PAUL G. ALLEN SCHOOL W

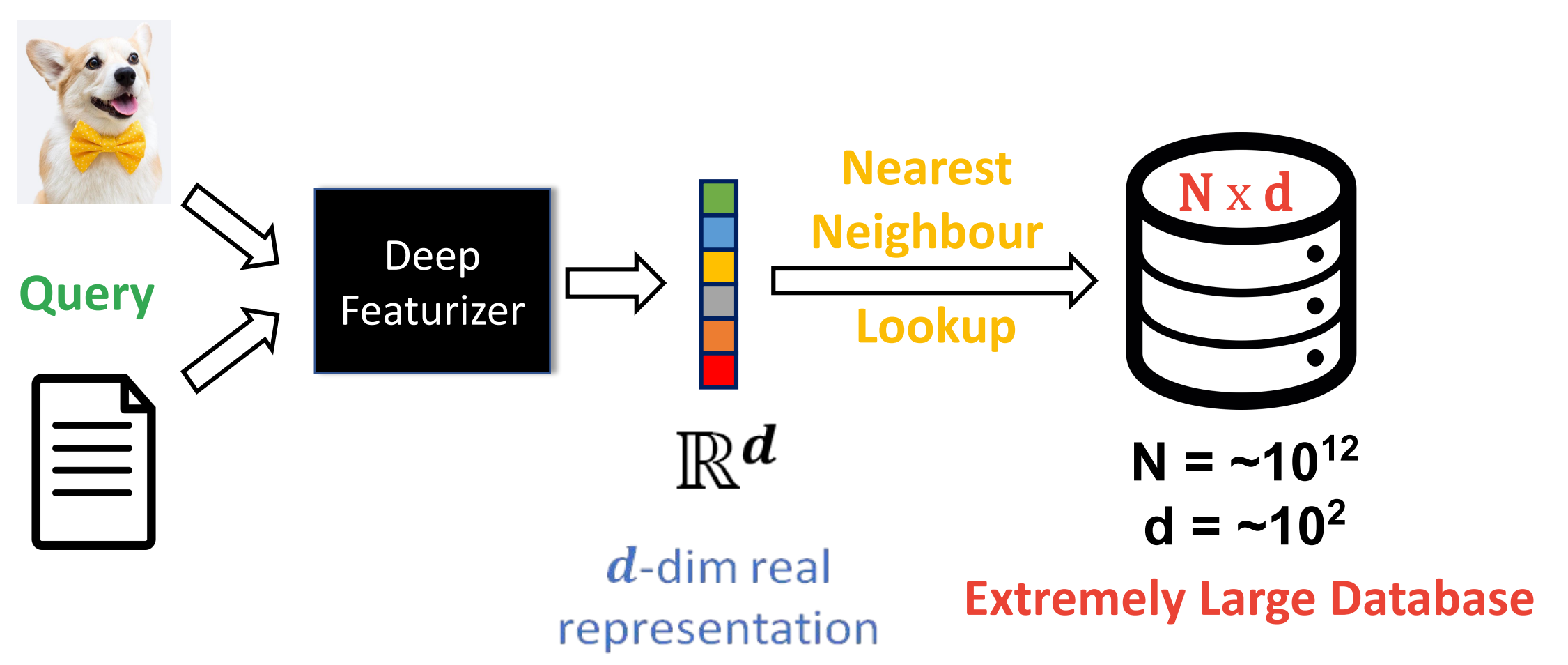
Google Research

Matryoshka Representation Learning

Aditya Kusupati*, Gantavya Bhatt*, Aniket Rege*,
Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen,
Sham Kakade, Prateek Jain, Ali Farhadi

