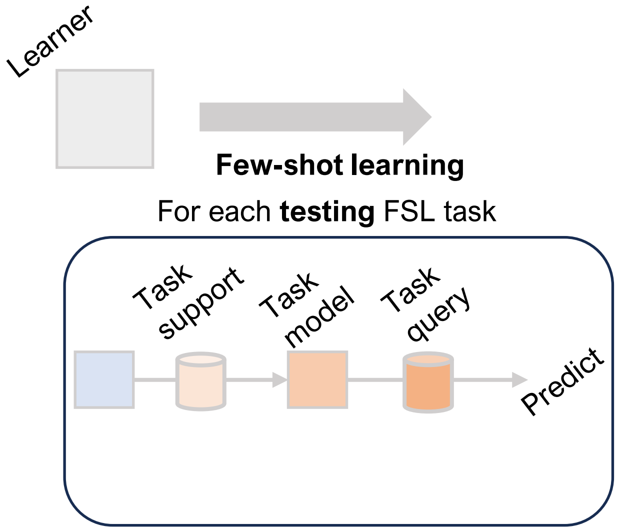
Sparse Meta-Tuning

Problem of interest

Few-shot classification (FSL): learning a model to perform classification from a few labelled examples

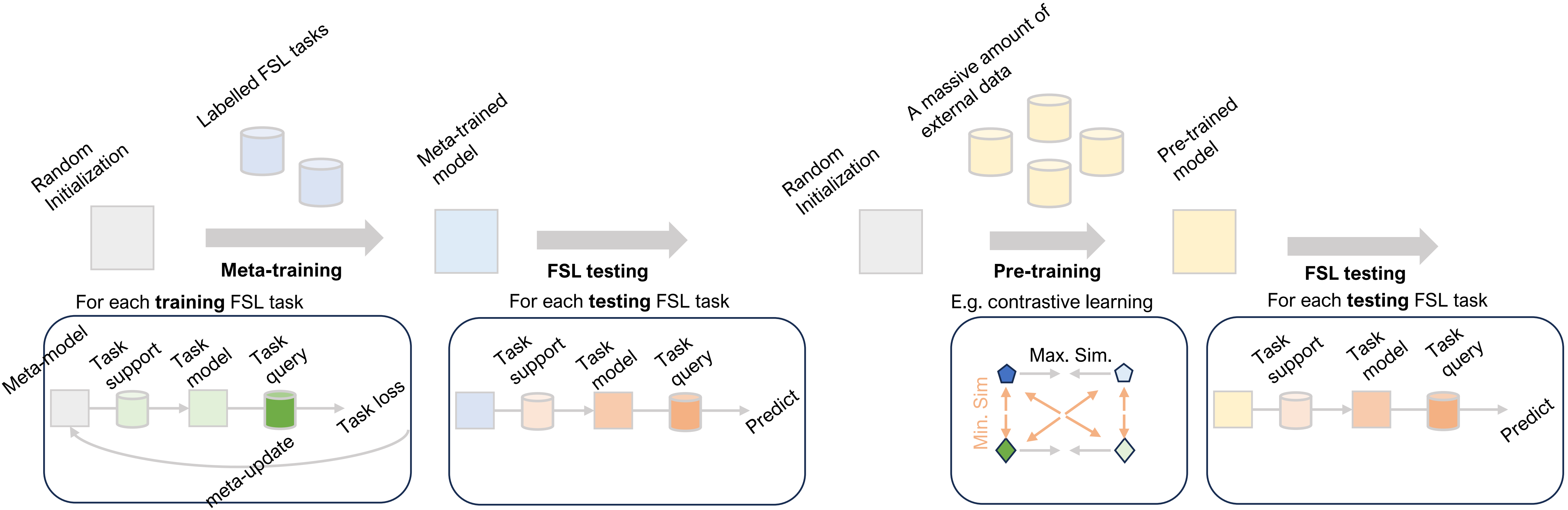


Problem of interest

Few-shot classification (FSL): learning a model to perform classification from a few labelled examples

Two typical approaches:

- **Meta-learning:** train from scratch over labelled few-shot task episodes by maximizing the FSL objective
- **Transfer-learning:** finetune a powerful pre-trained model directly on the few labelled example

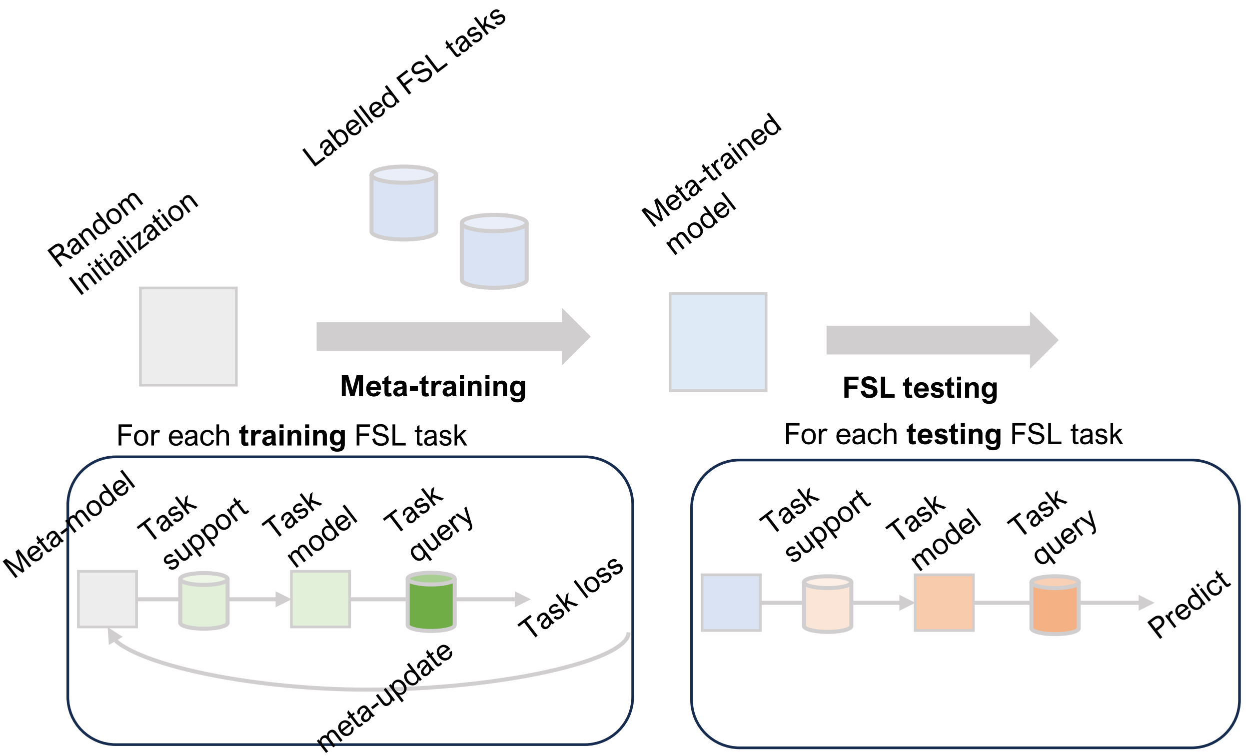


Problem of interest

Few-shot classification (FSL): learning a model to perform classification from a few labelled examples

Two typical approaches:

- **Meta-learning:** train from scratch over labelled few-shot task episodes by maximizing the FSL objective
- **Transfer-learning:** finetune a powerful pre-trained model directly on the few labelled example



Challenges:

- **Difficult optimization:** …second-order optimization…high complexity…from scratch…
- **Suboptimal performance**: typically get outperformed by transfer-learning approaches… especially nowadays in the era of big data and foundation models

