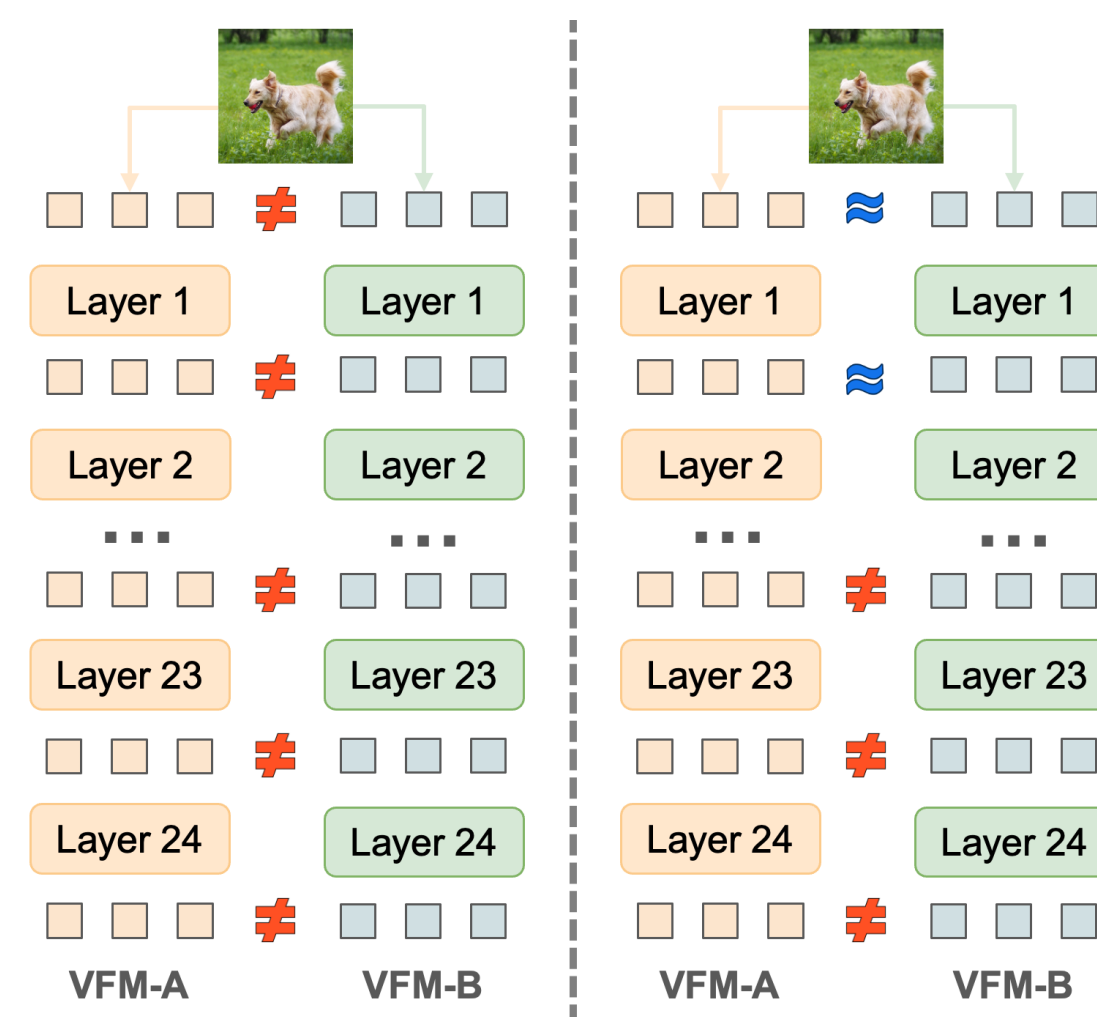


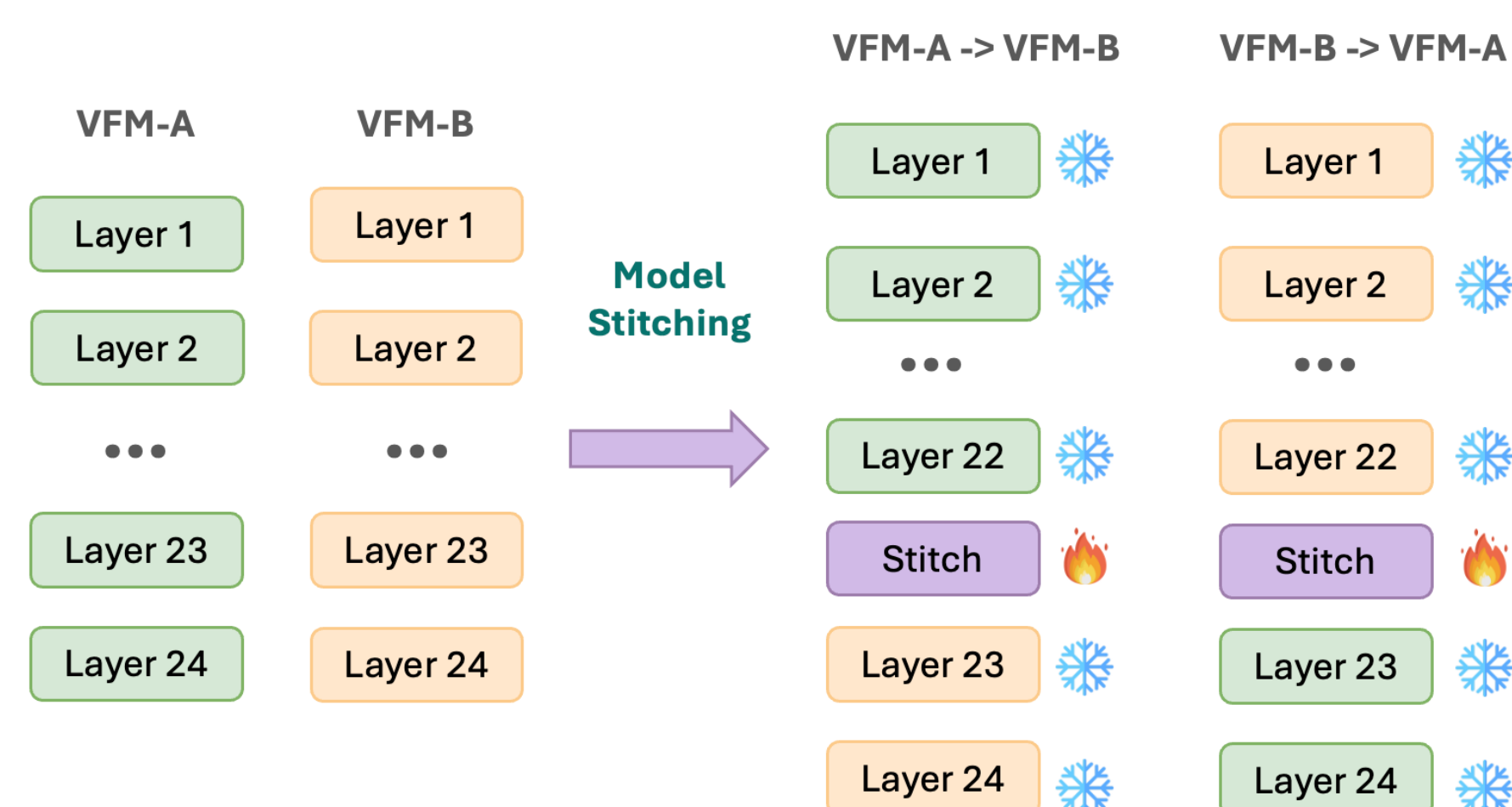
1 Overview

Vision Foundation Models (VFM)s, such as DINOv3 and SigLIP2, are trained with **different data**, **supervision**, and **objectives**.



Are VFMs fundamentally **different end-to-end**, or their internal representations are **similar** and **compatible** in some layers?

Model Stitching: connect the **shallow layers** of a model to the **deep layers** of another via a **stitch layer** (e.g., linear or MLP) to form a **stitched** model.