/RAIVNLab/MRL
bit.ly/mrl-paper

Motivation: Query-based Retrieval



Applicable for large-scale classification with millions of labels

Web-scale Challenges

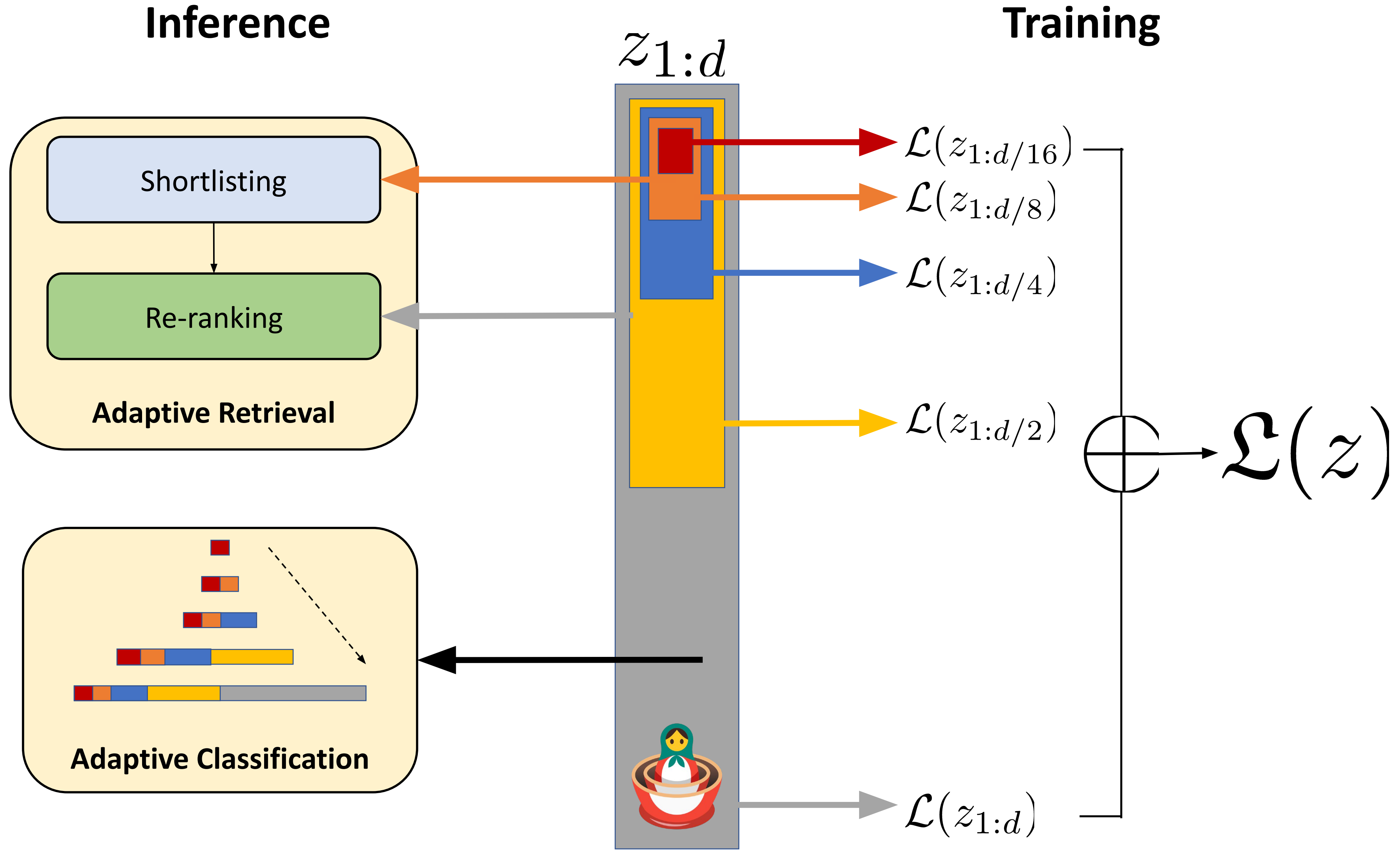
- Extremely large databases – 100s of TB
  - Linear dependence on representation size (*d*)
  - Embedding look-up much more expensive than featurization
- Require Approximate Nearest Neighbour Search (ANNS)
  - Post-hoc compressed index
- Incapable of Multi-Granularity
  - Use same *high-d* embedding for all tasks
  - Retrain a model for *low-d* based on deployment constraints
  - Eg: *2048-d* ResNet50 image representation for all tasks

Adaptive Deployment – Goals

- One representation vector for all downstream tasks
  - No post-hoc compression or expensive feature selection
  - No retraining for specific resource constraints
- Accurate and efficient *low-d* embeddings
  - Baked within the *high-d* embedding – Free
  - Reduced costs for expensive & high-recall shortlisting
  - As accurate as independently trained counterparts
- *High-d* embedding for cheap & precise re-ranking