Problem of interest

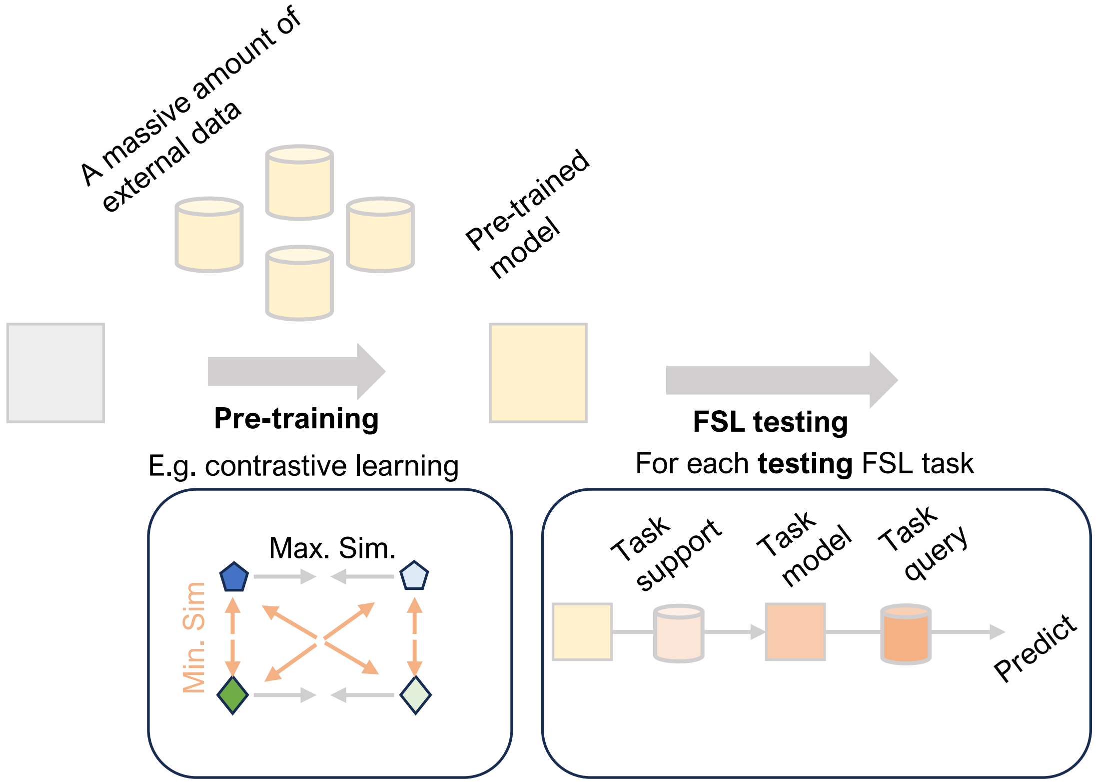
Few-shot classification (FSL): learning a model to perform classification from a few labelled examples

Two typical approaches:

- **Meta-learning:** train from scratch over labelled few-shot task episodes by maximizing the FSL objective
- **Transfer-learning:** finetune a powerful pre-trained model directly on the few labelled example

Challenges:

- **Misaligned objectives** between pre-training and downstream FSL …
- Therefore**, FSL performance can still be unsatisfactory** / suboptimal…



Problem of interest

Few-shot classification (FSL): learning a model to perform classification from a few labelled examples

Two typical approaches:

- **Meta-learning:** train from scratch over labelled few-shot task episodes by maximizing the FSL objective
- **Transfer-learning:** finetune a powerful pre-trained model directly on the few labelled example

Our interest: combine the best from both ends

Problem of interest

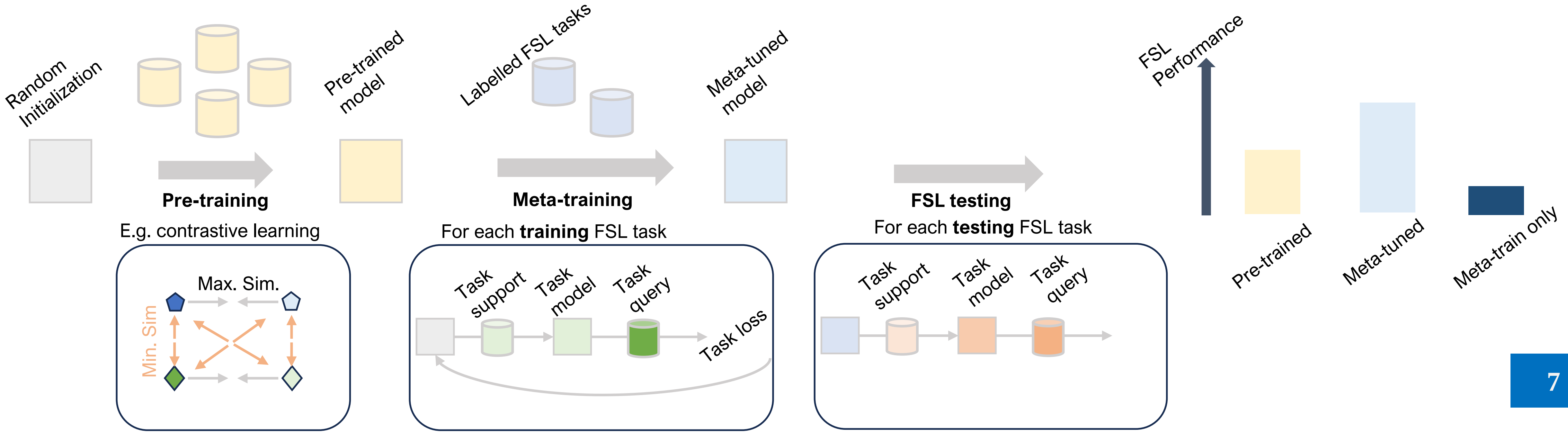
Few-shot classification (FSL): learning a model to perform classification from a few labelled examples

Two typical approaches:

- **Meta-learning:** train from scratch over labelled few-shot task episodes by maximizing the FSL objective
- **Transfer-learning:** finetune a powerful pre-trained model directly on the few labelled example

Our interest: combine the best from both ends

- **Meta-tuning:** meta-training starting from a pre-trained model



PMF: a previous approach

PMF[†] ≡ **P**re-train (DINO) → **M**eta-train (ProtoNet) → **F**ine-tune

- **Simple-yet-effective**: the state-of-the-art approach on the Meta-dataset benchmark

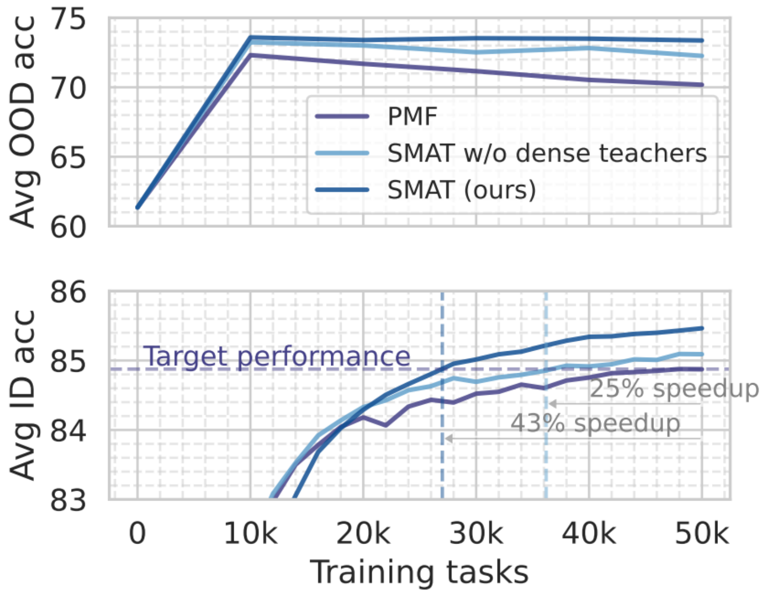
[†] Hu et. al, Pushing the Limits of Simple Pipelines for Few-Shot Learning, CVPR 2022

