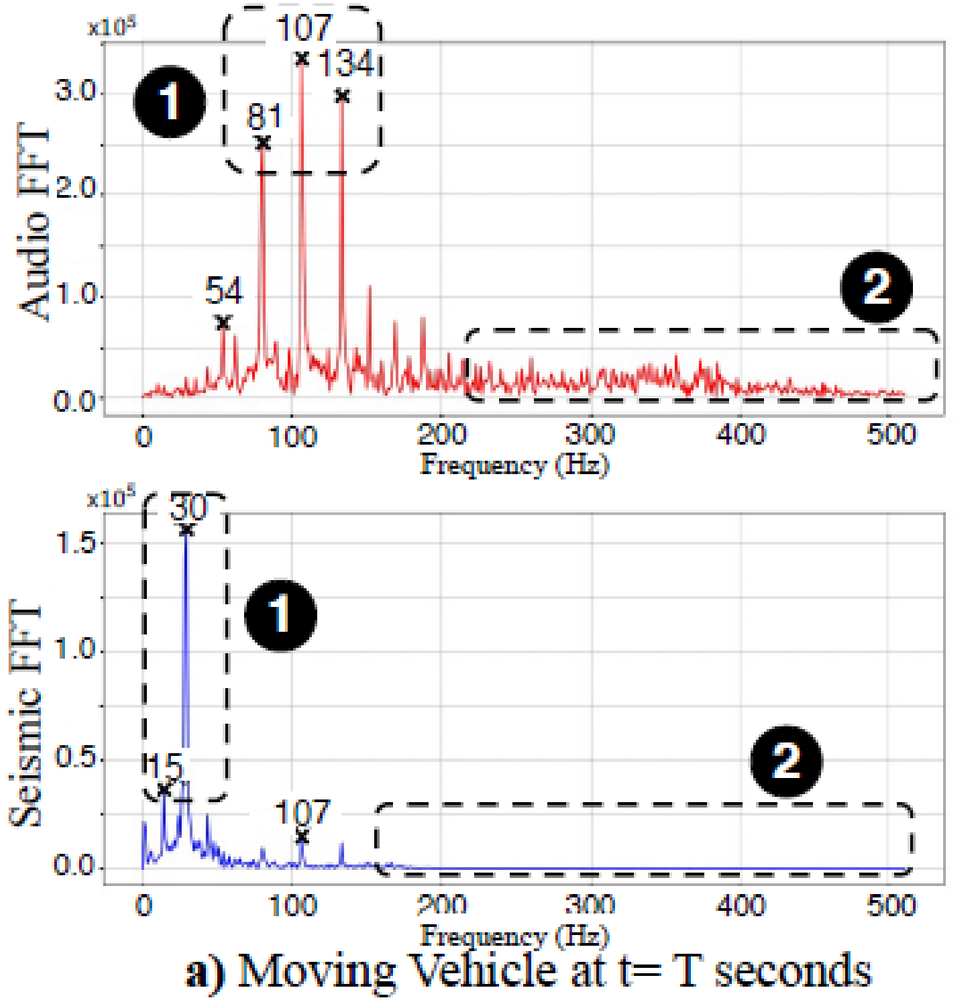
Key Takeaways & Practical Nuances

01	**Universality Beyond Audio**	• The 2D patch-masking paradigm extends natively to other frequency-domain signals (e.g., WiFi CSI, seismic data)
02	**The Phase Information Loss**	• Standard spectrograms discard signal phase, which is a critical limitation for complex wireless and spatial sensing tasks
03	**Deployment Computational Cost**	• The quadratic scaling of ViT attention makes real-time, resource-constrained edge inference highly prohibitive
04	**Preprocessing Brittleness**	• The patch semantics are rigidly tied to the specific STFT hyperparameters (window size, hop length) used during pre-training

Could it improve?

We can:

- Inject frequency-domain priors
- Improve with limited labels
- Make it more suitable for IoT time-series