Matryoshka Representation Learning - MRL

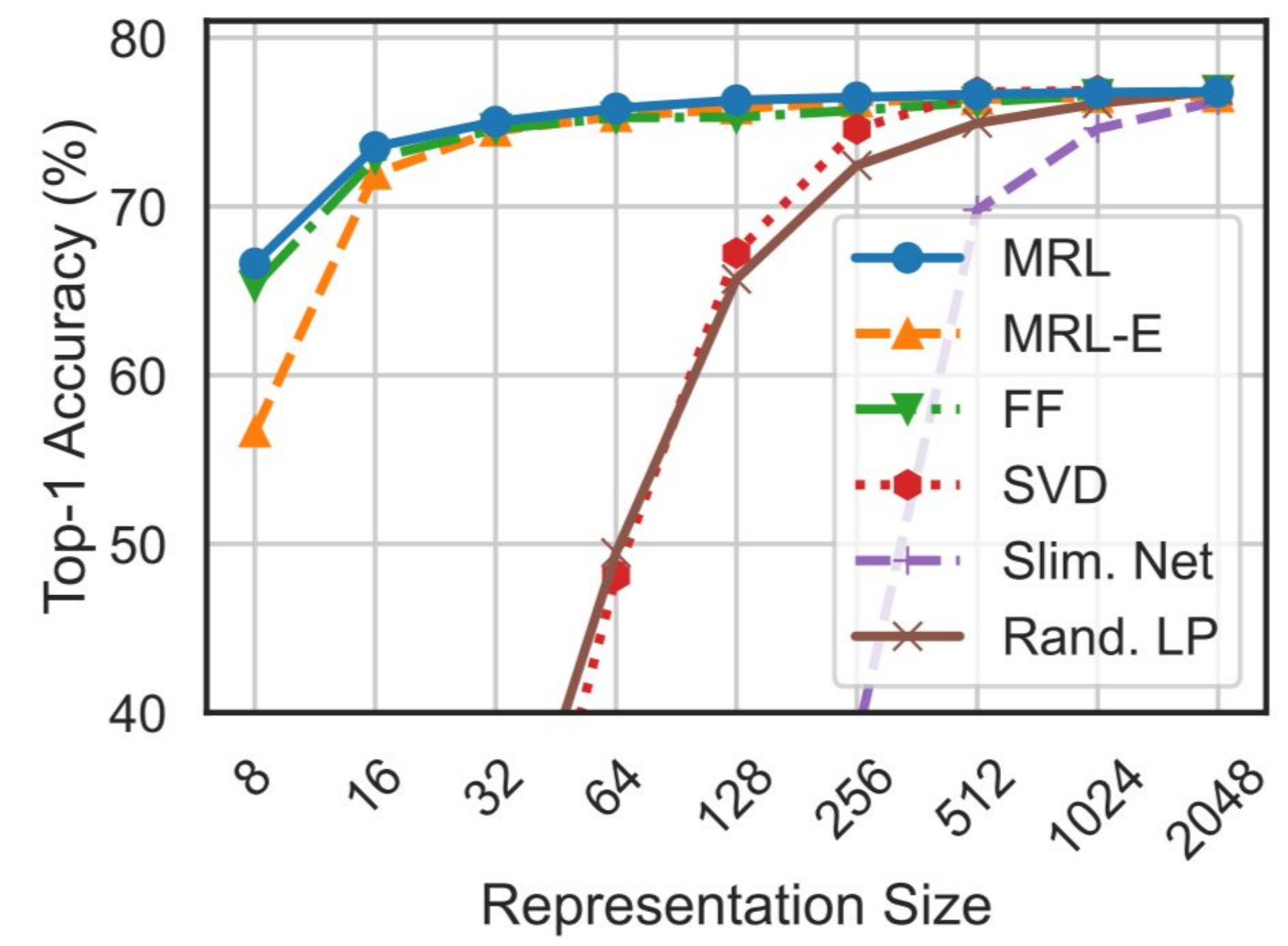
- Solve the same learning task at **multiple granularities** $(log(d))$
- Easily adaptable to any representation learning setup
  - Scale, modality and task agnostic – **1B images with ease**
- **First *k* dims** form the required *low-d* embeddings
  - As accurate as retrained *low-d* counterparts
- Enable **adaptive** deployment
  - Accurate large-scale classification & retrieval based on constraints