PMF: a previous approach

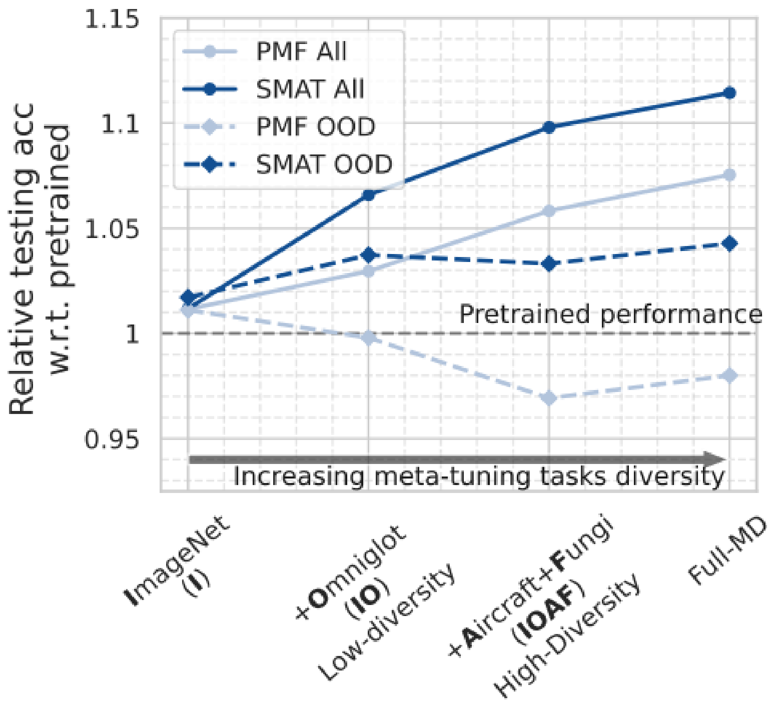
PMF† ≡ **P**re-train (DINO) → **M**eta-train (ProtoNet) → **F**ine-tune

- **Simple-yet-effective**: the state-of-the-art approach on the Meta-dataset benchmark
- However, it still suffers from **two major drawbacks**

Meta-overfitting



Task interference



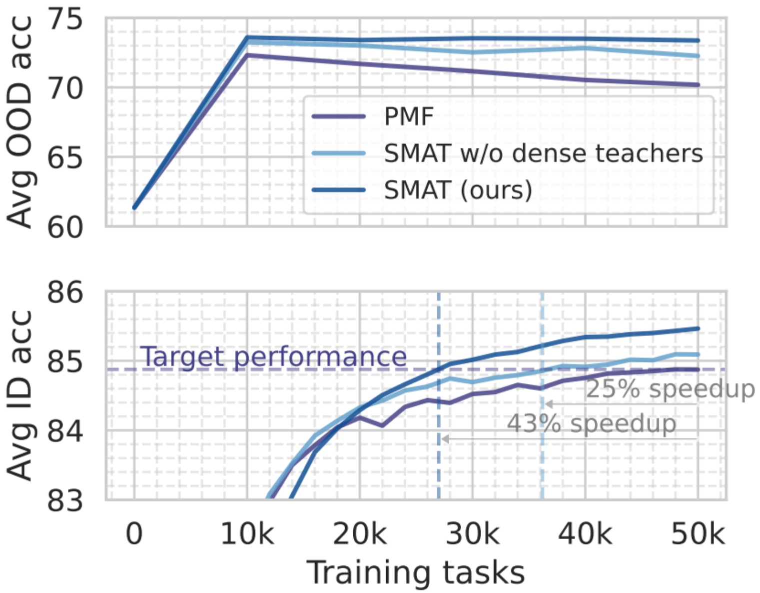
† Hu et. al, Pushing the Limits of Simple Pipelines for Few-Shot Learning, CVPR 2022

PMF: a previous approach

PMF[†] ≡ **P**re-train (DINO) → **M**eta-train (ProtoNet) → **F**ine-tune

- **Simple-yet-effective**: the state-of-the-art approach on the Meta-dataset benchmark
- However, it still suffers from **two major drawbacks**

Meta-overfitting



Improved ID generalization performance

But, at a **significant cost of OOD** performance

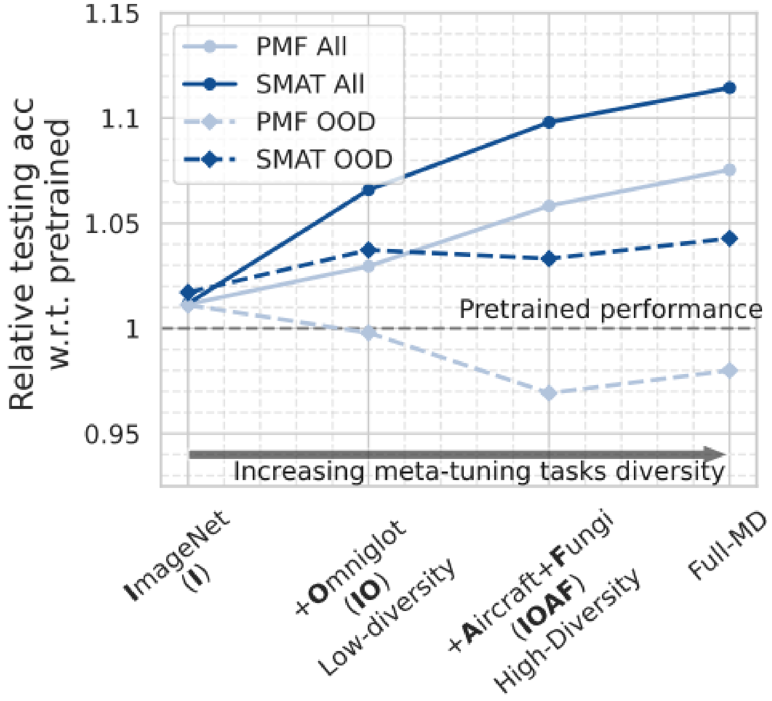
[†] Hu et. al, Pushing the Limits of Simple Pipelines for Few-Shot Learning, CVPR 2022

PMF: a previous approach

PMF† ≡ **P**re-train (DINO) → **M**eta-train (ProtoNet) → **F**ine-tune

- **Simple-yet-effective**: the state-of-the-art approach on the Meta-dataset benchmark
- However, it still suffers from **two major drawbacks**

Task interference



⬆ **Improved overall** generalization performance

⬇ But, the improvement **scales less-well** with more meta-training datasets

† Hu et. al, Pushing the Limits of Simple Pipelines for Few-Shot Learning, CVPR 2022

How can we improve it?

Our goal: a meta-tuning method that pushes toward the ID/OOD performance Pareto front

Sparse **M**eTa-**T**uning (SMAT) **key ideas**:

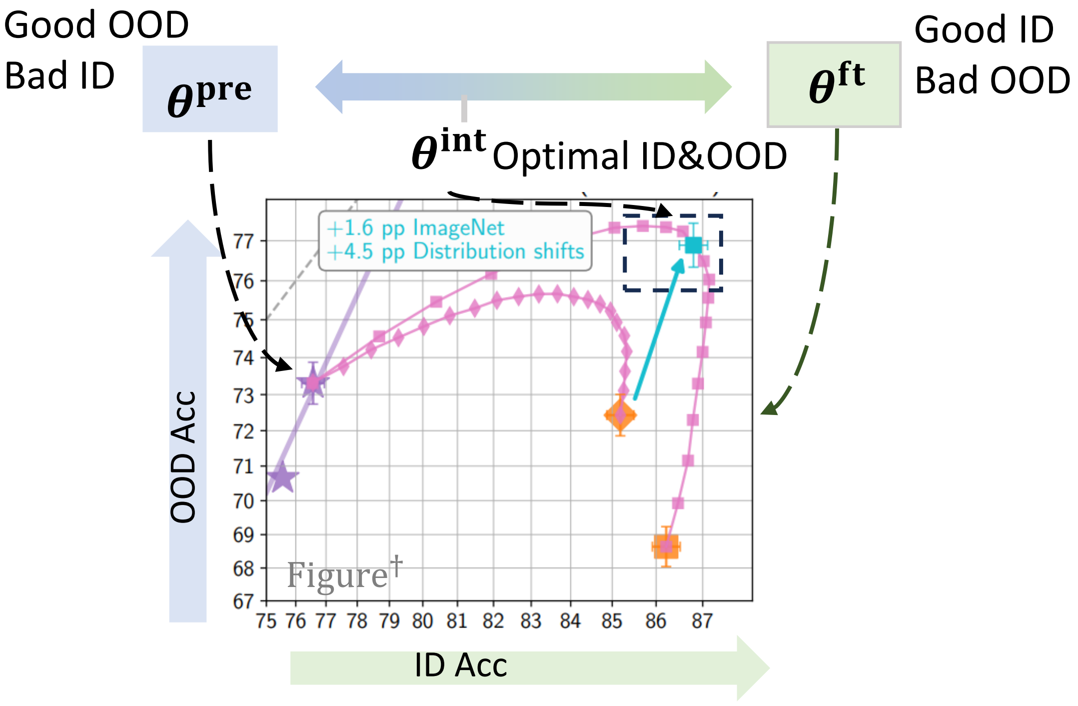
1. **Meta-learn sparsely interpolated experts**:
   - Sparse weight interpolation finds optimal ID and OOD performance
2. **Mixture-of-expert (MoE)-inspired model:**
   - A balanced point between fully task-agnostic and task-specific