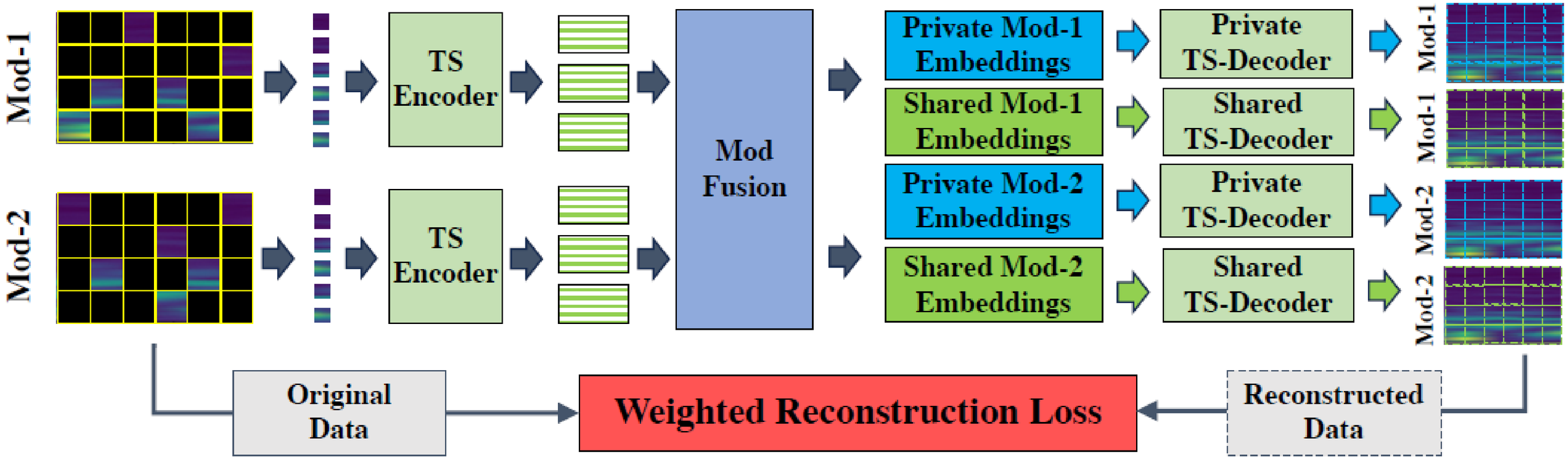
a) Moving Vehicle at t= T seconds

Frequency-Aware MAE

Design Changes:

- Local window attention (frequency-aware modelling)
- Temporal window shifting (capture drift)
- Frequency-weighted reconstruction loss
- Factorized multimodal fusion

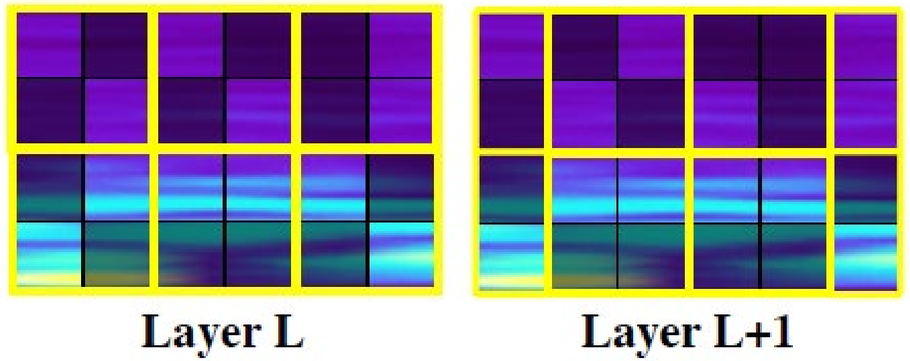
Workflow Implementation



Modelling Local Structure and Temporal Drift

Methodology:

- Attention to local frequency windows
- Windows shift across time in next layer

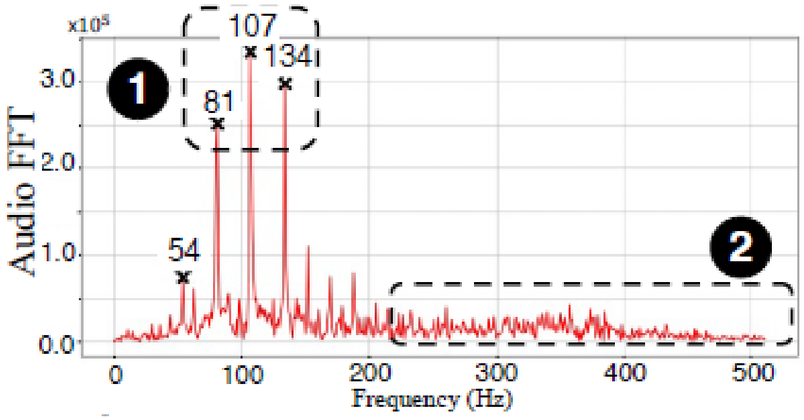


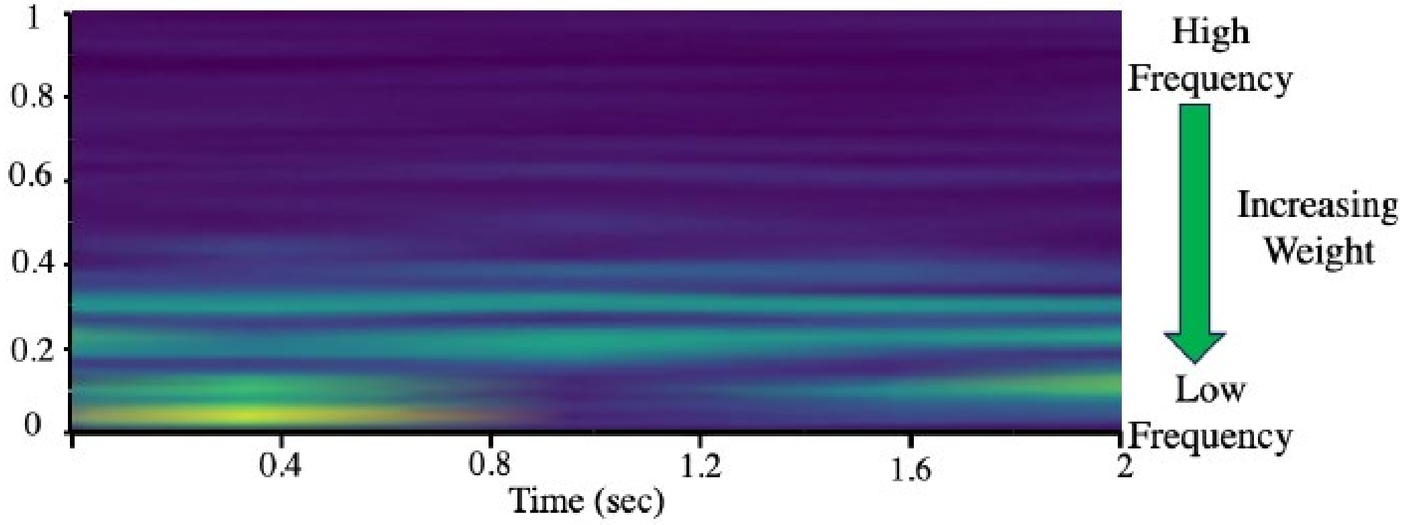
Benefits:

- Captures harmonic drift
- Reduces computation

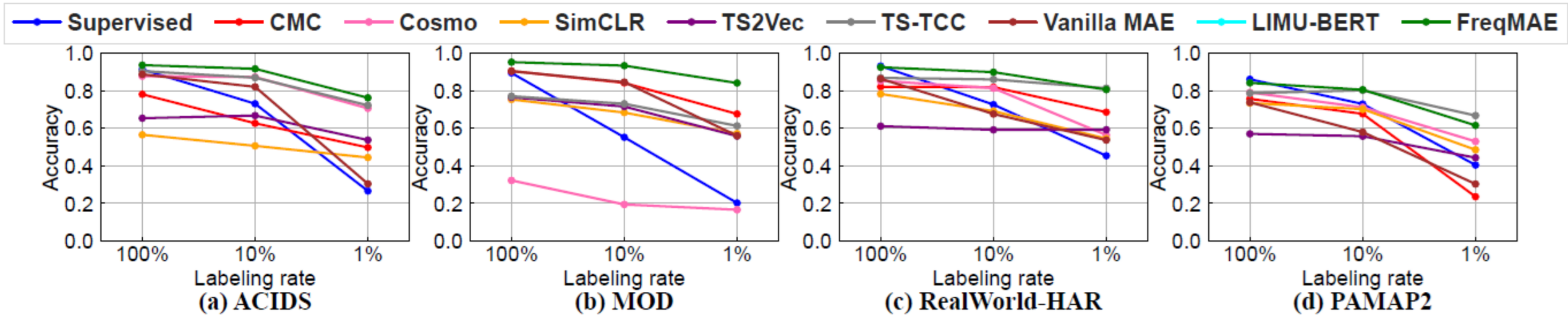
Not All Frequencies Are Equally Informative

- Uniform MSE assumes equal importance
- Information usually band concentrated
- Weighted loss prioritizes informative regions