Classification Accuracy

ImageNet OVA

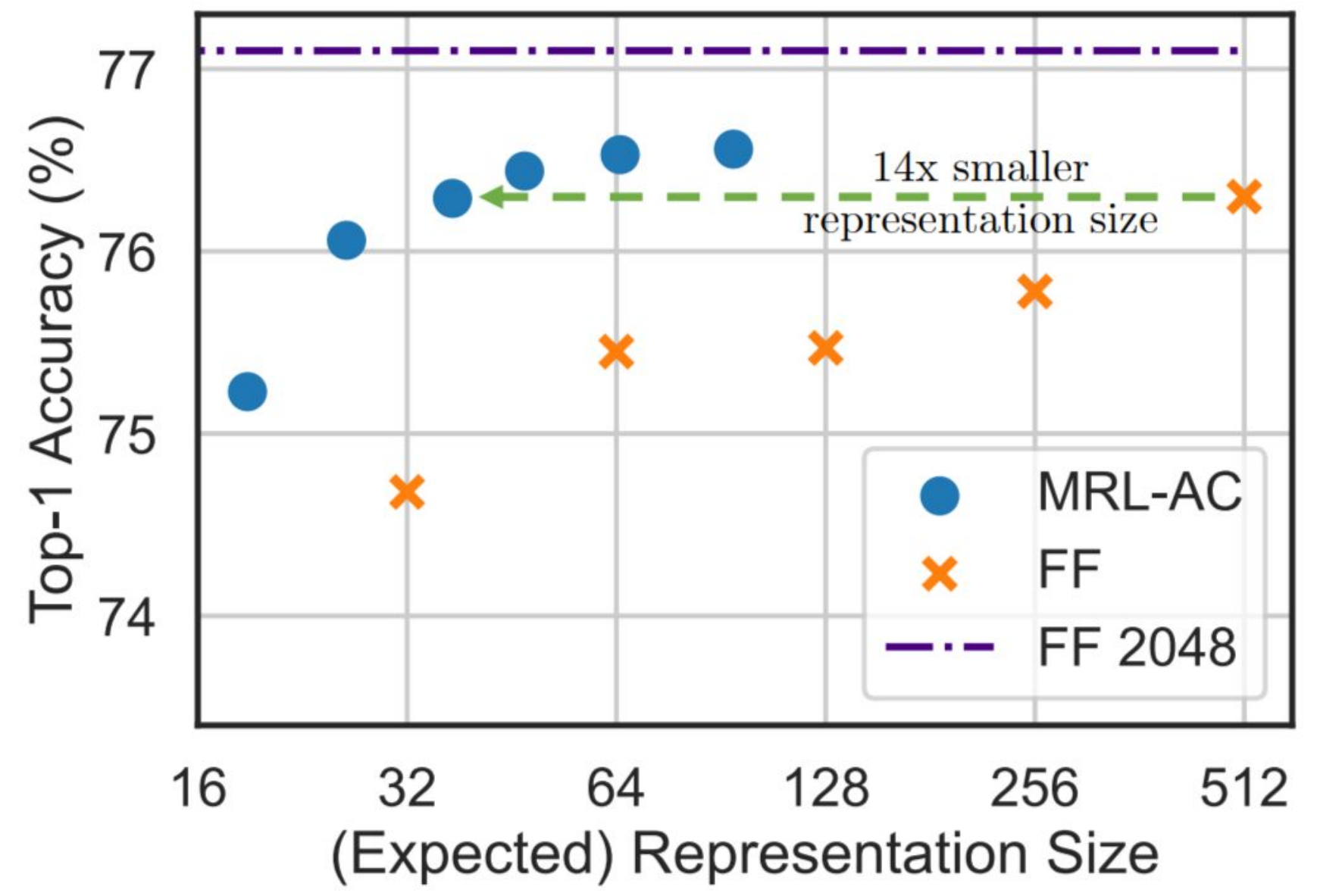
- ResNet50: Same accuracy as independently trained *low-d* models (FF)