† Hu et. al, Pushing the Limits of Simple Pipelines for Few-Shot Learning, CVPR 2022

How can we improve it?

Sparse **M**eTa-**T**uning (SMAT) **key ideas**:

1. **Meta-learn sparsely interpolated experts**:
   - Sparse weight interpolation finds optimal ID and OOD performance
2. Mixture-of-expert (MoE)-inspired model:
   - A balanced point between fully task-agnostic and task-specific

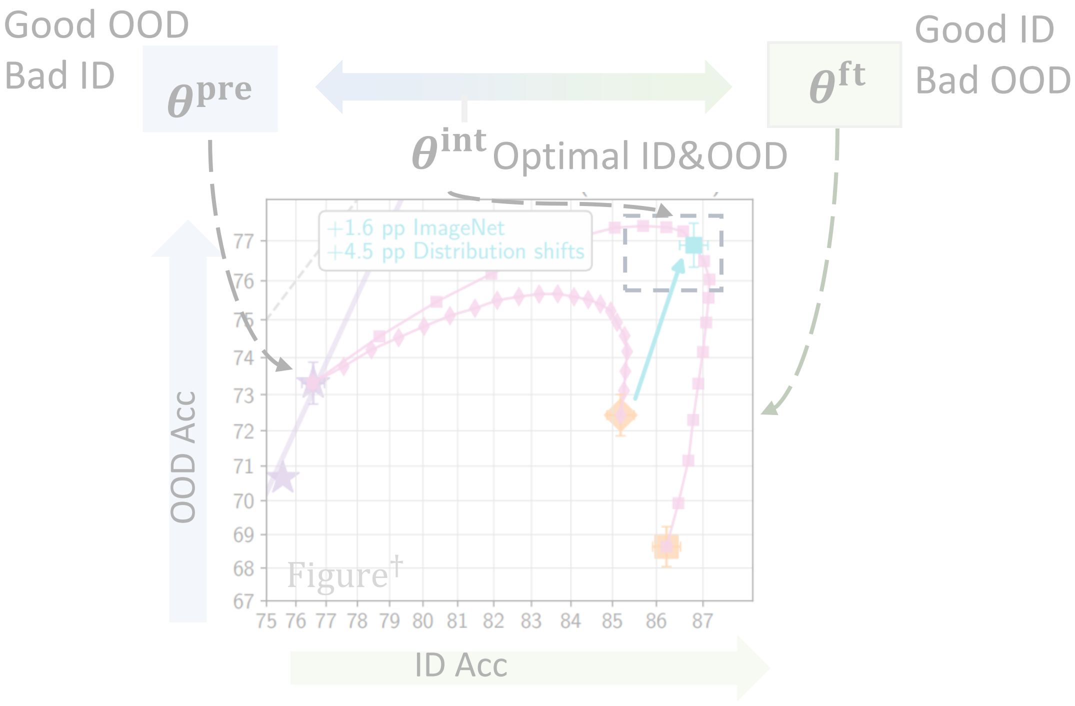


† Wortsman et. al, Robust fine-tuning of zero-shot models, CVPR 2022

How can we improve it?

Sparse **M**e**T**a-**T**uning (SMAT) **key ideas**:

1. **Meta-learn sparsely interpolated experts**:
   - Sparse weight interpolation finds optimal ID and OOD performance
2. **Mixture-of-expert (MoE)-inspired model:**
   - A balanced point between fully task-agnostic and task-specific



