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

SUMMARY

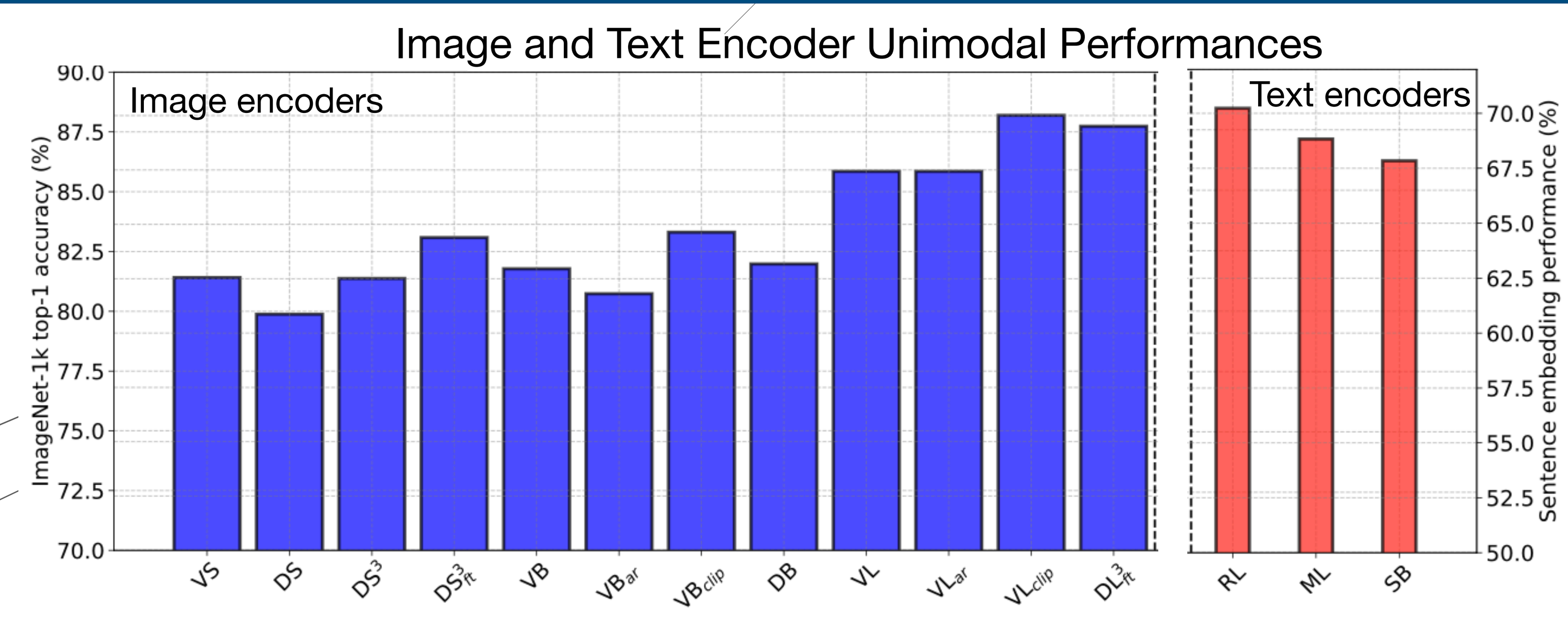
Contrastive vision-language models (VLMs) like CLIP: align encoders of image-text modalities via an InfoNCE loss.

