Hyper-Align: Efficient Modality Alignment via Hypernetworks

Jaisidh Singh[1,2,3,5] Diganta Misra[2,3] Boris Knyazev[6] Antonio Orvieto[2,3,4]

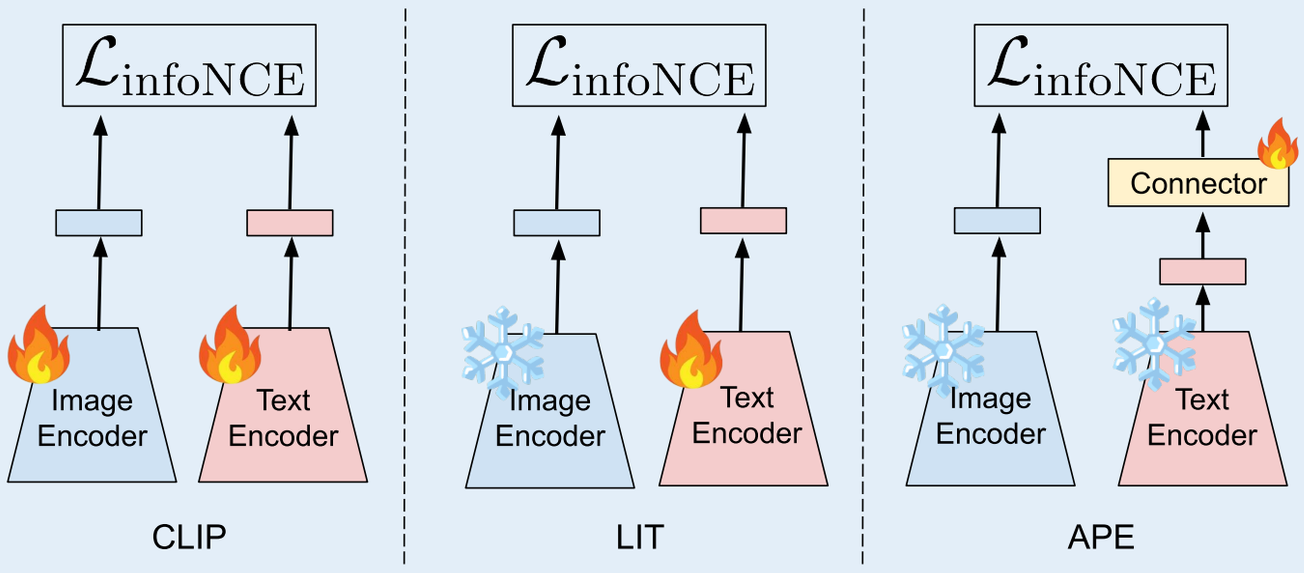
[1]University of Tübingen [2]ELLIS Institute Tübingen [3]MPI for Intelligent Systems, Tübingen
[4]Tübingen AI Center [5]Zuse School ELIZA [6]SAIT AI Lab Montreal

ICLR 2025 Workshop on Weight Space Learning

Background

Contrastive VLMs pretraining schemes can
- Train encoders end-to-end
- Train only modality connectors between pretrained encoders