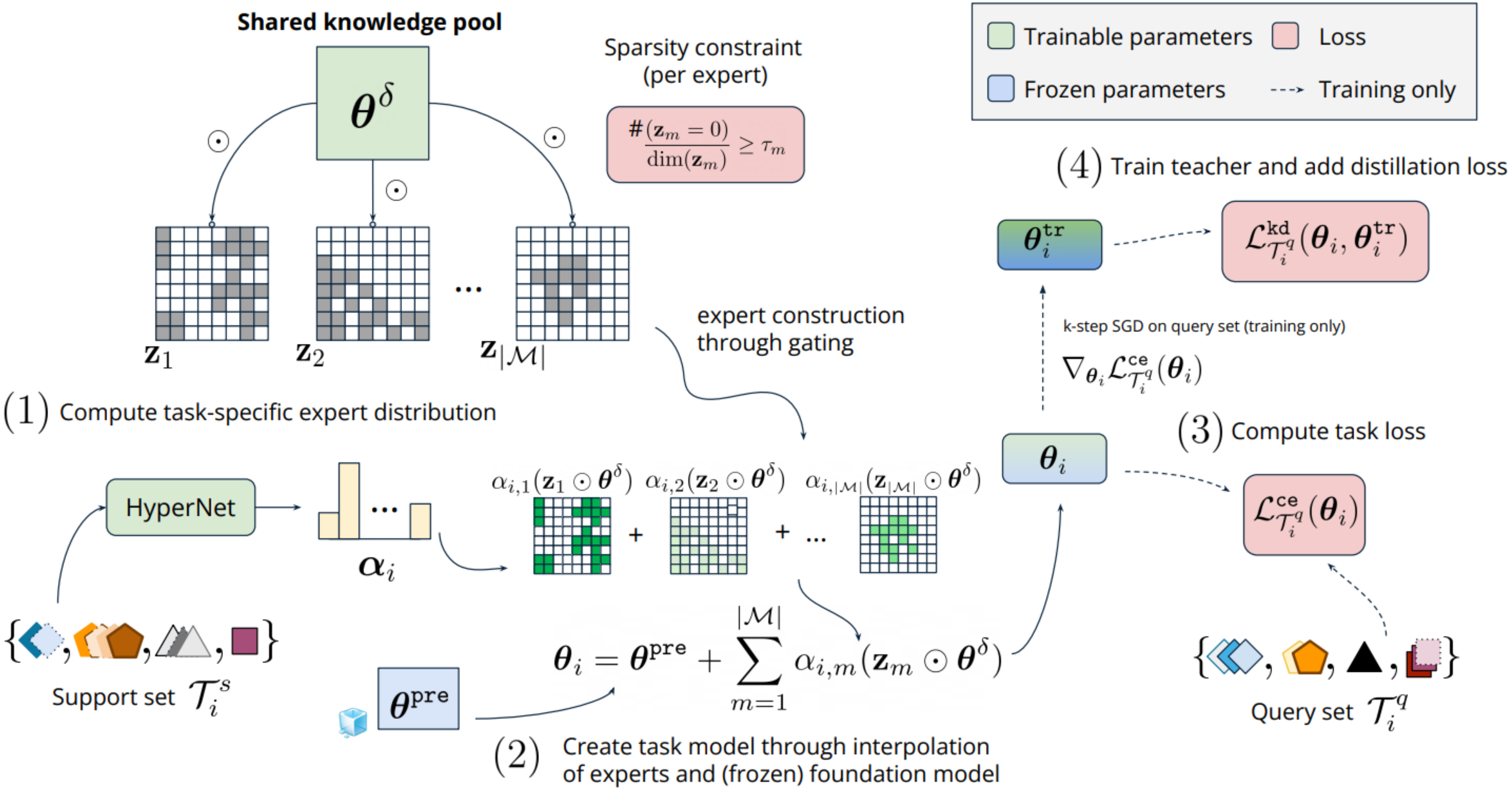
Fully **task-agnostic**
E.g., PMF

Partitioned:
task-specific & task-agnostic
E.g., SMAT

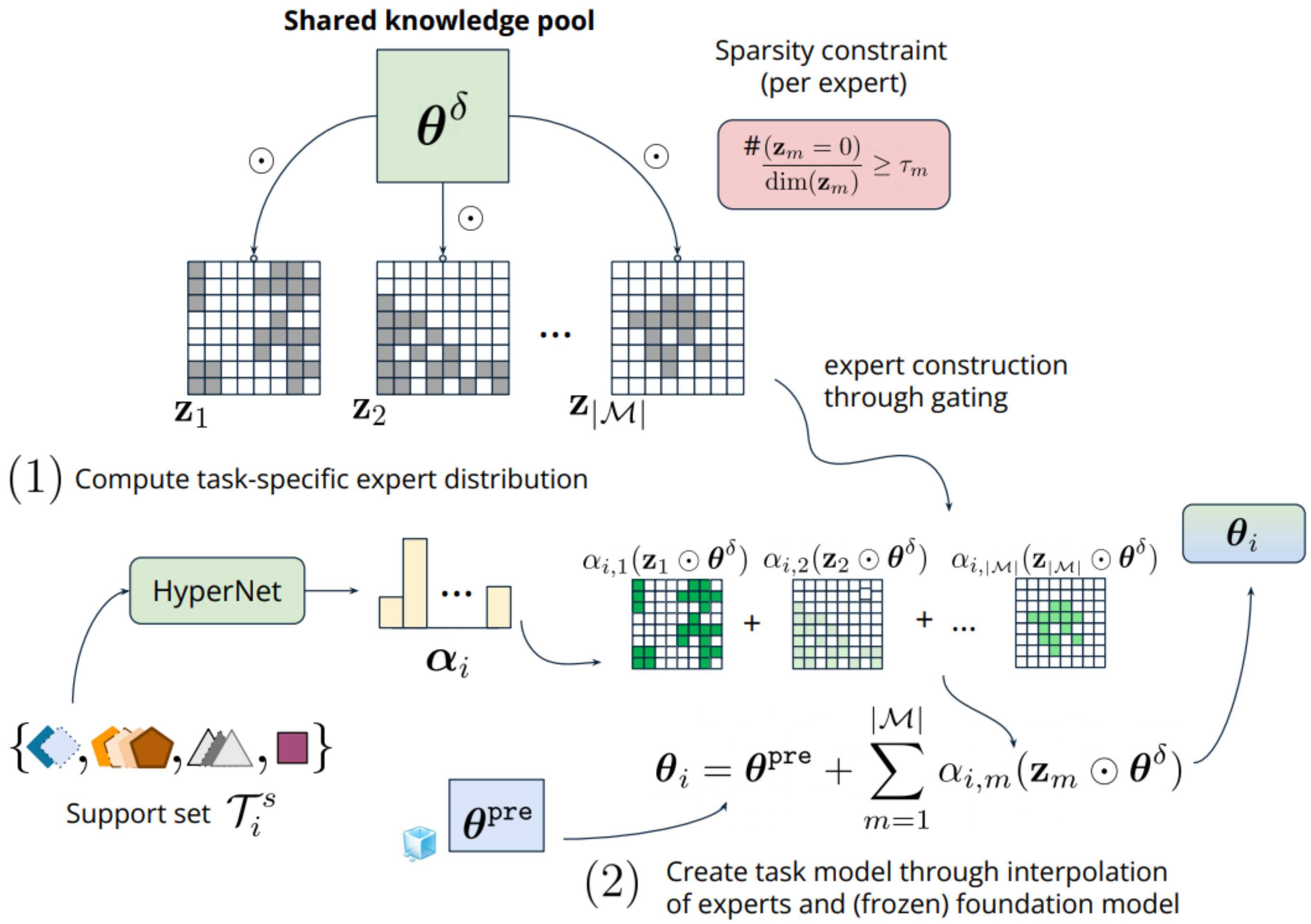
Fully **task-specific**
E.g., Task experts

$\theta_1 = \theta_2 = \theta_3$	$\theta_i = \sum_m \alpha_{i,m} \cdot \theta_m^{\text{expert}}$	θ_1 θ_2 θ_3
High interference **Allow** transfer	**Reduced** interference **Allow** transfer	**No** interference **No** transfer

SMAT: overview



SMAT: meta-training



Task-specific expert merging

- $|\mathcal{M}|$ distinct sparse experts, $(\boldsymbol{z}_m \odot \boldsymbol{\theta}^{\delta})$
- a common merging rule defined by a hypernetwork that outputs task-specific expert distribution $\boldsymbol{\alpha}_i$ based on the task support set.