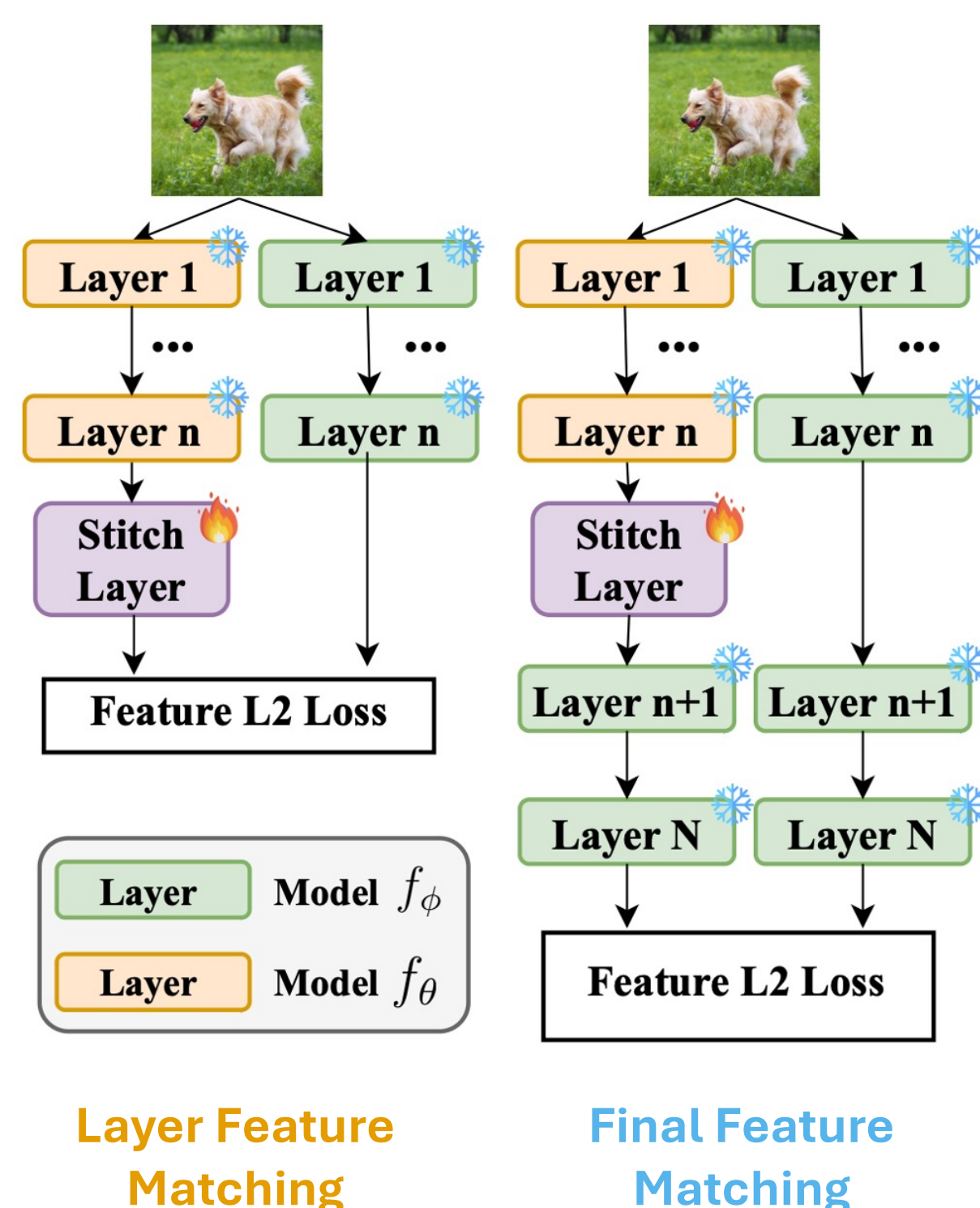
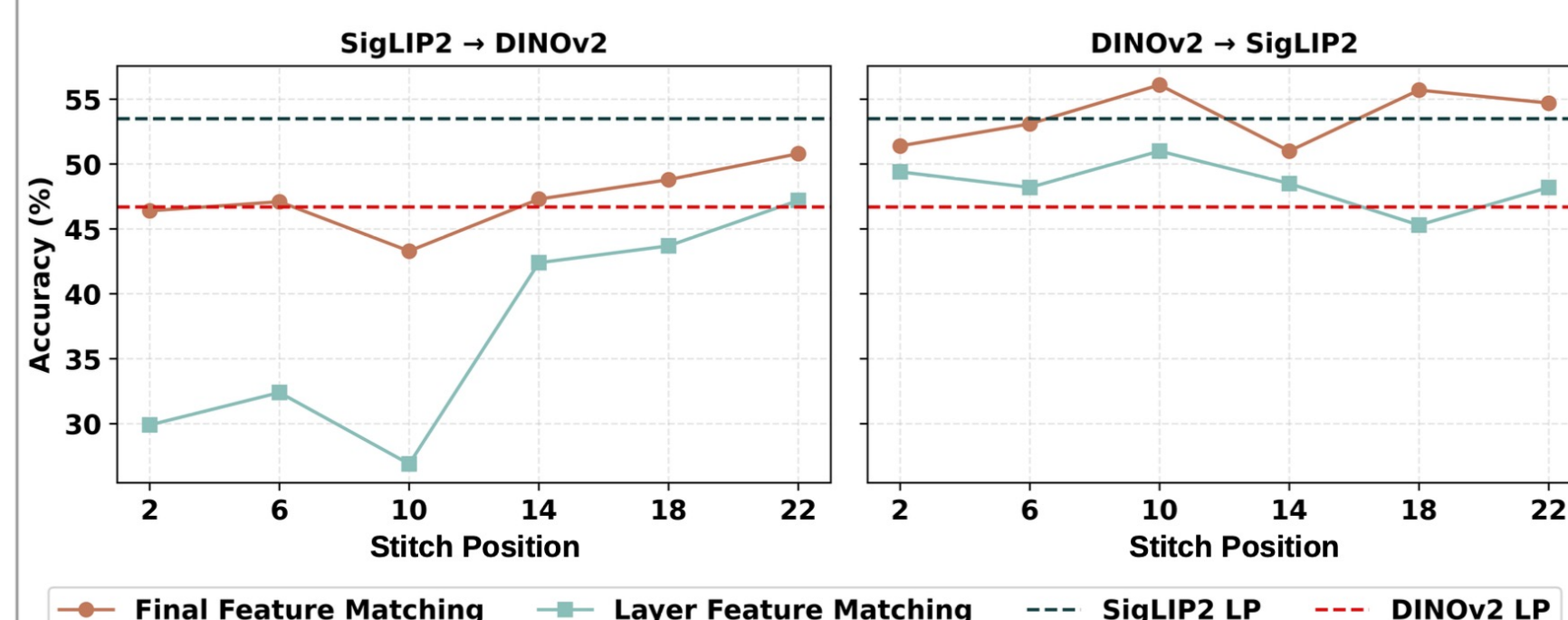
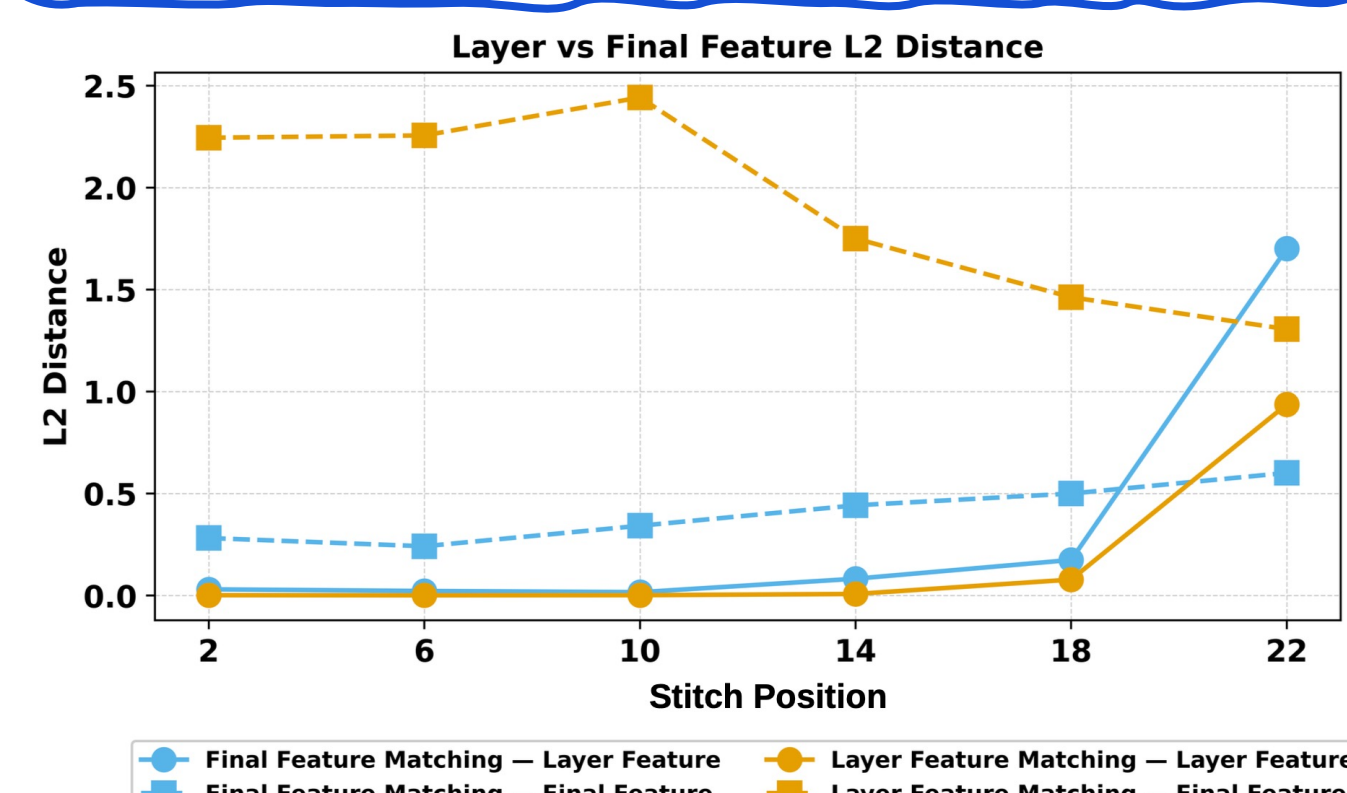
If the **stitched** model has **minimal performance drop**, it means their internal representations are **similar**.