Adaptive Classification

ImageNet-1K

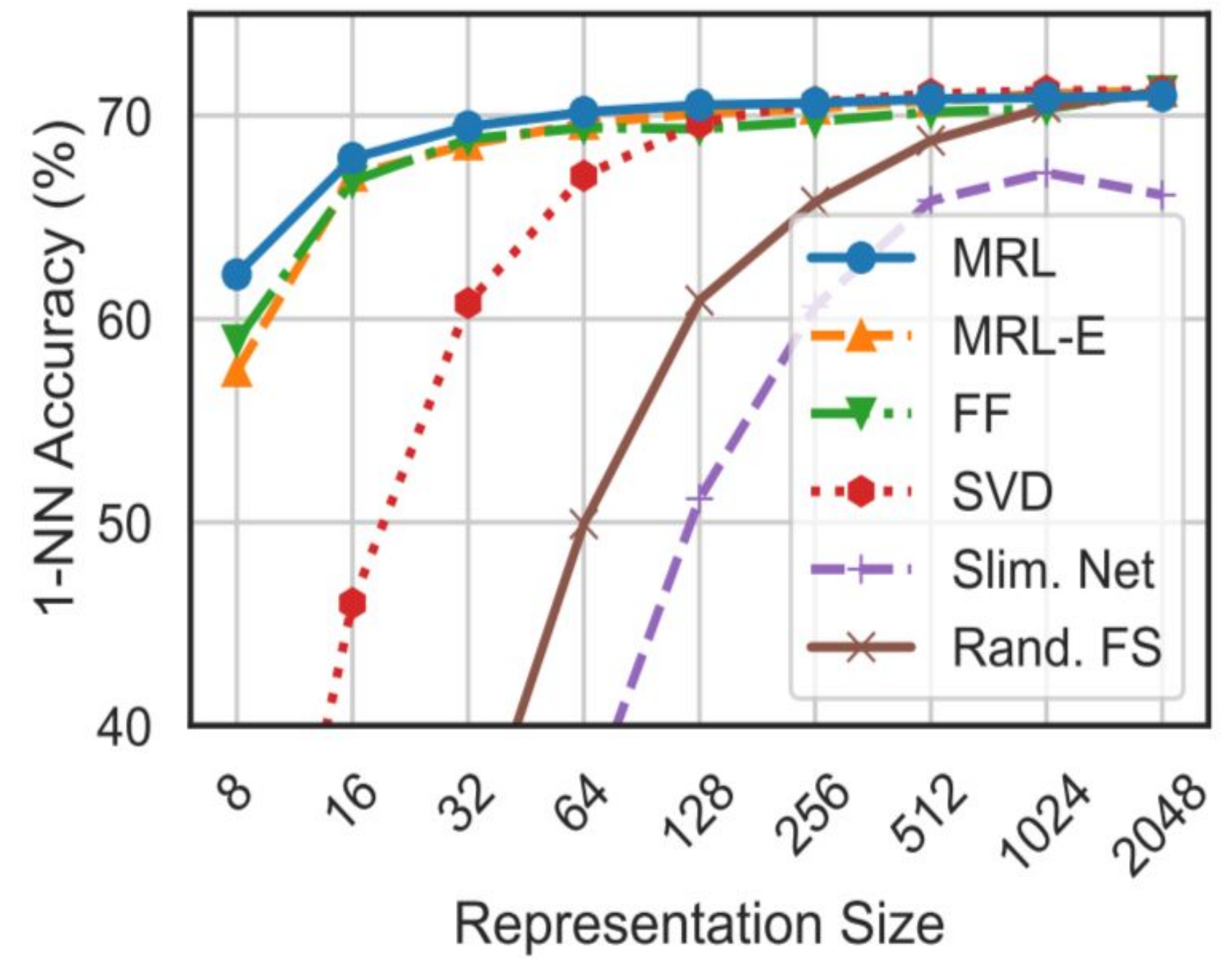
- ResNet50-MRL model with cascades
- 14x smaller embedding size for same accuracy