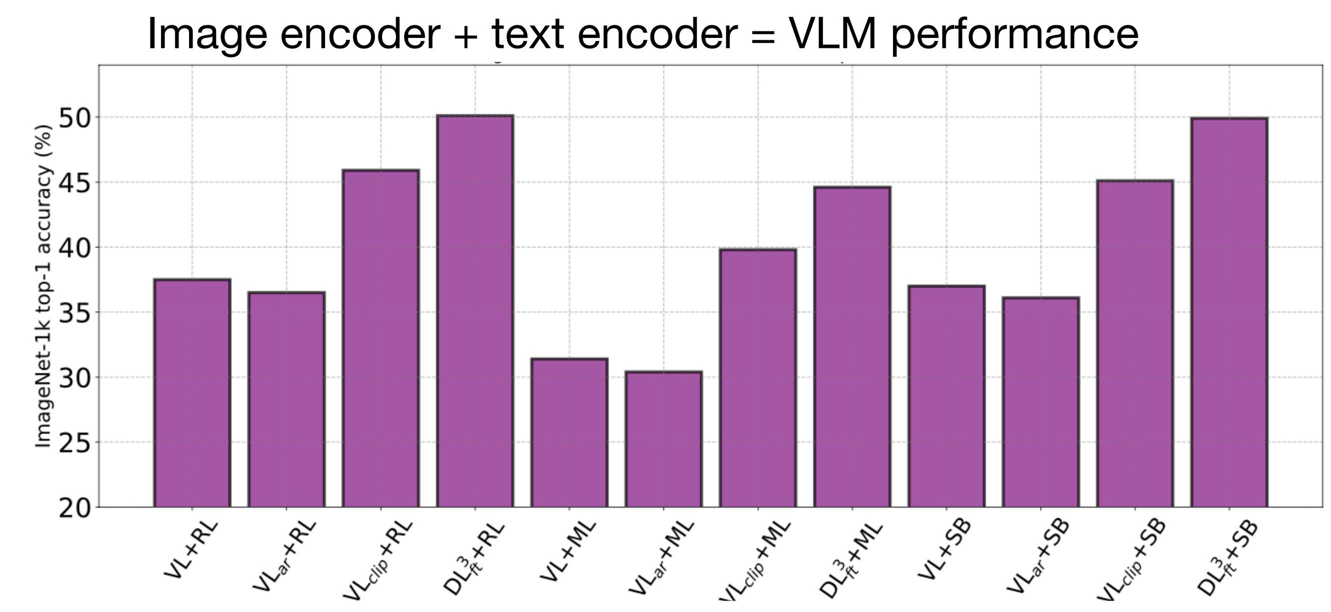
Background: Instead of training VLMs end-to-end,
- APE trains a modality connector (MLP) between pretrained encoders
- outperforms CLIP at significantly lower costs.

Problem: unimodal performance ≠ multimodal performance.
- Finding the optimal pair in N image and M text encoders requires searching all $N \times M$ combinations.
- Training any one combination needs massive data volumes.
- Hence, training all combinations individually becomes unfeasible.

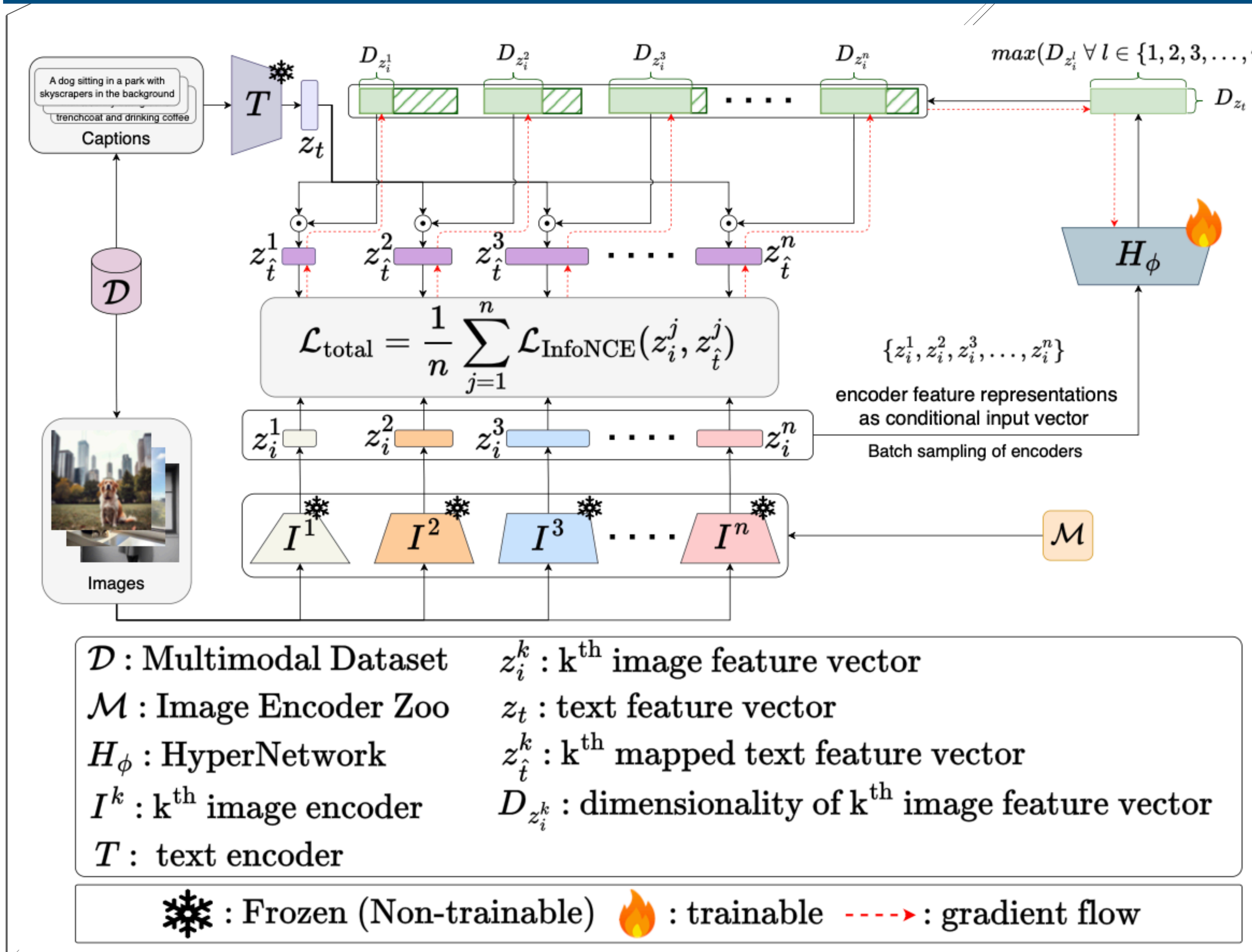
Proposed solution (Hyper-Align): Use a hypernetwork to learn all $N \times M$ modality connectors together, instead of learning them individually (APE).