APE outperforms CLIP at much lower computational cost

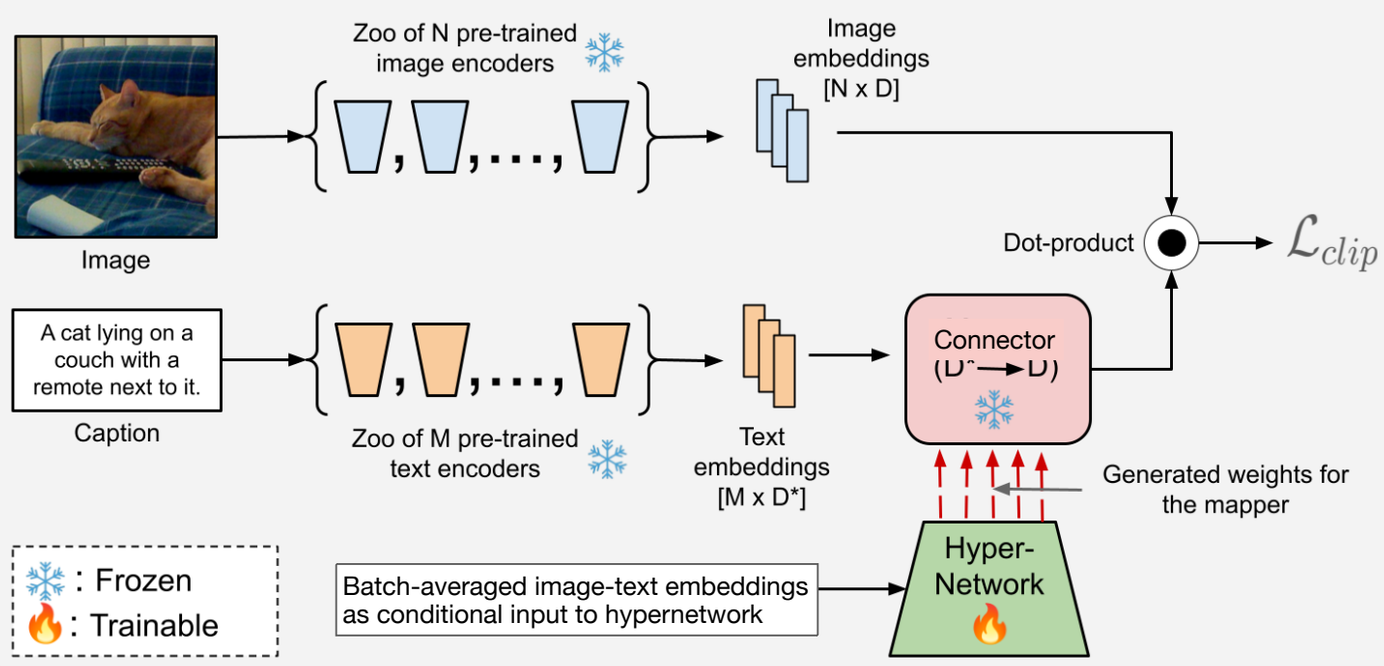


Proposed Solution: Hyper-Align

Learn $N \times M$ connectors *together* by showing a hypernetwork data from N image and M text encoders.

Result: Hyper-Align finds optimal pair in $N \times M$ combinations at 8x smaller computational budgets with negligible performance drop.

Methodology Overview

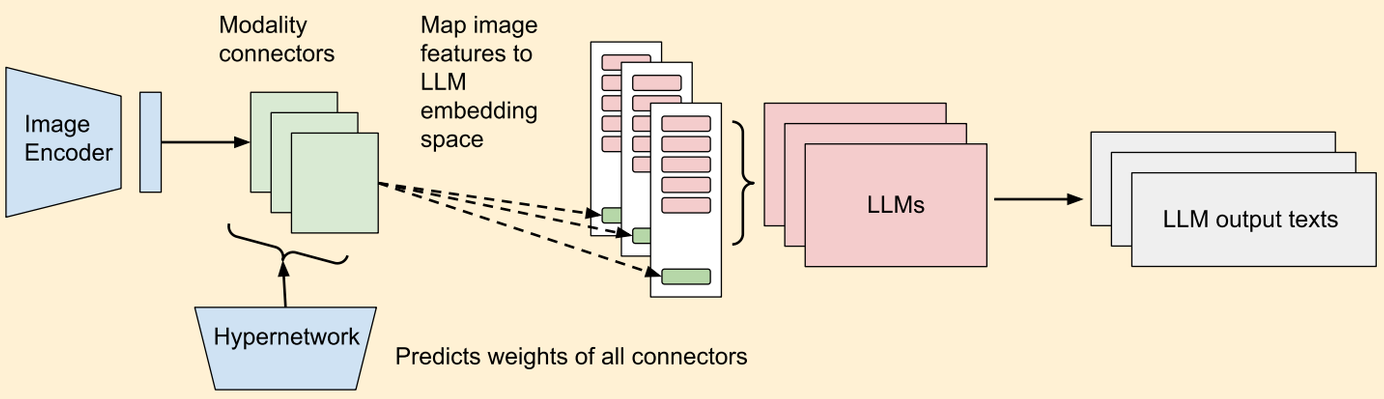


Conclusion

Parameter prediction via hypernetworks afford efficient multimodal feature alignment, via modality connectors