Representation Quality

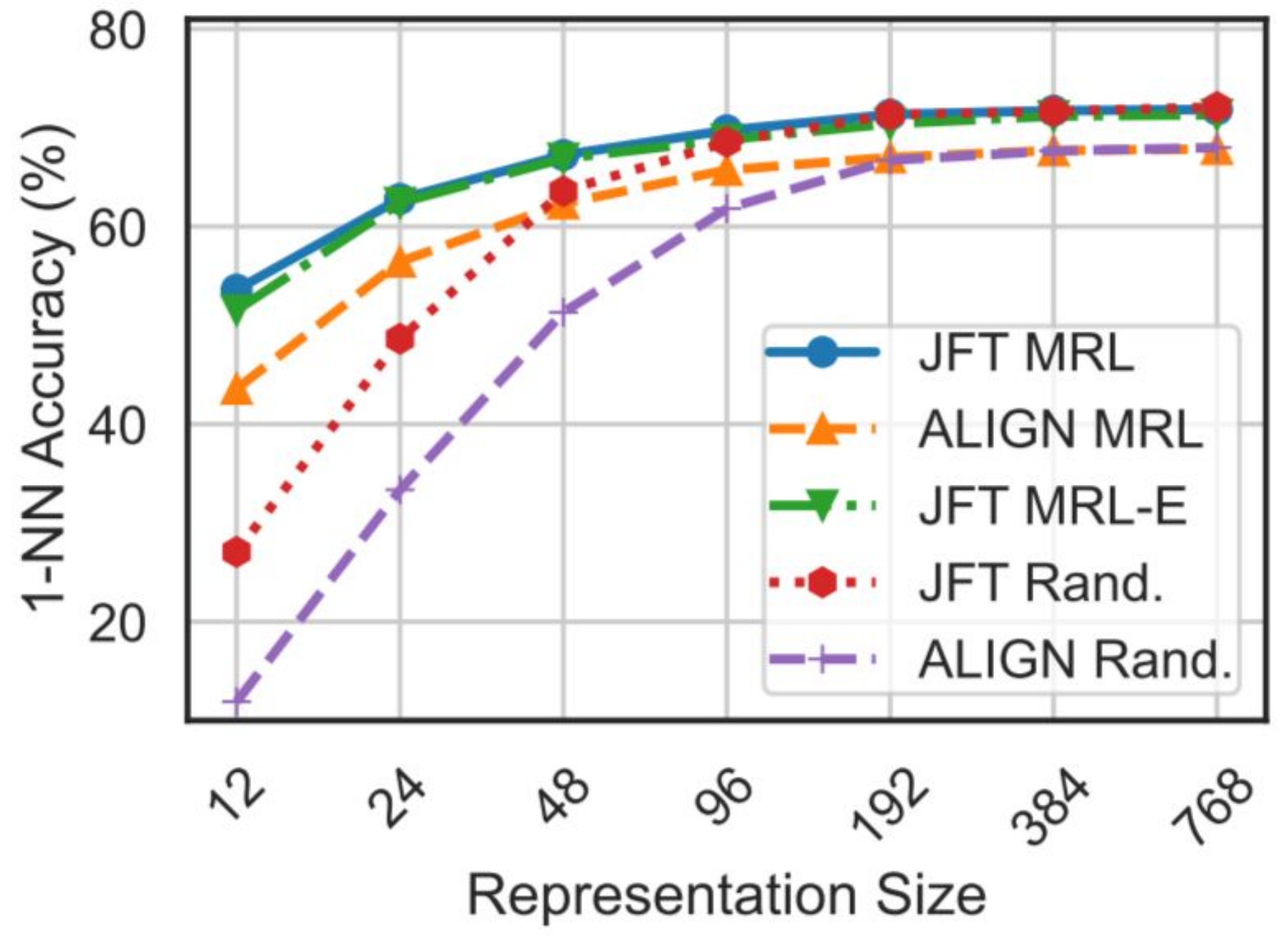
ImageNet k-NN

- ResNet50 models trained on ImageNet-1K
- Other baselines fall off drastically at *low-dimensions*