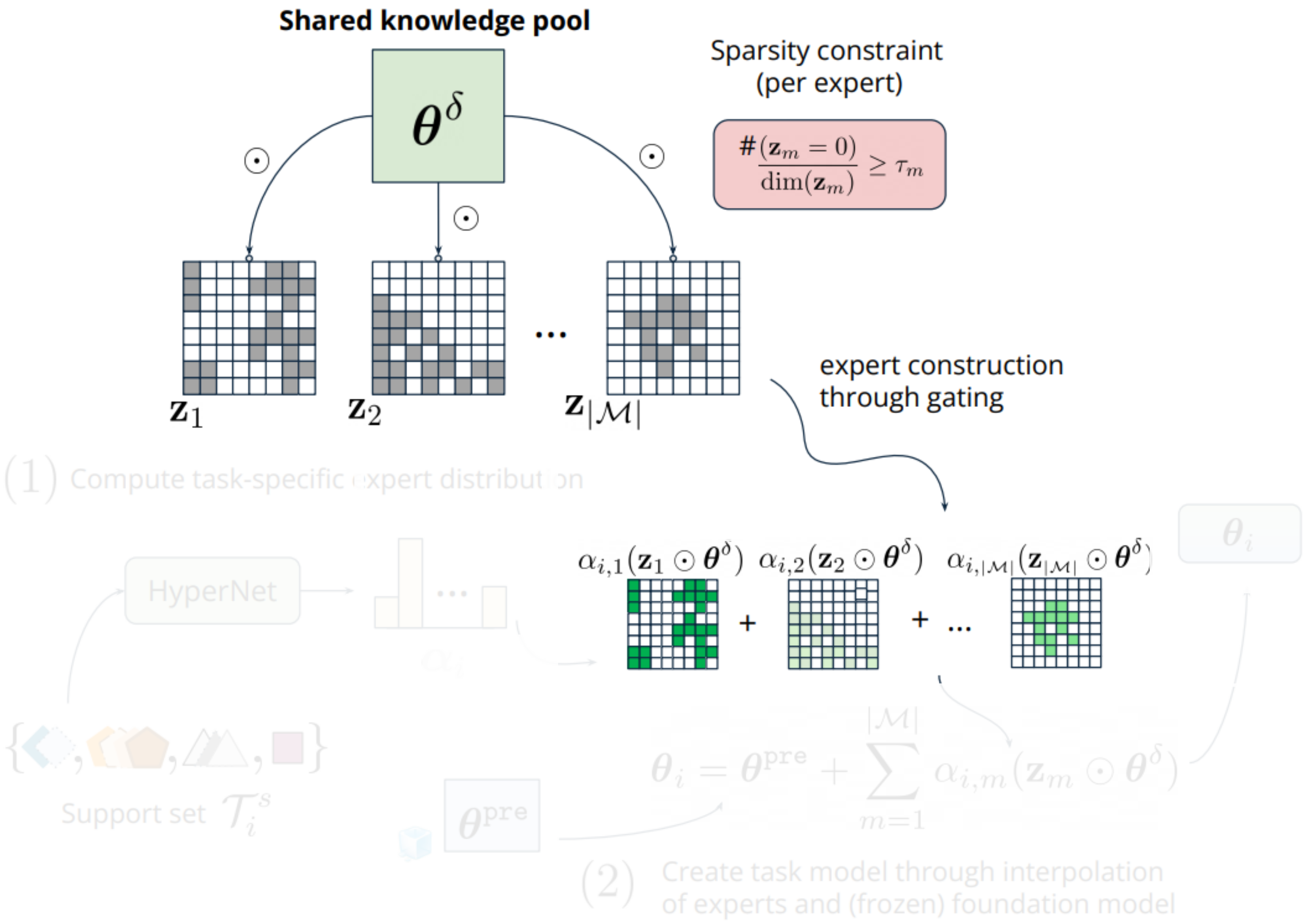
Meta-learn sparse reparameterization

- $\{\boldsymbol{z}_m\}$: **distinct**, **sparse**, binary gates $\in [0,1]^{|\theta|}$
- $\boldsymbol{\theta}^{\delta}$: a **shared**, **dense** reparameterization
- Thus, meta-learn $\boldsymbol{z}_m$ end-to-end effectively discover *where-to-share* and *where-to-specialize.*
- **No bias** as in hand-crafted task-specific/task-agnostic partition

Interpolation between common θ^{pre} and θ^{δ}

- Sparsity of $\{\boldsymbol{z}_m\}$(also α_i) controls the relative interpolation strength
- $\{\boldsymbol{z}_m\}$ define the big picture:
  - More 1s = More like the meta-tuned model
  - More 0s = More like the pre-trained model
- α_i allows local, task-specific variation:
  - Different expert distribution across tasks

SMAT: meta-training



Task-specific expert merging

- $|\mathcal{M}|$ distinct sparse experts, $(\mathbf{z}_m \odot \theta^\delta)$
- a common merging rule defined by a hypernetwork that outputs task-specific expert distribution α_i based on the task support set.

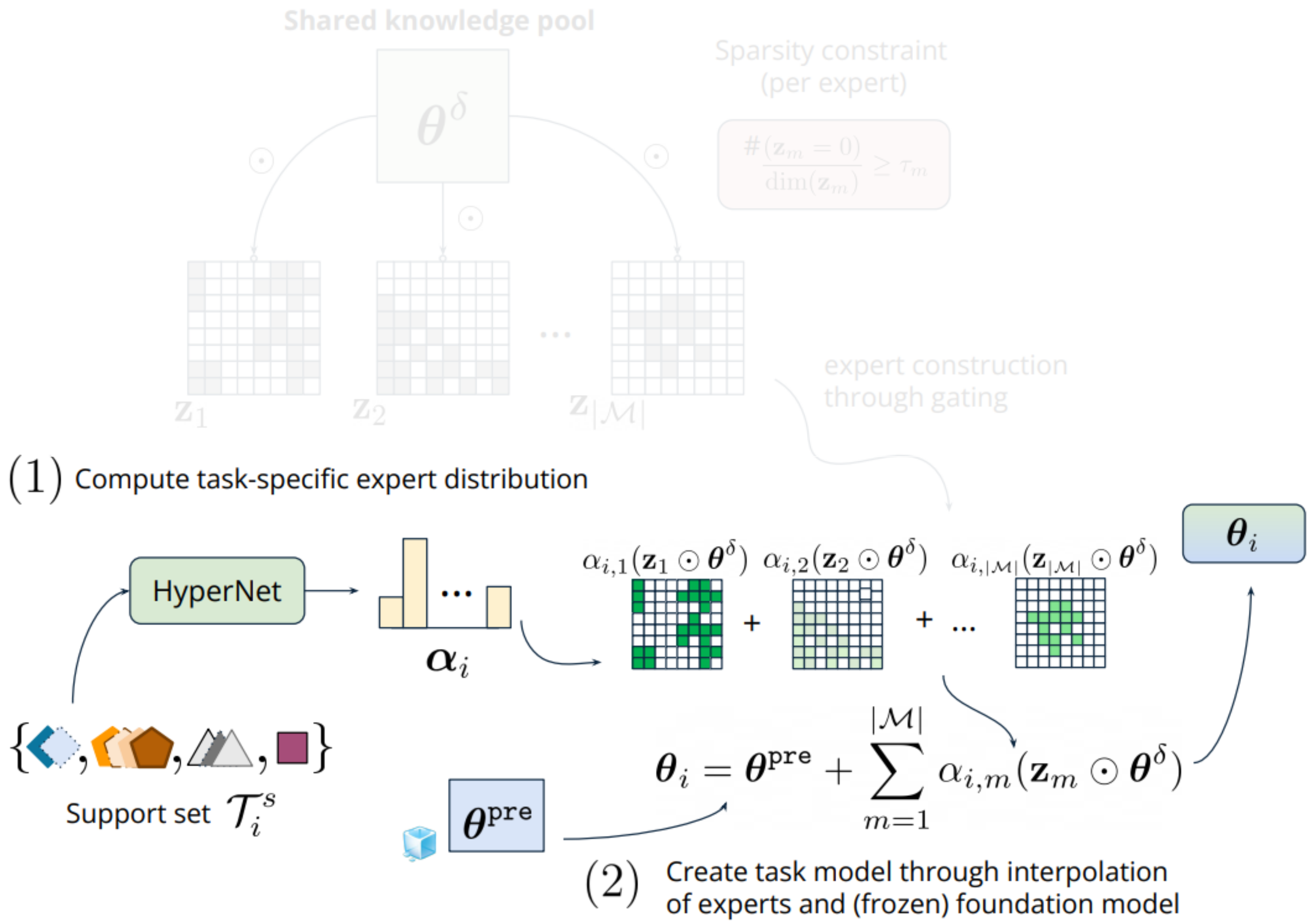
Meta-learn sparse reparameterization

- $\{\mathbf{z}_m\}$: **distinct**, **sparse**, binary gates $\in [0,1]^{|\theta|}$
- θ^δ: a **shared**, **dense** reparameterization
- Thus, meta-learn $\mathbf{z}_m$ end-to-end effectively discover *where-to-share* and *where-to-specialize.*
- **No bias** as in hand-crafted task-specific/task-agnostic partition

Interpolation between common θ^{pre} and θ^δ

- Sparsity of $\{\mathbf{z}_m\}$(also α_i) controls the relative interpolation strength
- $\{\mathbf{z}_m\}$ define the big picture:
  - More 1s = More like the meta-tuned model
  - More 0s = More like the pre-trained model
- α_i allows local, task-specific variation:
  - Different expert distribution across tasks

SMAT: meta-training



Task-specific expert merging

- $|\mathcal{M}|$ distinct sparse experts, $(\mathbf{z}_m \odot \theta^\delta)$
- a common merging rule defined by a hypernetwork that outputs task-specific expert distribution α_i based on the task support set.