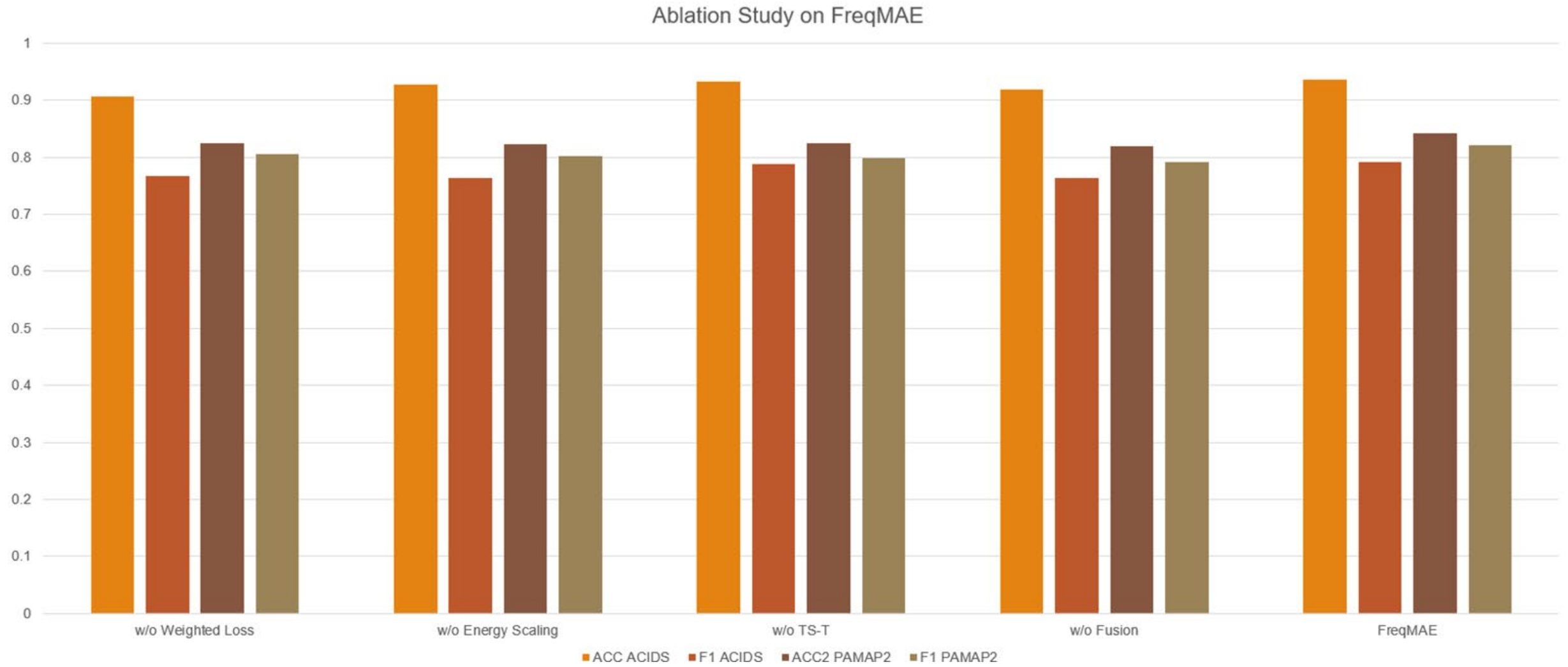


Does Frequency-Aware MAE Help?



- Better performance under label scarcity

Does Frequency-Aware MAE Help?



Still Static Assumptions

- Masking remains random
- Frequency weighting predefined
- Importance estimated heuristically

- Can masking itself become adaptive?

PhyMask: Physically-Informed Adaptive Masking

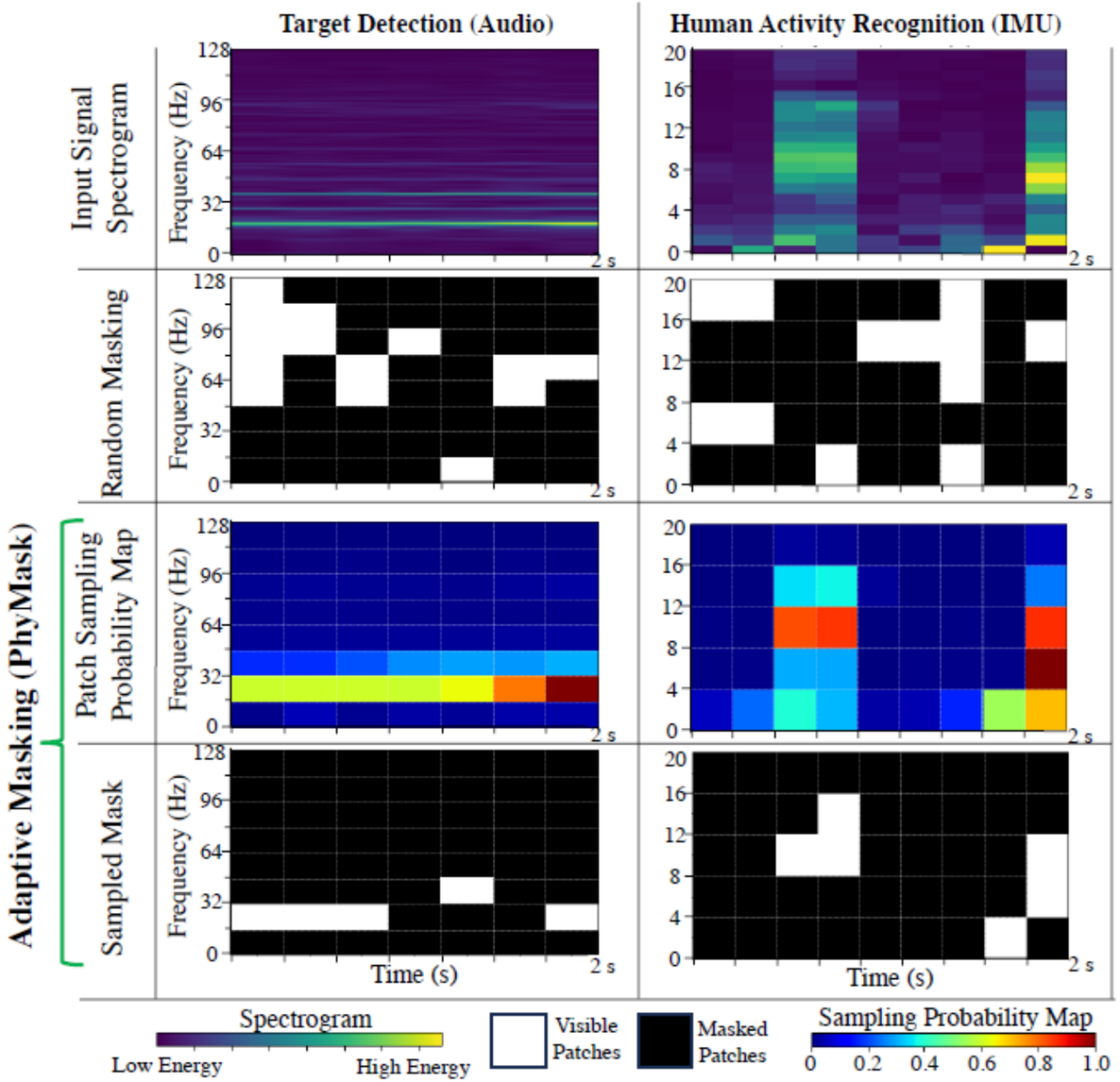
Instead of weighting reconstruction loss, we can:

- Adapt masking based on signal characteristics
- Masking becomes data-aware
- Reduce wasted reconstruction effort

Static Masking vs Adaptive Masking

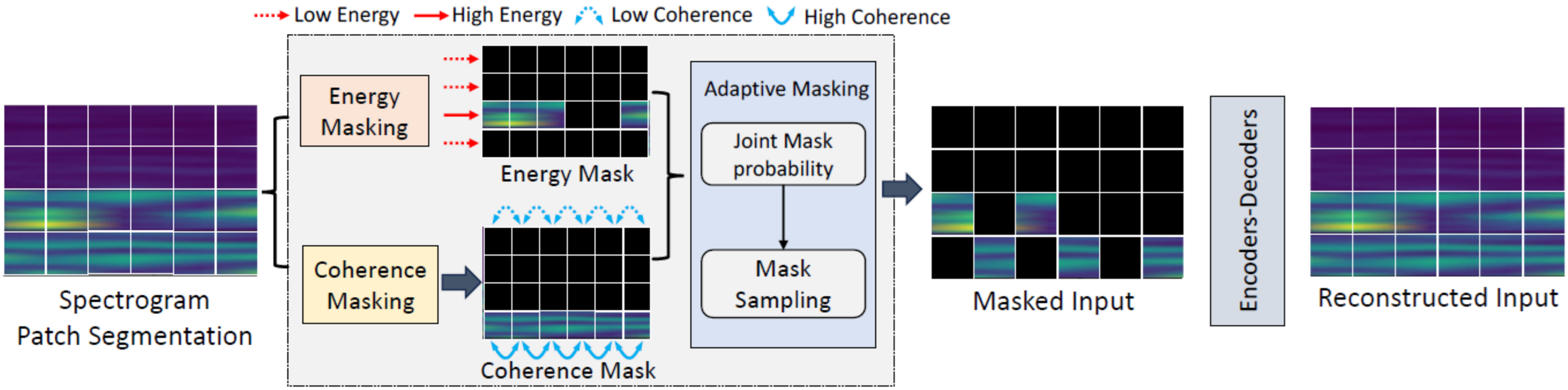
- Random masking treats all patches equally
- Frequency-weighted loss still reconstructs everything
- Important regions should be reconstructed more often
- Uninformative regions shouldn’t dominate training

