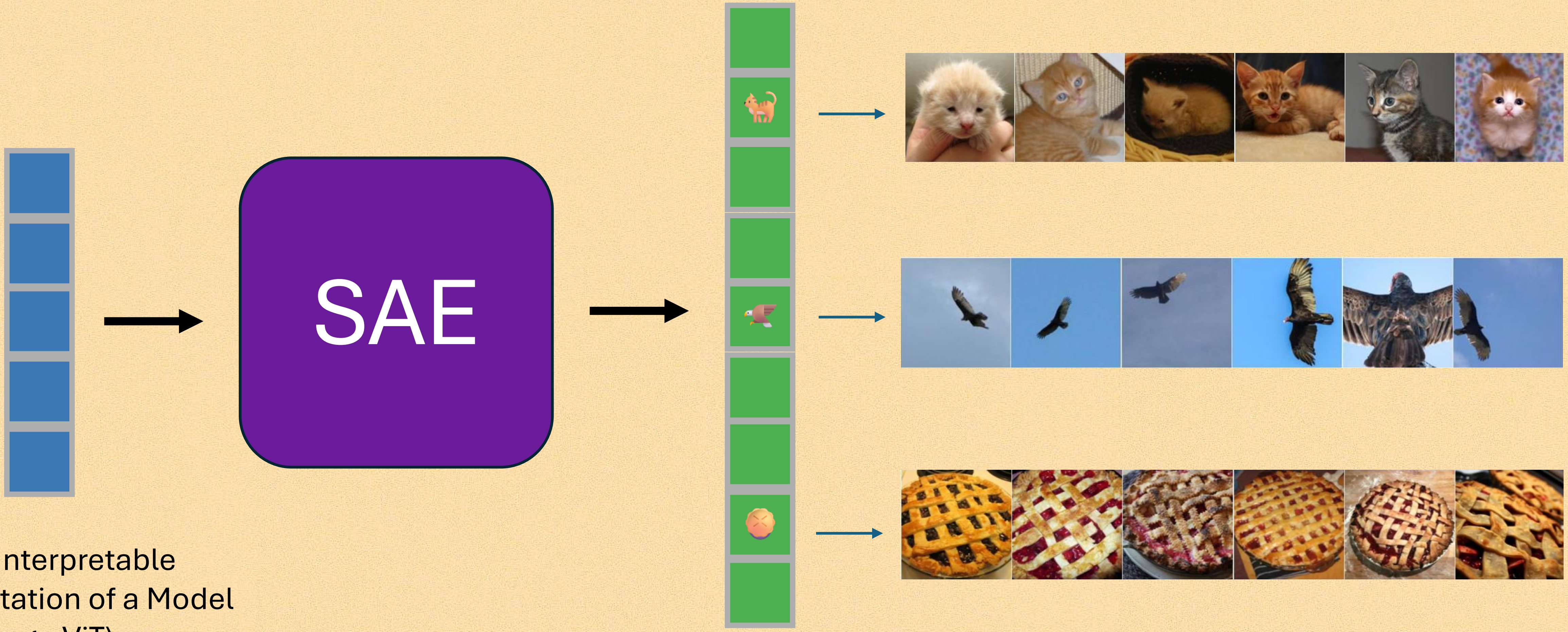
SPARC

Concept-Aligned Sparse Autoencoders for Cross-Model and Cross-Modal **Interpretability**

*"Scaling **Interpretability** across models"*

Ali Nasiri-Sarvi • Hassan Rivaz • Mahdi S. Hosseini
Concordia University • Mila-Quebec AI Institute

Sparse Autoencoders Make Neural Representations Interpretable



SAE

Non-Interpretable Representation of a Model (e.g., ViT)

Interpretable SAE Representation

What if we want to understand the representation of multiple models?

🔍 SAEs Help… But Only One Model at a Time (comparing many gets expensive)

✅ **Comparing** Metrics? Easy.

Model A	94.2%
Model B	91.8%
Model C	87.5%

📊 **Run benchmarks → Compare numbers**
Done!

🤔 **Understanding** what shared concepts multiple models learn? Hard.

Model A	SAE's interpretable latents	🐱	🐕		🚗	🏠		⚽		🌳	
Model B	SAE's interpretable latents		🚗		🌳	🎵		🐱	🐕		🏠
Model C	SAE's interpretable latents	🏠	⚽	🎵		🐱	🐕			🔥	

🔬 **Each SAE model outputs insights, not numbers**

🔀 **Same concepts, different latent positions across SAEs**

🧑‍💻 **Model insights should be manually compared**

⚠️ Scaling Problem

Thousands of neurons to analyze per model
Matching concepts by hand
Time-consuming, error-prone

🧠 Expert Bottleneck

Some domains require expert interpretation (e.g., medical)
High cost and **slow progress**