- With appropriate stitching, different VFMs are not just **compatible** — they can be **complementary**.
- Stitching **fuse** VFMs' **strengths** and enable more **efficient** multi-VFM system.

2 Naïve Stitch Fails

Naïve Stitching

Layer Feature Matching: Match **local** features at the stitch position.
Problem: **Small** errors at the stitch position can be amplified by frozen target layers -> **large final** feature differences.



Remedy
Final Feature Matching: Match **final** features between the stitched and target models.

Naïve Stitching

Task-Loss Training: Optimize the downstream objective directly.

- Problem:**
- For shallow stitching, gradients backprop through a **long, fixed** transformation to update only the **stitch** layer.
 - Optimization challenges for **randomly initialized** **stitch** layer,

DINOv2	SigLIP2	Layer Feature Matching	Final Feature Matching	Task Loss Training
46.7%	53.4%	49.4%	51.4%	25.1%

Task-Loss Training

DINOv2 -> SigLIP2 stitched at layer 2

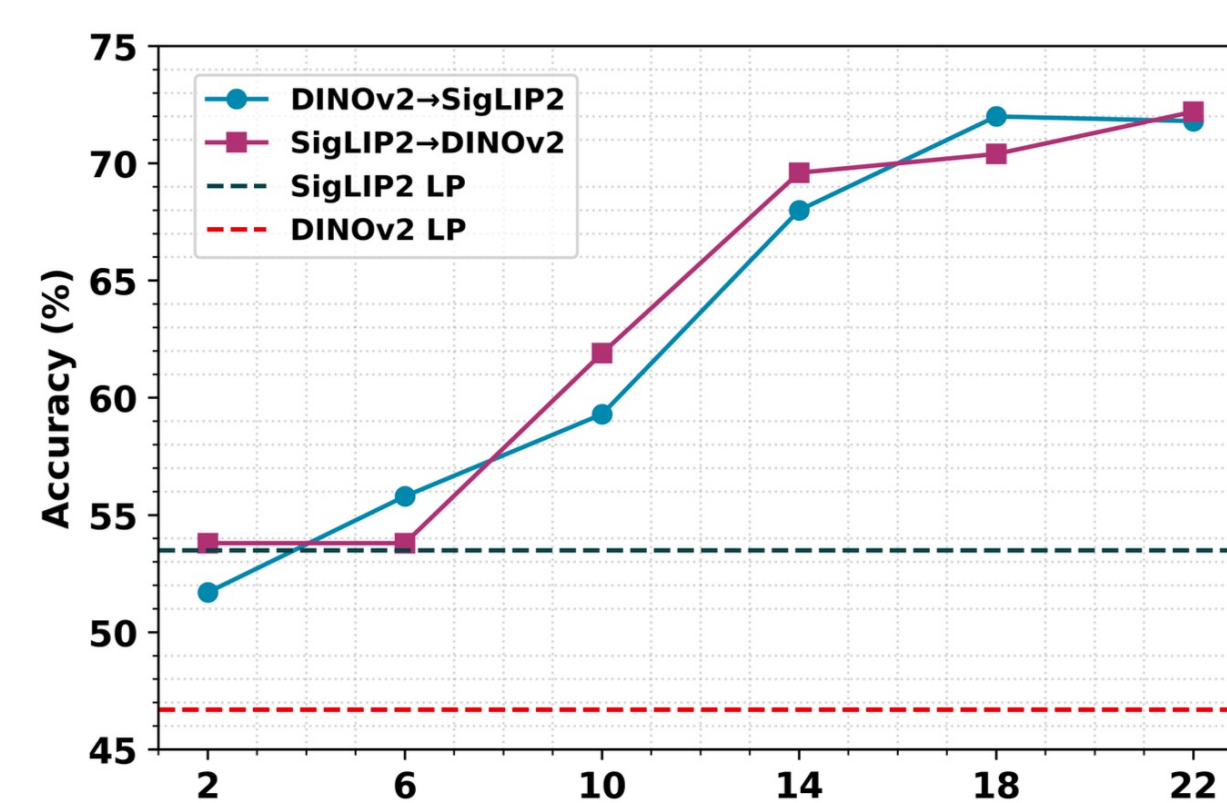
3 Our Stitching Recipe

Stage 1 Final Feature Matching Pretraining