Physically-Guided Mask Selection

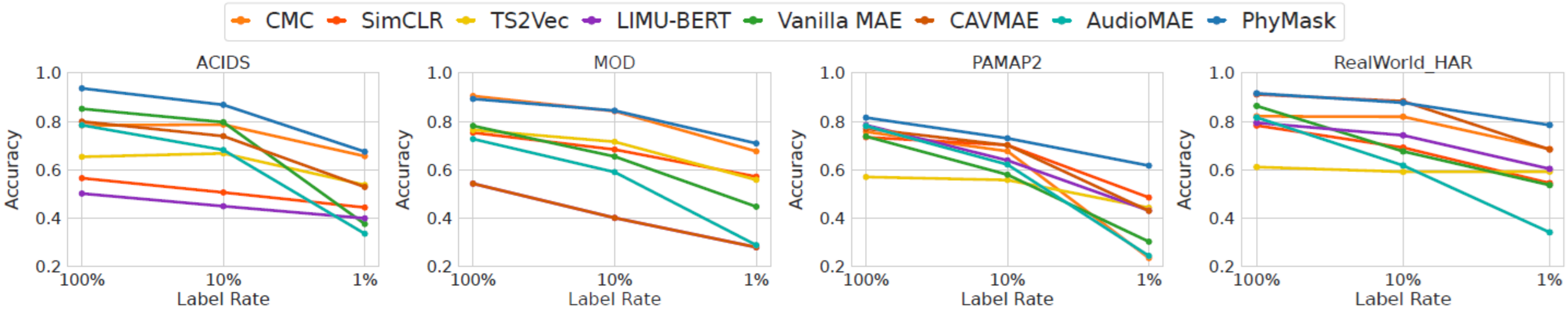


- Estimate patch importance (energy / physical signal metrics)
- Adjust masking probability dynamically
- Harder regions masked more often
- Easy/noisy regions masked less often

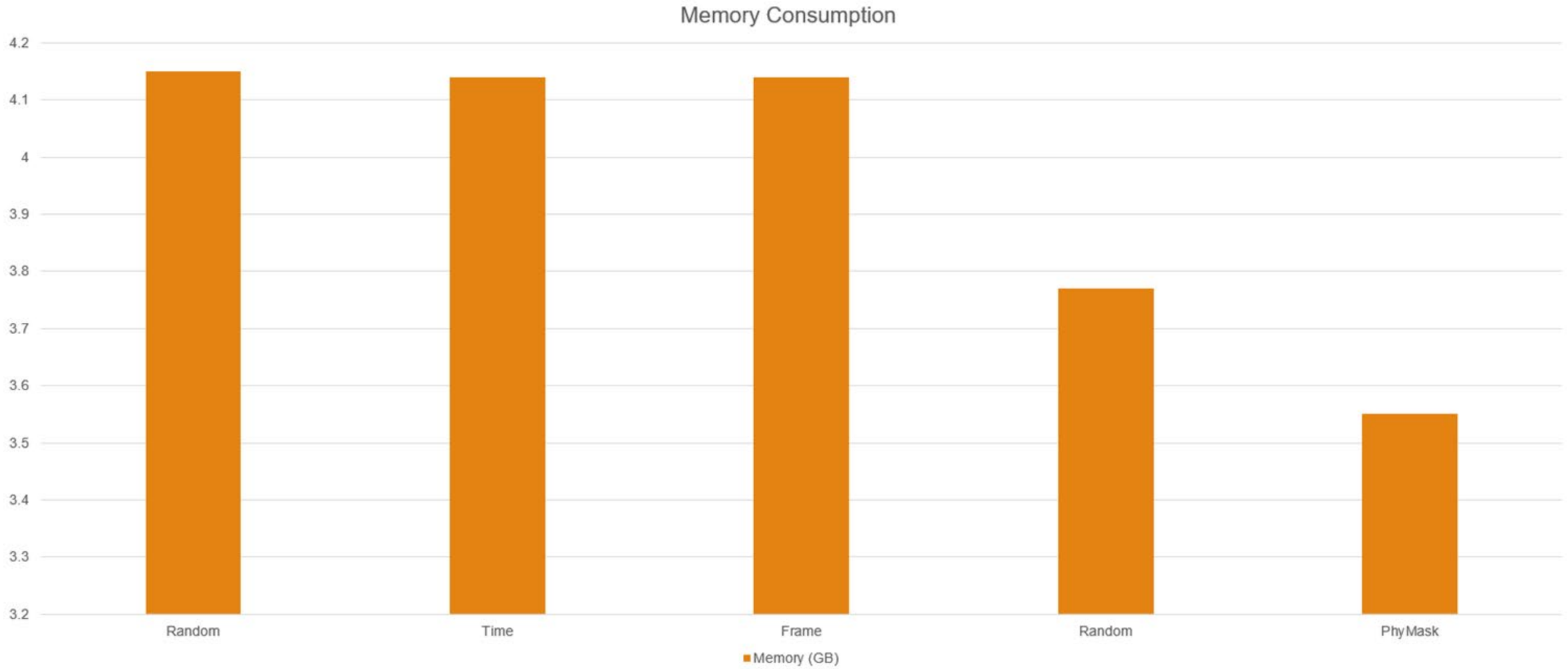
Workflow Implementation



Does PhyMask Help?



Does PhyMask Help?