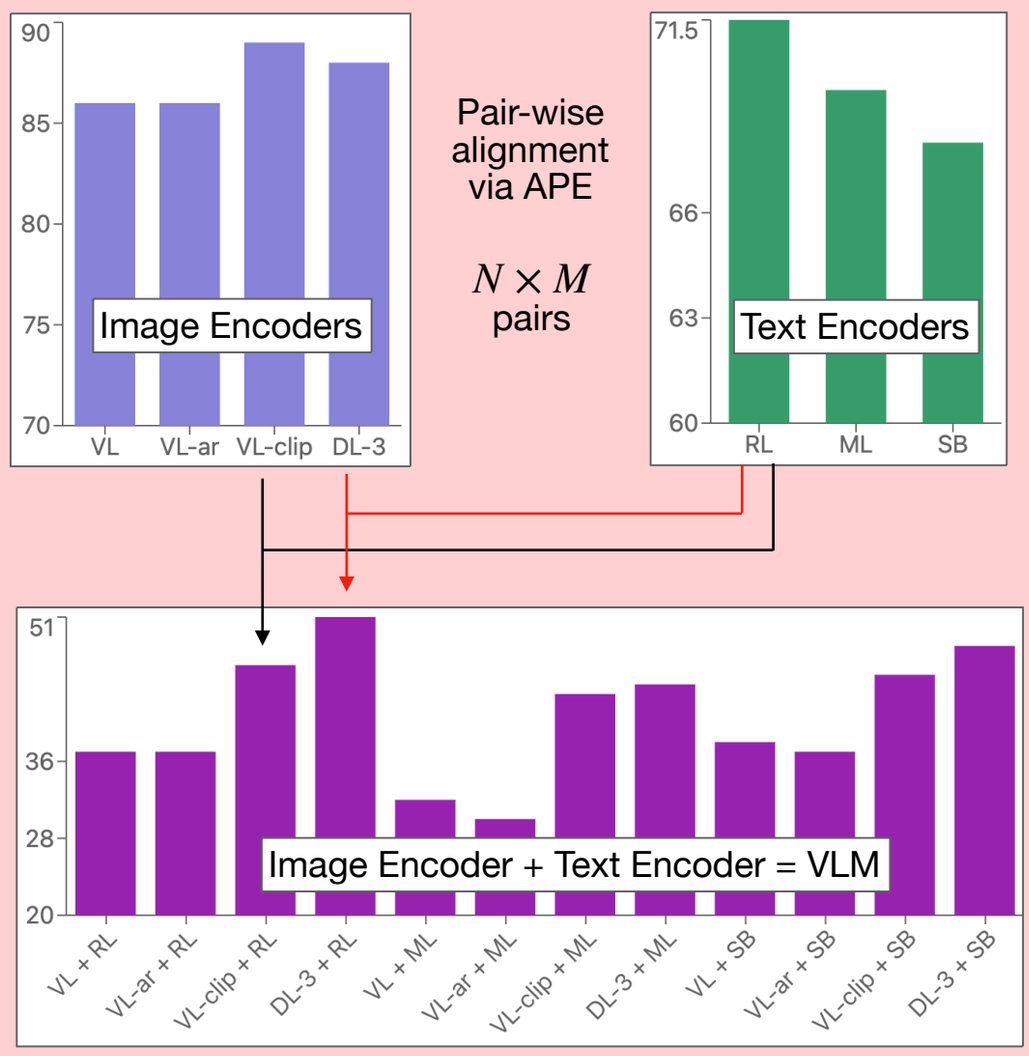
Possibilities for Future Work

Hyper-Align between image encoders and LLMs for efficient feature to embedding alignment in MLLMs



Research Problem

Unimodal Performance ≠ Multimodal Performance

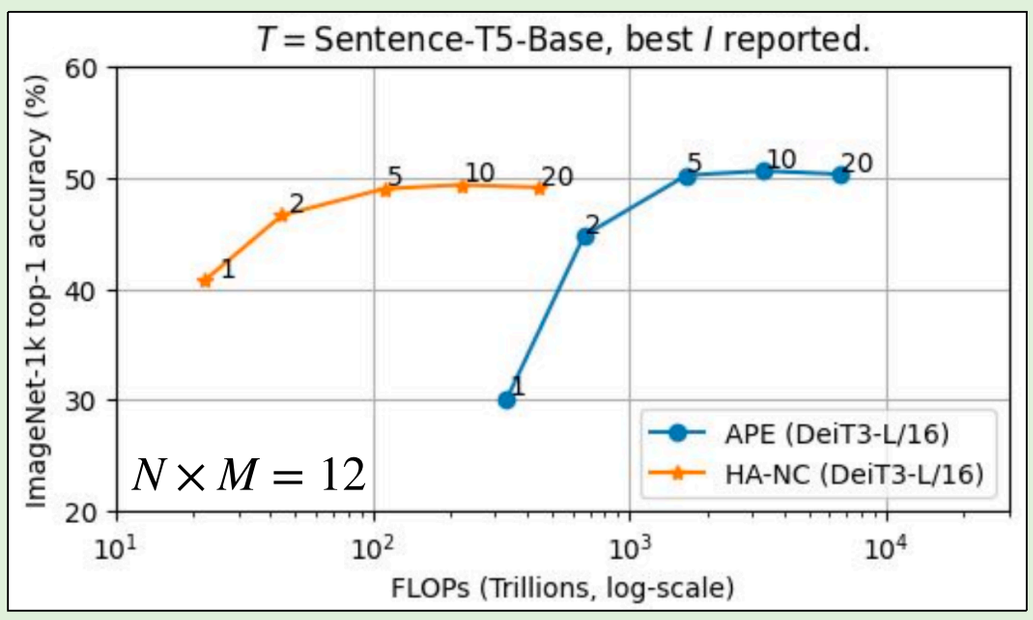


Need to search $N \times M$ connectors to find optimal VLM in N image & M text encoders.

Computationally expensive even with linear layer modality connectors + APE

Main Results

- I : Image encoder (best reported)
- T : Text encoder (fixed)
- $12 \leq N \leq 30$
- $M = 1$
- Modality connector : linear layer



Scale encoder count & diversity