Label-free pre-train: match the target model's final representation

Stage 2 Task-Loss Fine-Tuning

Adapt to downstream task from a strong initialization

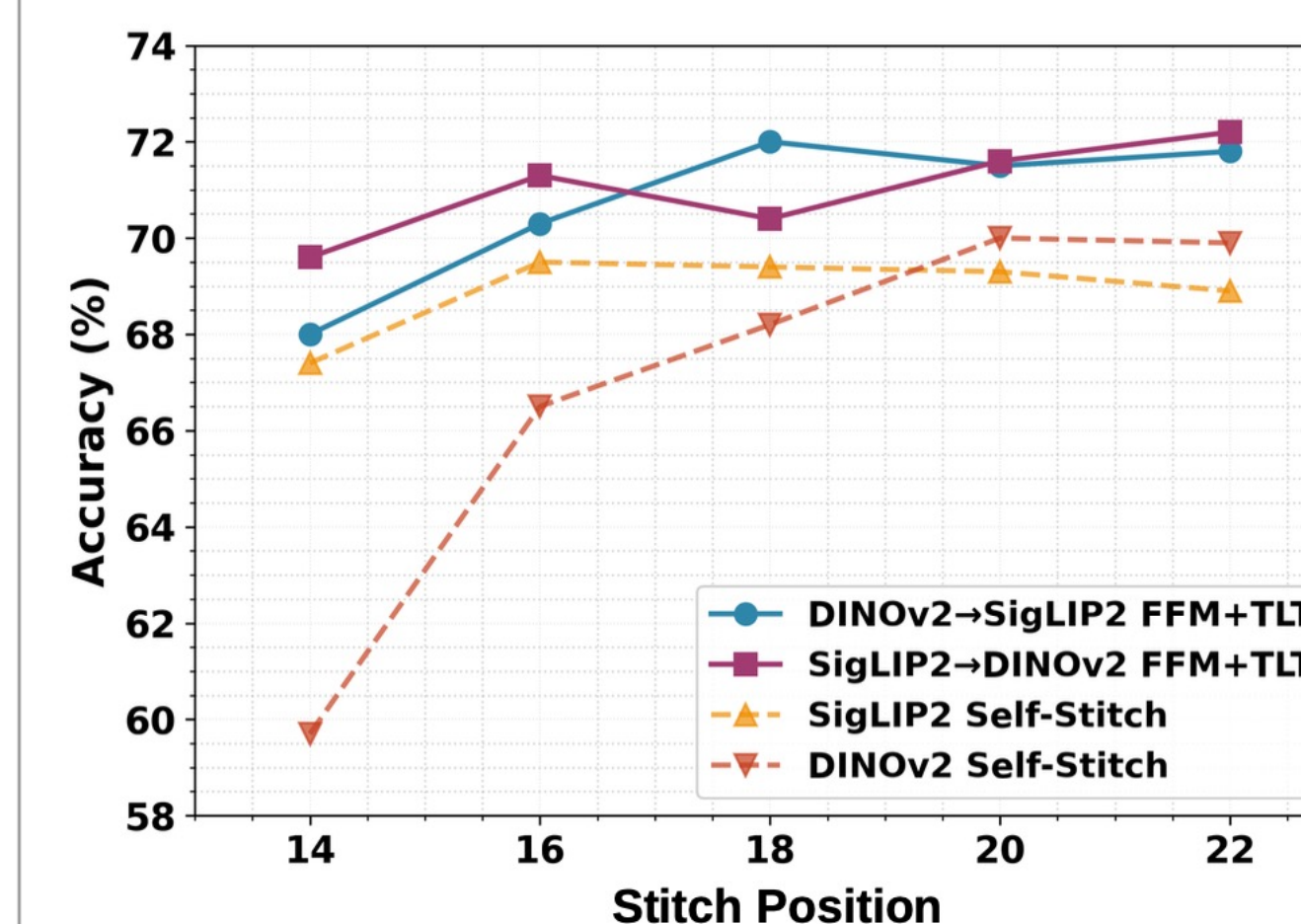


Stitch model outperform original VFMs on fMoW. LP: Linear Probing.

4 Knowledge Fusion, Not Just Capacity

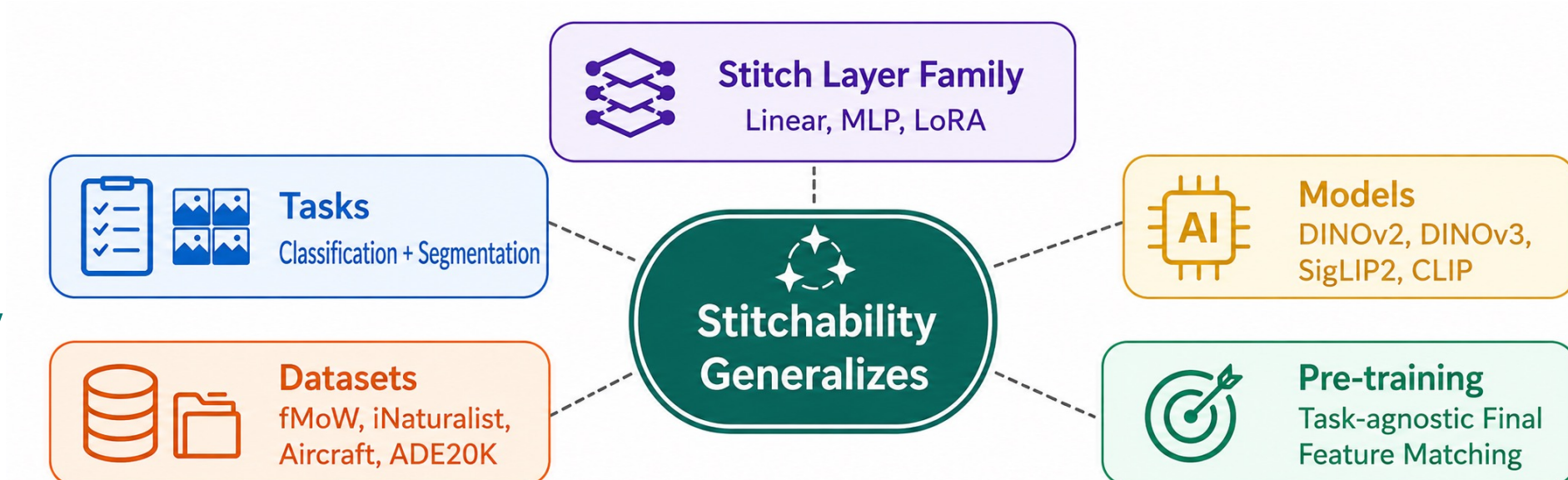
Self-Stitch Baseline:

- Insert **same** **stitch layer** at **same** **stitch position** into source and models with **same** training data, and **objective**.
- Isolate whether improvement is only from the stitch layer.