Efficiency and Edge Considerations

- Avoid reconstructing mostly noise regions
- Reduce unnecessary computation
- Improve convergence speed
- More deployment-friendly

Adaptive Masking: Benefits and Risks

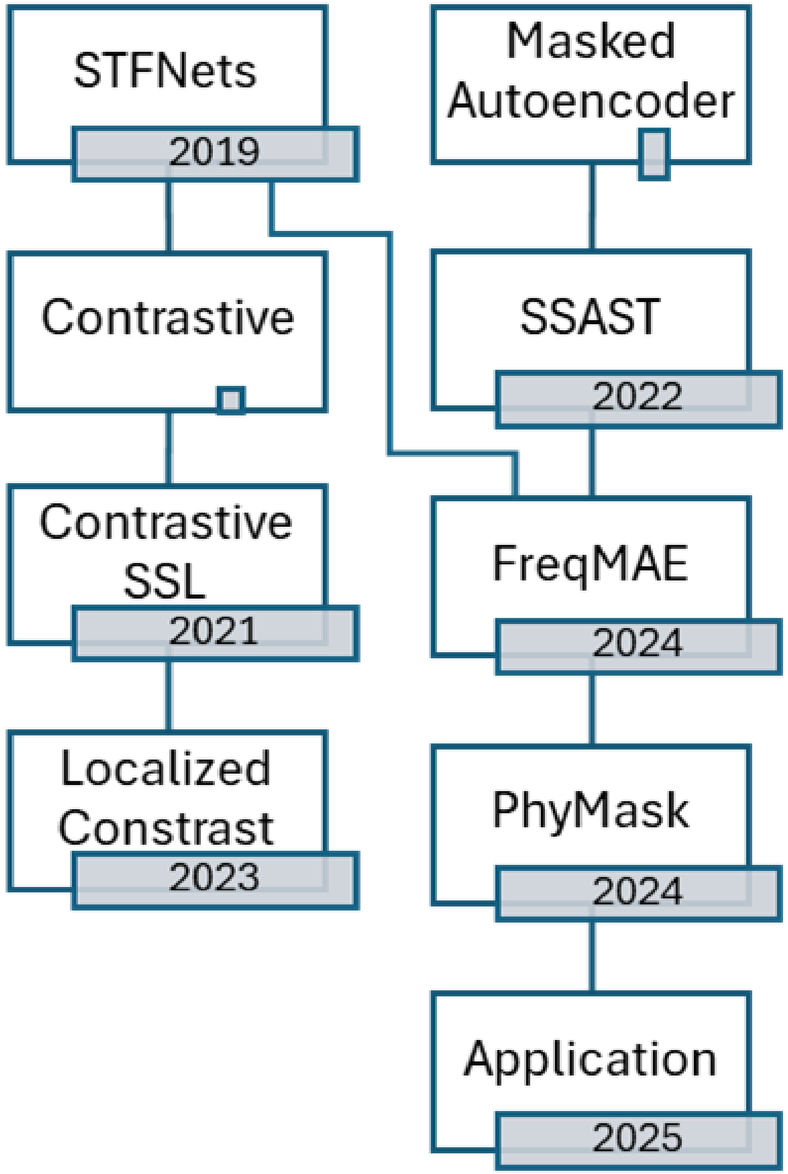
Benefits:

- More versatile than static weighting
- Less assumption-driven
- Better resource allocation

Risks:

- Importance estimation may be wrong
- Could bias toward dominant patterns
- May suppress rare events
- Adds complexity