Result: Compared to APE (on a linear modality connectors), Hyper-Align
- Is comparable in performance
- Yields an 8x reduction in FLOP cost.





METHODOLOGY

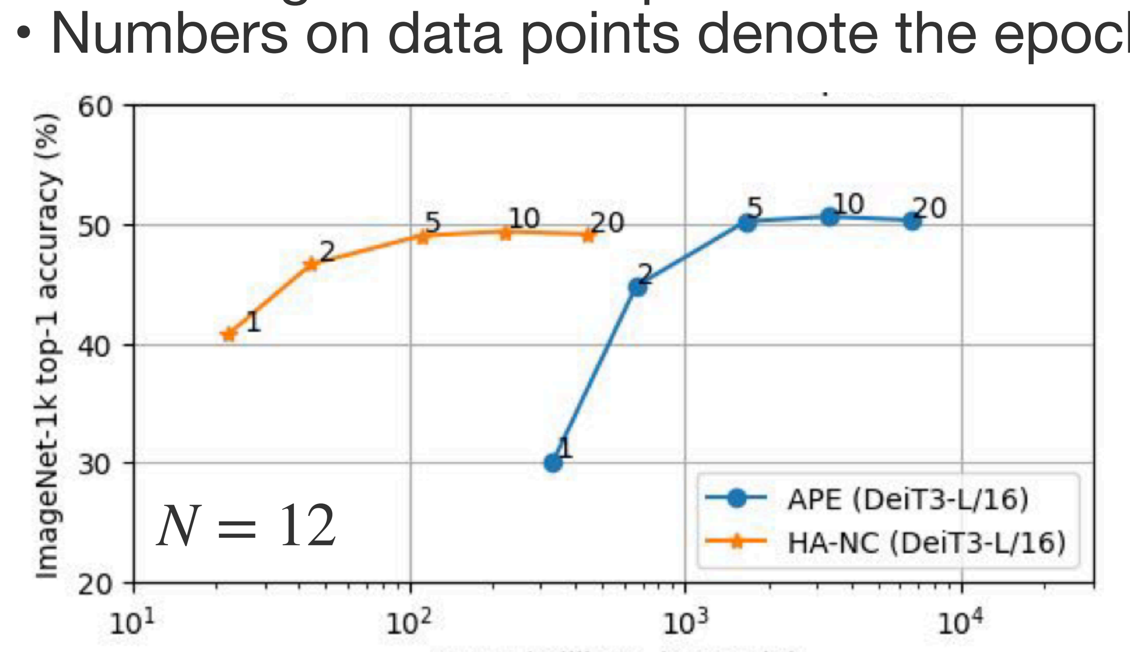


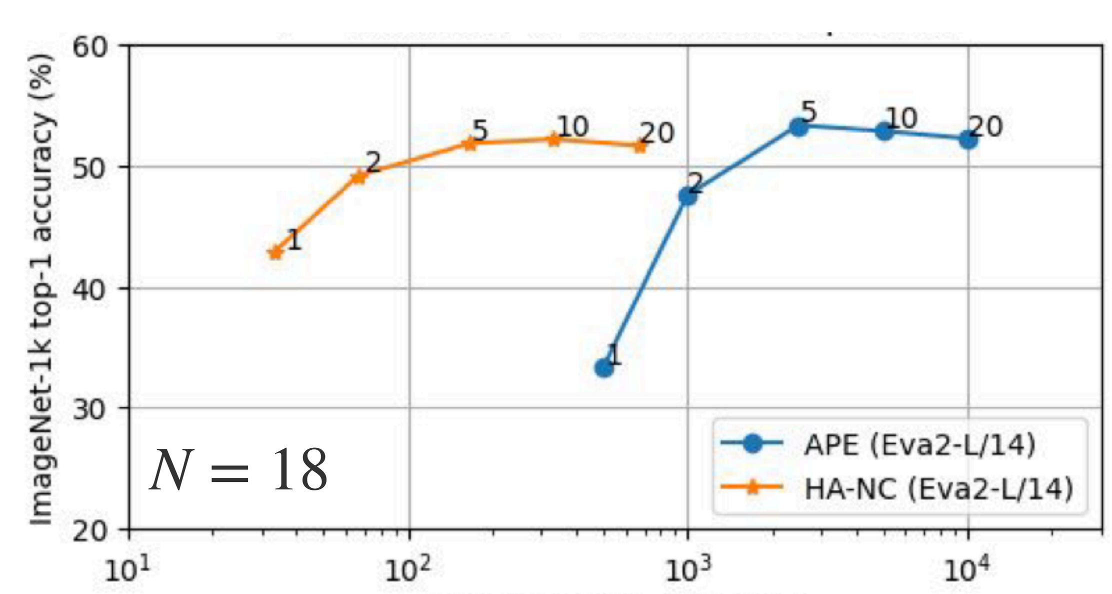
APE: train a linear layer $f_\theta : \mathbb{R}^{D_{z_t}} \rightarrow \mathbb{R}^{D_{z_i}}$ between encoders.
- z_i and z_t are embeddings of an image-caption sample
- training objective is $L_{APE} = L_{\text{InfoNCE}}(f_\theta(z_t), z_i)$

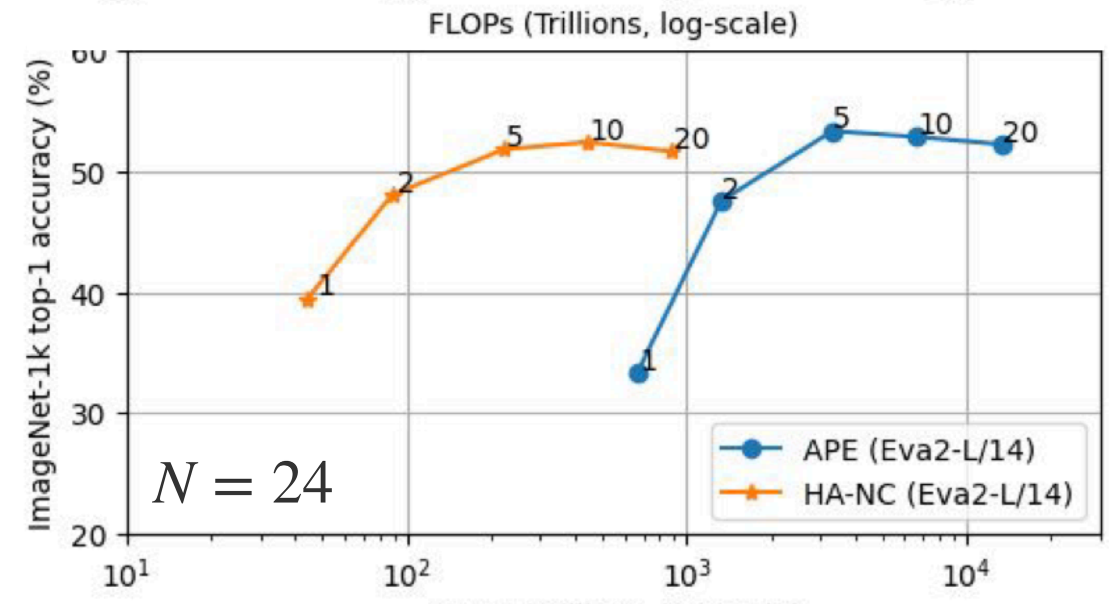
Hyper-Align: hypernetwork H_ϕ uses a conditional input c_j to predict the parameters of the j^{th} modality connector.
- z_i^j = batch of embeddings from j^{th} image encoder, $j \in \{1, \dots, N\}$
- z_t = batch of embeddings from 1 text encoder
- $H_\phi(c_j) = \theta_j$ where c_j = batch-average(z_i^j)
- training objective is $L_H = \Sigma_{j=1}^{N} \left(L_{\text{InfoNCE}}(f_{H_\phi(c_j)}(z_t), z_i^j)\right) / N$
- H_ϕ = MLP which predicts several parameter spaces (via slicing).
- We train H_ϕ on a mini-batch $B < N$ of image encoders per step to efficiently scale up the number of combinations (N).