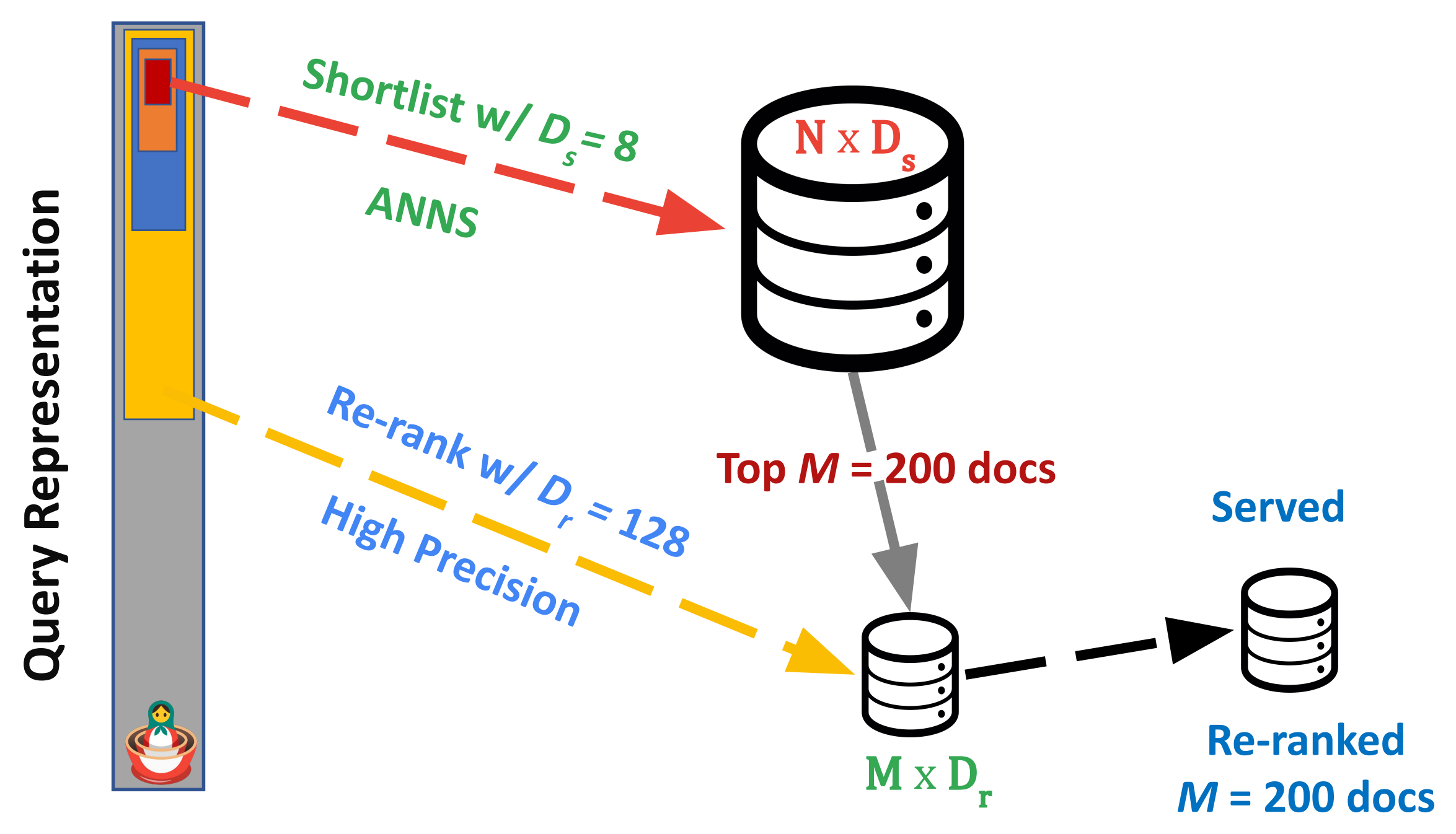


ViT+JFT & ALIGN k-NN

- ViT-B/16 models trained on JFT-300M and ALIGN (V+L)
- Scales to **1B images** w/o accuracy drop