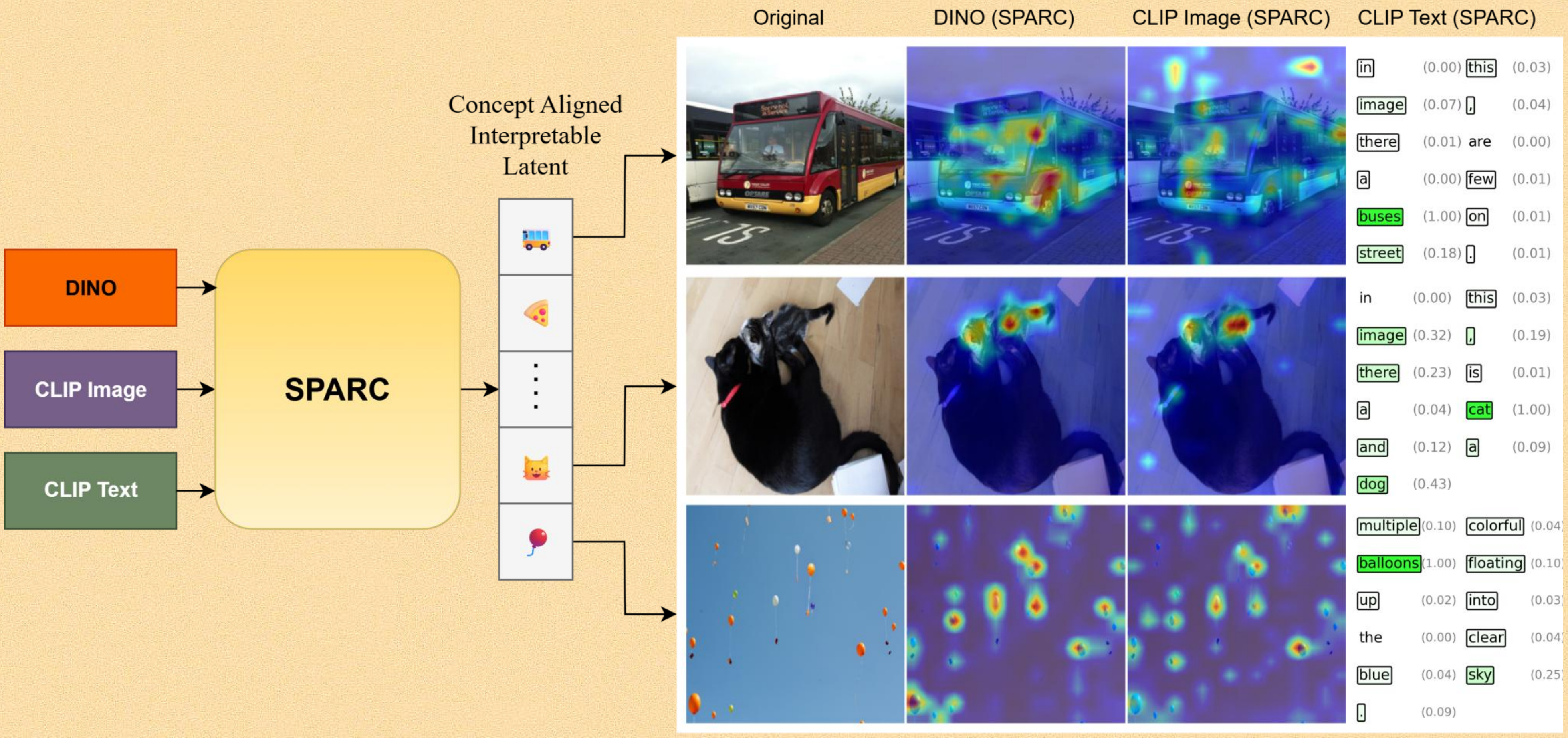
📈 O(n^2) Time Consumption

Each model needs **(n-1) comparisons**
Manual concept matching
Complexity explodes with models

SPARC: A unified solution

Sparse **A**utoencoders for **R**epresentation of **C**oncepts

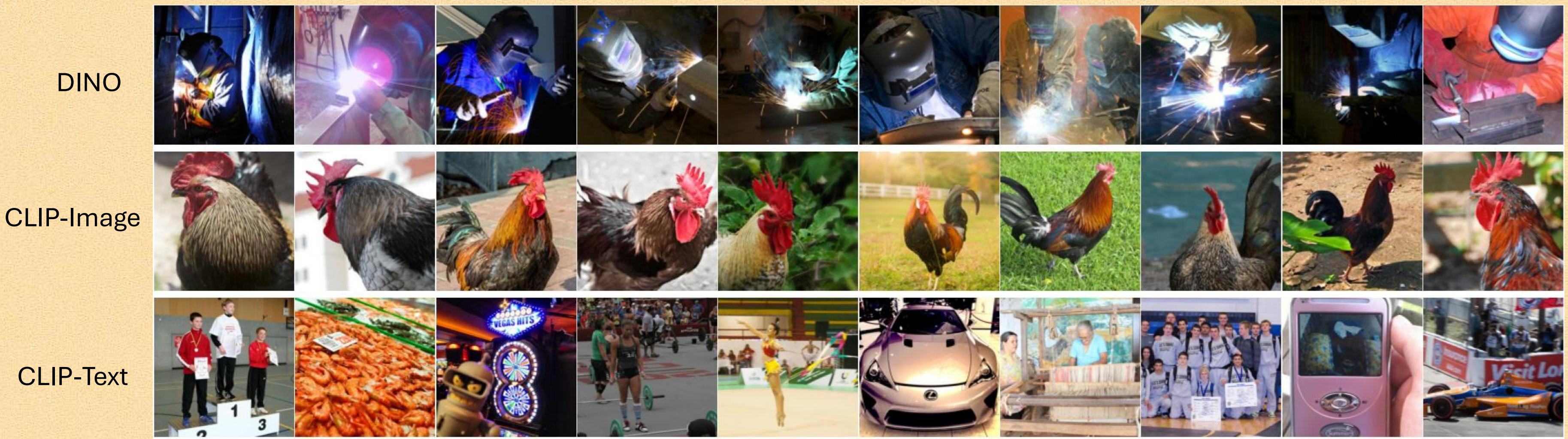
- Key idea: Build a **single, shared interpretable latent space** for multiple models
- Forces concept alignment directly through shared architecture
- Works across different model architectures and modalities (vision, language)



SPARC allows direct comparison across models and modalities without manual concept matching

SPARC aligns the learned SAE Concepts

Before SPARC: The same latent index encodes different concepts in each stream (welding, roosters, random captions) → no shared meaning.



After Applying SPARC

After SPARC: The same latent index now retrieves the same concept (kittens) in all streams.