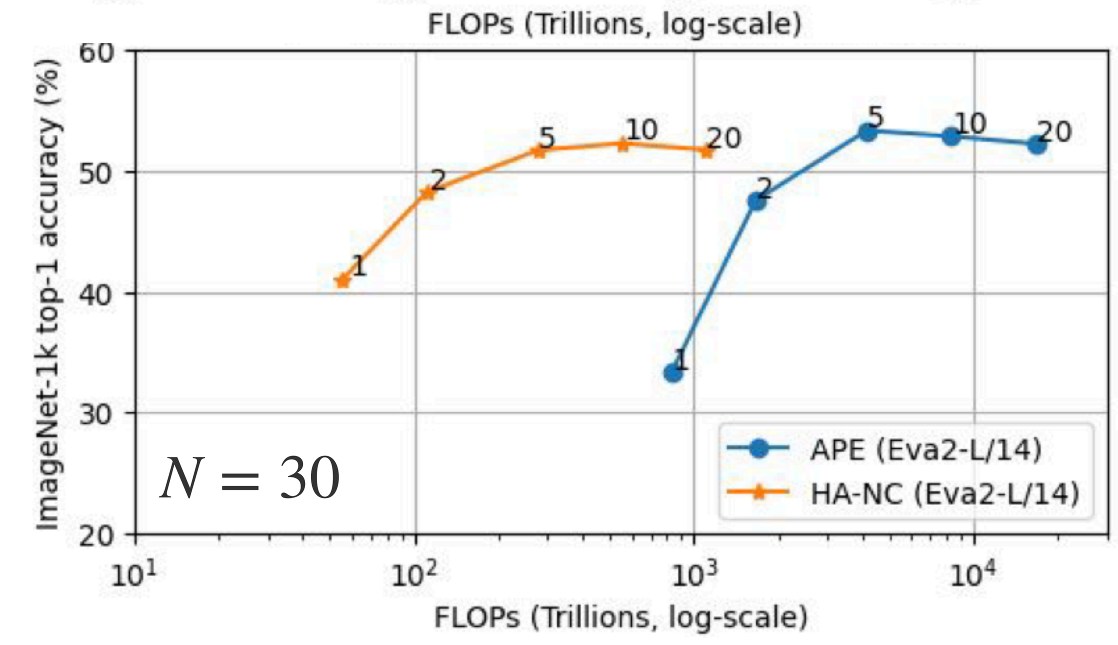
EXPERIMENTS AND RESULTS

Scaling up no. of combinations:
- $M = 1$ (sentence-t5-base) and N varies from 12 to 30
- Best image encoder reported at each value of N
- Numbers on data points denote the epochs at which the VLM was evaluated.









Search over various image encoder scales:
- N = 30 equally split among 3 feature dims
- Parameter count ↑ as feature dim ↑
- Best ImageNet accuracy is shown per scale.

Scale type	Range	Method: Ours	Method: APE
Feature dim	384	36.75	38.36
	768	42.83	45.44
	1024	51.92	53.86
Param. count	< 30M	36.75	38.36
	30M $- 120$M	43.04	44.84
	> 120M	51.92	53.86

CONCLUSION

Parameter prediction via hypernetworks can
- **efficiently search image-text encoder pairs** for optimal VLMs, under constraints.
- Future work can use Hyper-Align on **image encoders and LLMs to create MLLMs.**

Corresponding author: `jaisidh.singh@student.uni-tuebingen.de`