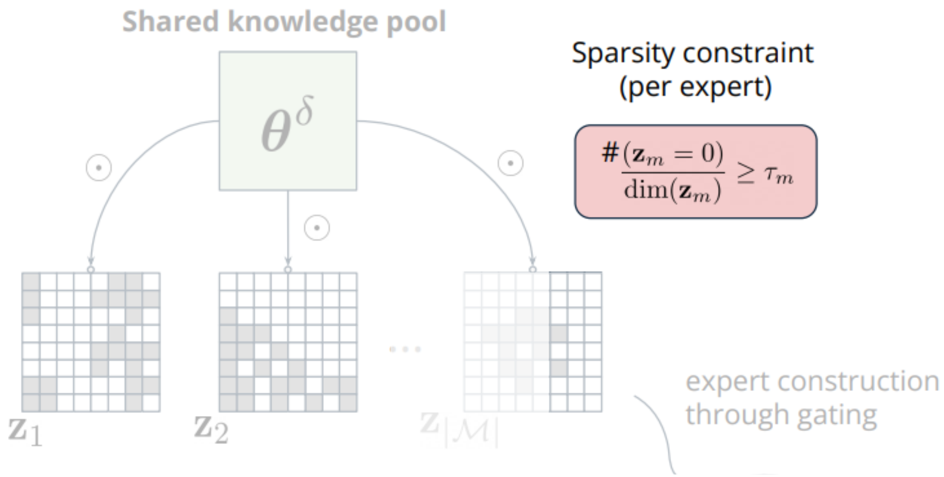
Meta-learn sparse reparameterization

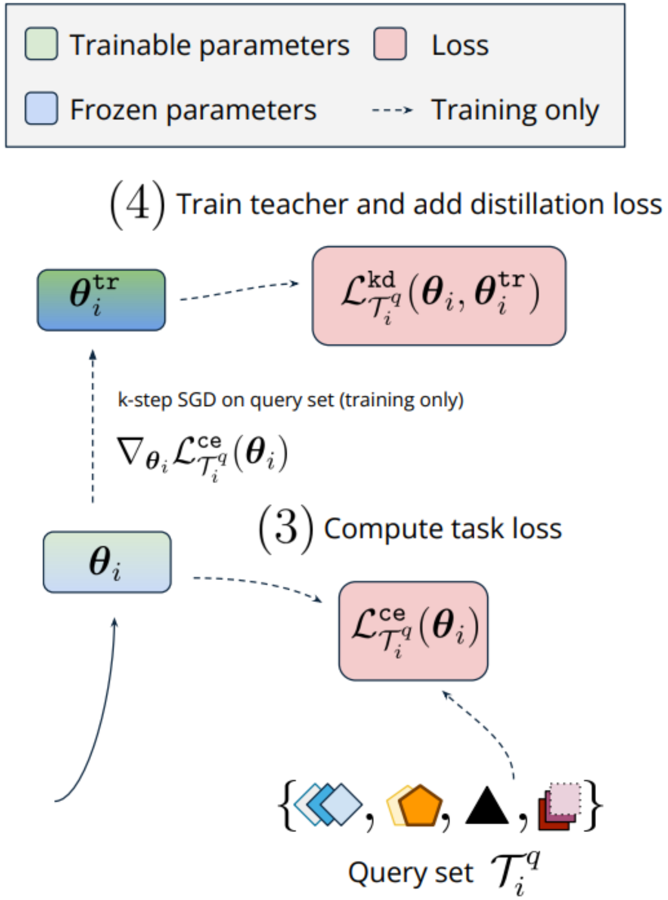
- $\{\mathbf{z}_m\}$: **distinct**, **sparse**, binary gates $\in [0,1]^{|\theta|}$
- θ^δ: a **shared**, **dense** reparameterization
- Thus, meta-learn $\mathbf{z}_m$ end-to-end effectively discover *where-to-share* and *where-to-specialize*.
- **No bias** as in hand-crafted task-specific/task-agnostic partition

Interpolation between common θ^{pre} and θ^δ

- Sparsity of $\{\mathbf{z}_m\}$(also α_i) controls the relative interpolation strength
- $\{\mathbf{z}_m\}$ define the big picture:
  - More 1s = More like the meta-tuned model
  - More 0s = More like the pre-trained model
- α_i allows local, task-specific variation:
  - Different expert distribution across tasks

SMAT: meta-training

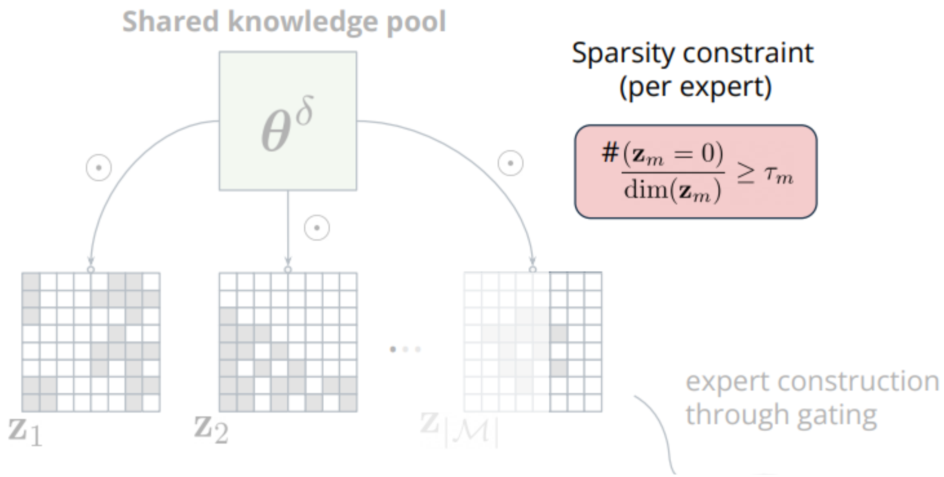


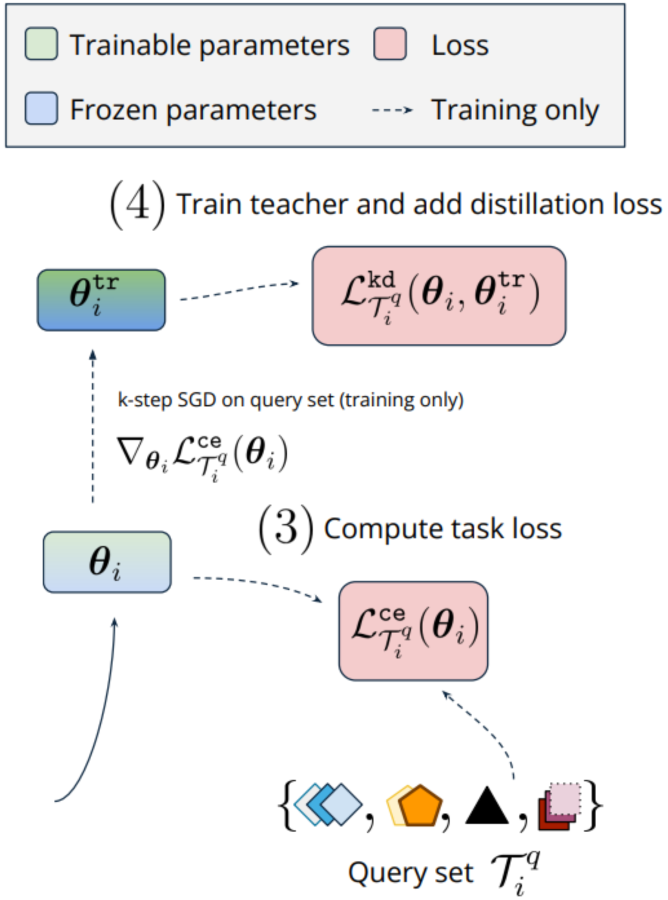


Task-specific dense teachers

- $\boldsymbol{\theta}_i^{tr}$: a **unconstrained, highly task-specific** teacher
- $\boldsymbol{\theta}_i$: merged model, **implicitly constrained** by sparsity constraints on individual experts.
- **Knowledge distillation** enforces the student to mimic the teacher, therefore encourages specialization and cooperation among experts.