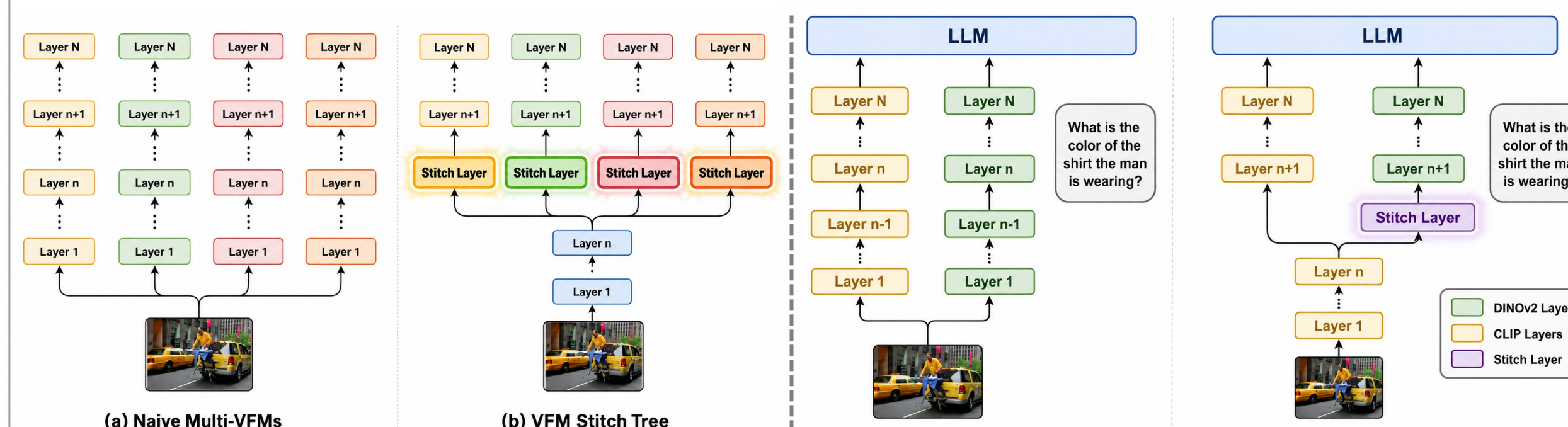


Stitching heterogeneous VFMs is not only **feasible** but also has the potential to fuse **complementarity** between them.

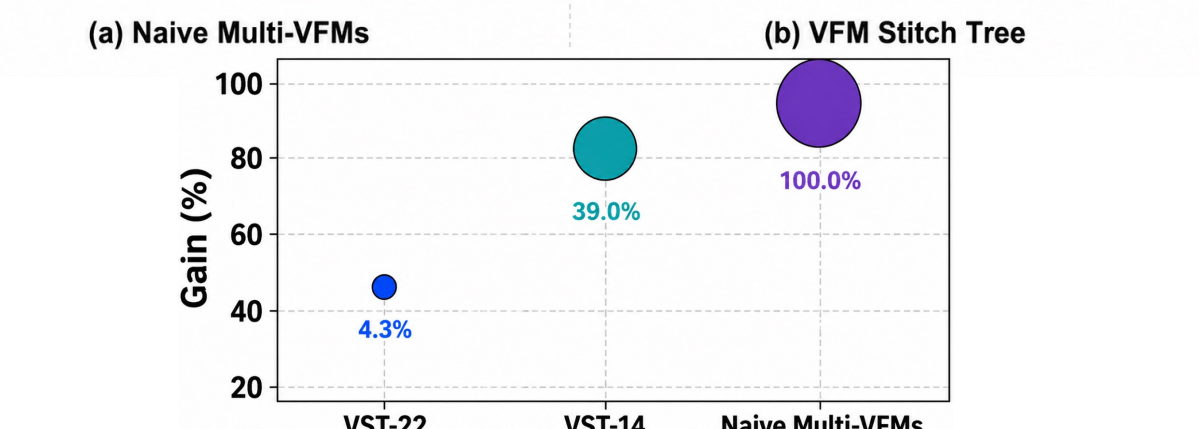
5 Generality



6 Applications: VFM Stitch Tree (VST)



VST: Share shallow layers, retain **specialized** deep branches, connected with **stitch** layer



VST in MoF-LLaVA recovers most two-VFM gains with **lower** overhead

7 Takeaways

- 1 Stitch layer **training** matters
- 2 Stitching **fuses** strengths
- 3 Stitching enables **efficient** systems

Model Stitching, from an analytical tool to a practical framework for combining the **complementary strengths** of VFMs.



Paper + Code

Work done during internship at Amazon.