Synthesis

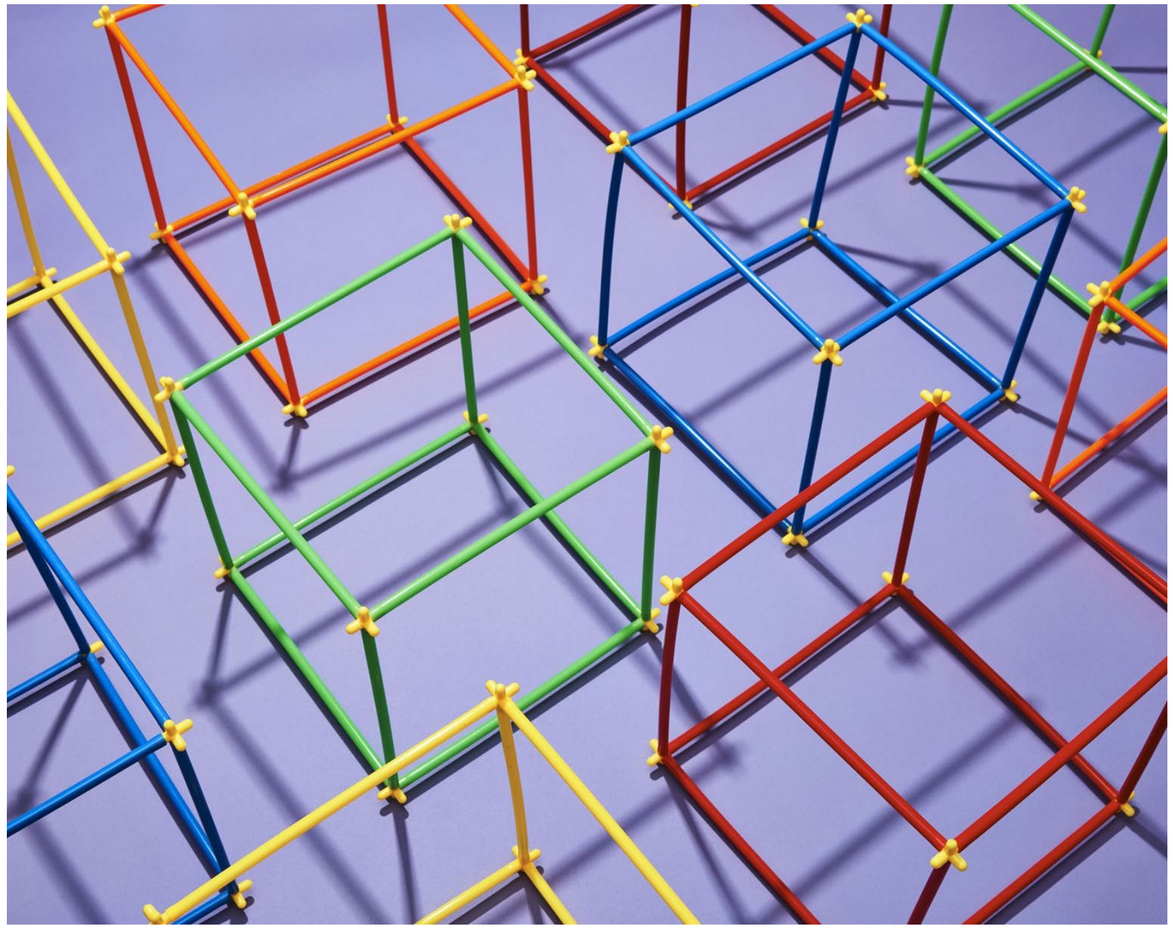


Two Paradigms, One Direction

What Was the Field Actually Solving?

Structure:

- Respect signal structure
- Learn without annotation
- Improve representation granularity
- Improve stability and scalability
- Optimize for deployment



Q&A

References

Papers Lists

1. Shuochao Yao, Ailing Piao, Wenjun Jiang, Yiran Zhao, Huajie Shao, Shengzhong Liu, Dongxin Liu, Jinyang Li, Tianshi Wang, Shaohan Hu, Lu Su, Jiawei Han and Tarek Abdelzaher, "STFNets: Learning Sensing Signals from the Time-Frequency Perspective with Short-Time Fourier Neural Networks," In Proc. The Web Conference (WWW), San Francisco, CA, May 2019.

2. Dongxin Liu, Tianshi Wang, Shengzhong Liu, Ruijie Wang, Shuochao Yao, and Tarek Abdelzaher. "Contrastive self-supervised representation learning for sensing signals from the time-frequency perspective." In 2021 International Conference on Computer Communications and Networks (ICCCN), pp. 1-10. IEEE, 2021.

3. Yuan Gong, Cheng-I. Lai, Yu-An Chung, and James Glass. "Ssast: Self-supervised audio spectrogram transformer." In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, no. 10, pp. 10699-10709. 2022

4. Setareh Rahimi Taghanaki, Michael Rainbow, and Ali Etemad. "Self-supervised human activity recognition with localized time-frequency contrastive representation learning." IEEE Transactions on Human-Machine Systems 53, no. 6 (2023): 1027-1037.

5. Denizhan Kara, Shengzhong Liu, Jinyang Li, Dongxin Liu, Tianshi Wang, Ruijie Wang, Yizhuo Chen, Yigong Hu, Tarek Abdelzaher, "FreqMAE: Frequency-Aware Masked Autoencoder for Multi-Modal IoT Sensing," In Proc. The Web Conference (WWW), May 2024.

6. Denizhan Kara, Tomoyoshi Kimura, Yatong Chen, Jinyang Li, Ruijie Wang, Yizhuo Chen, Tianshi Wang, Shengzhong Liu, Lance Kaplan, Joydeep Bhattacharyya, Tarek Abdelzaher, "PhyMask: An Adaptive Masking Paradigm for Efficient Self-Supervised Learning in IoT," In Proc. 22nd ACM Conference on Embedded Networked Sensor Systems (SenSys), Hangzhou, China, November 2024.