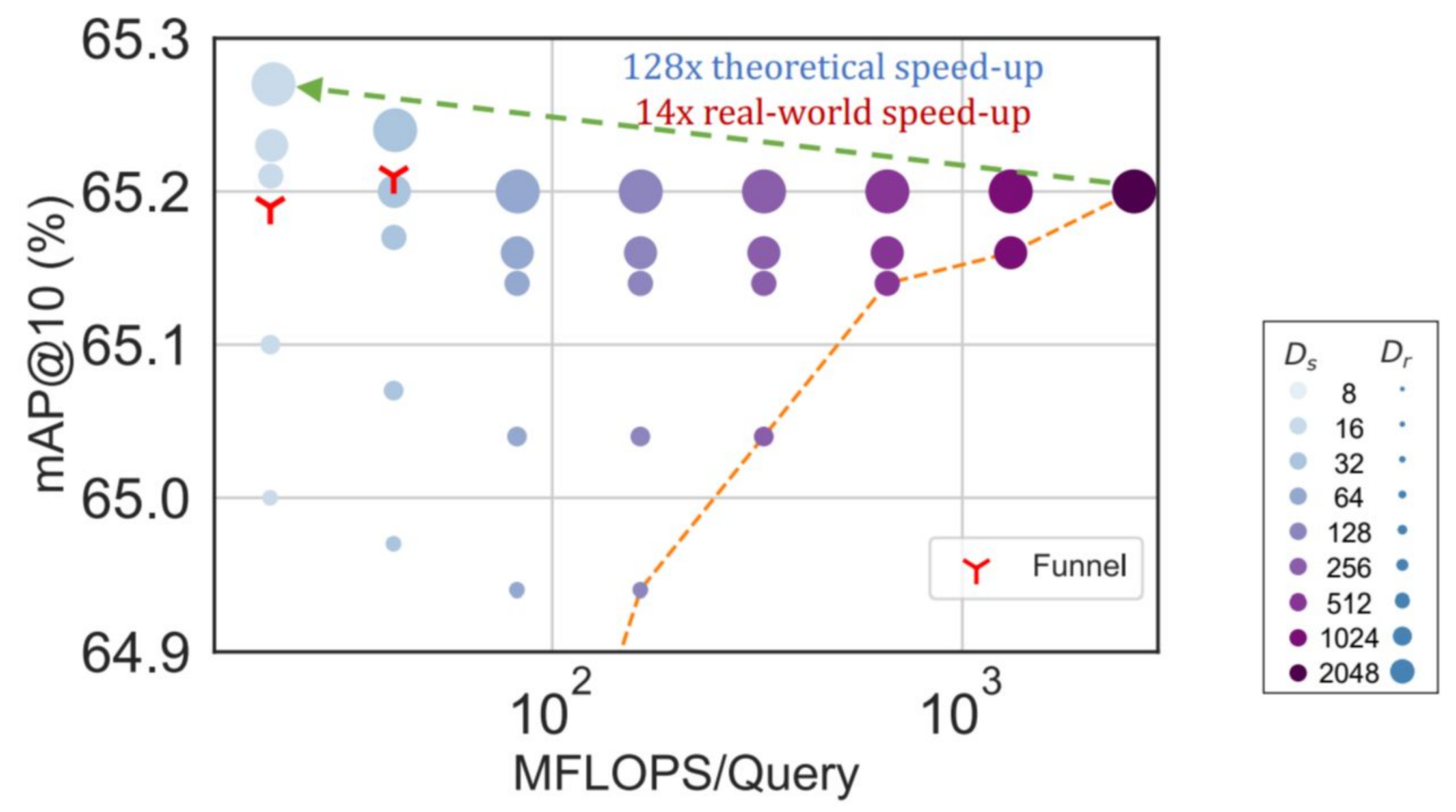


Adaptive Retrieval



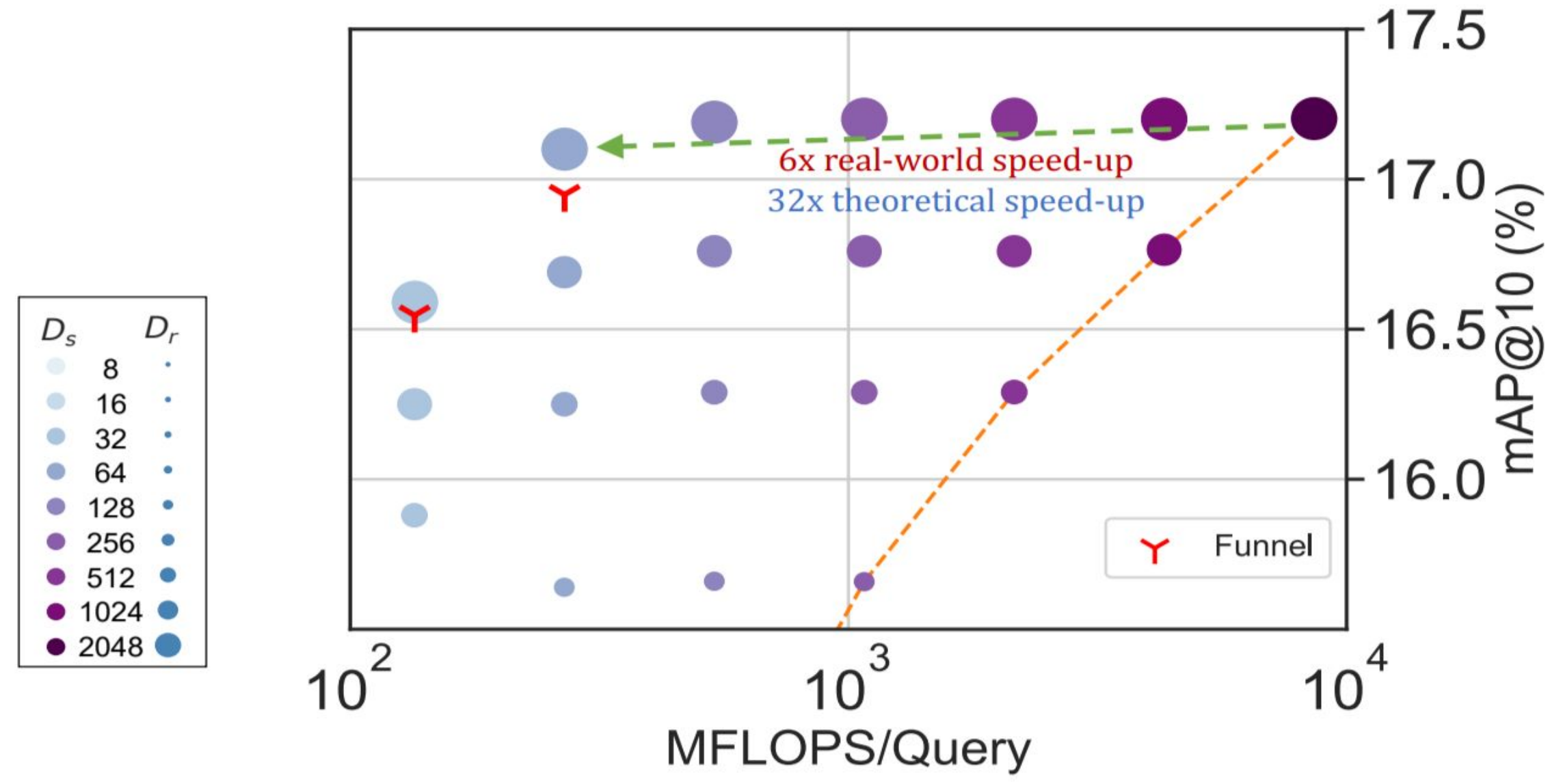
ImageNet-1K

- 14x real-world speed-up for the best mAP@10
- All real-world implementations use HNSW for shortlisting



ImageNet-4K (Try it!)

- 6x real-world speed-up for the best mAP@10
- Funnel retrieval alleviates the need for optimal D_s & D_r


mAP@10 (%)
10^2
10^3
10^4
MFLOPS/Query
6x real-world speed-up
32x theoretical speed-up
Funnel
D_s
D_r