SMAT: meta-training





Everything together into a Lagrangian

$$\min_{\boldsymbol{\theta}^\delta, \boldsymbol{\zeta}, \boldsymbol{\Phi}} \max_{\boldsymbol{\lambda} \geq 0} \mathbb{E}_{\mathcal{T}_i \sim \mathcal{P}_{ID}} \left[\mathcal{L}^{\mathrm{ce}}_{\mathcal{T}_i^q}(\boldsymbol{\theta}_i) + \mathcal{L}^{\mathrm{kd}}_{\mathcal{T}_i^q}(\boldsymbol{\theta}_i, \boldsymbol{\theta}_i^{\mathrm{tr}}) \right]$$

$$+ \sum_{m=1}^{|\mathcal{M}|} \lambda_m \Big(\frac{1}{|\phi_m|} \sum_{k=1}^{|\phi_m|} \tau - Q_{\phi_m}(s_k \leq 0) \Big)$$

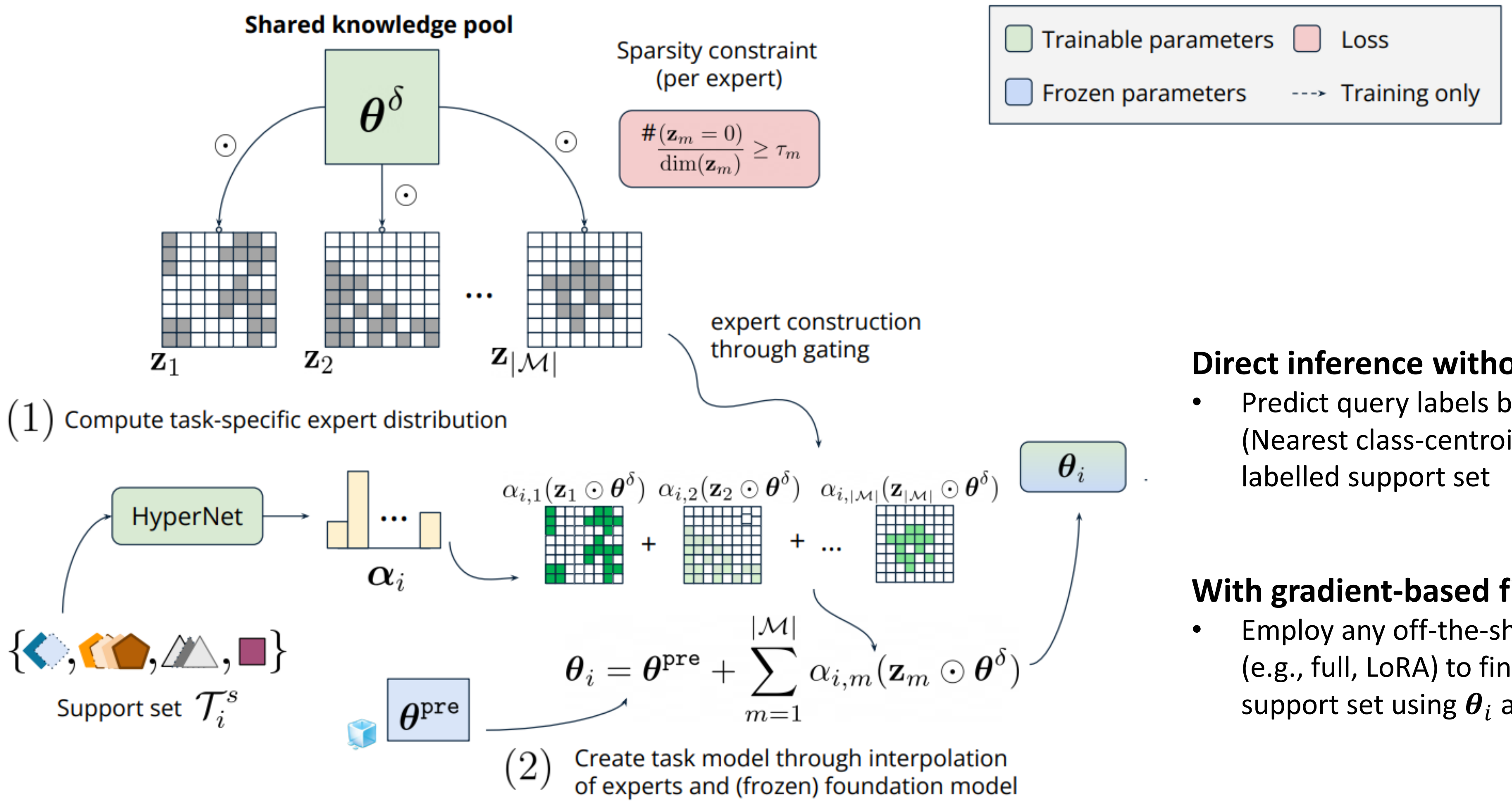
$$\text{where} \quad \theta_i = \theta^{\mathrm{pre}} + \sum_{m=1}^{|\mathcal{M}|} \alpha_{i,m} (z_m \odot \theta^\delta),$$

$$z_m \sim q_{\phi_m}; \alpha_i \sim h_\zeta(\mathcal{T}_i^s),$$

Q_{ϕ_m}: The CDF of the variational distribution for optimization z_m

We share the sparsity constraint τ for all experts for simplicity

SMAT: inference



Direct inference without fine-tuning

- Predict query labels by constructing a ProtoNet (Nearest class-centroid) classifier with $\boldsymbol{\theta}_i$ and the labelled support set

With gradient-based fine-tuning

- Employ any off-the-shelf fine-tuning techniques (e.g., full, LoRA) to fine-tune the model on the support set using $\boldsymbol{\theta}_i$ as an initialization.

Experimental results

Comparing with the SOTA on Meta-dataset

- **SMAT (Ours) outperforms baselines under all evaluation scenarios**

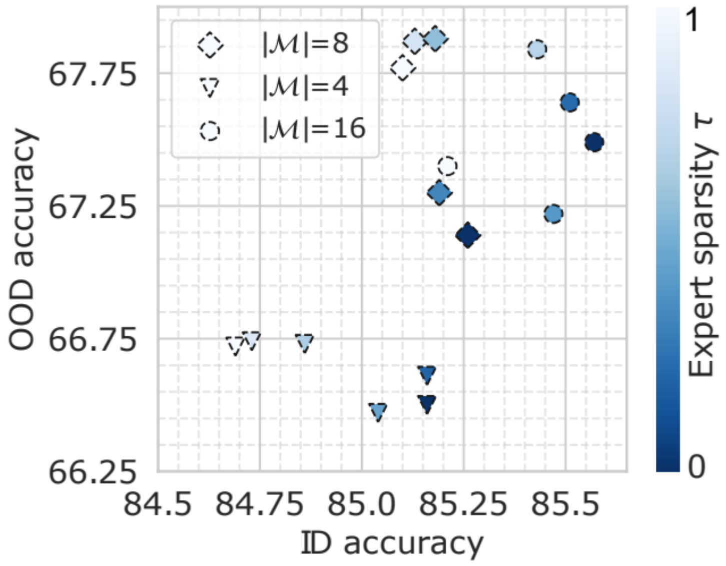
Table 1. Few-shot testing results on the Meta-dataset benchmark and additional OOD testing datasets for methods using DINO-ViT-Small backbone. † and ‡ respectively indicate published results in †(Hu et al., 2022) and ‡(Basu et al., 2023). Gray indicates our method.

	w/o fine-tuning				with gradient-based fine-tuning						
Datasets	†Pre	†PMF	SoftMerge	SMAT	†Pre+full	†PMF+full	SoftMerge+full	SMAT+full	‡Pre+LoRA	PMF+LoRA	SMAT+LoRA
ImageNet	73.48	73.54	74.33	**74.94**	73.54	74.59	74.71	75.24	74.22	73.54	**75.72**
Aircraft	62.17	88.33	88.80	**89.49**	75.4	88.33	90.6	**90.78**	80.8	89.75	90.71
Omniglot	54.33	**91.79**	91.24	89.54	78.7	91.79	92.01	90.83	80.8	**92.78**	90.99
CUB	85.37	91.02	91.54	**92.48**	85.4	91.02	91.95	**92.48**	85.8	91.17	92.42
DTD	83.67	81.64	80.98	**85.86**	86.9	86.61	86.84	**88.34**	86.8	86.73	88.28
Quickdraw	60.59	**79.23**	78.98	78.83	73.6	**79.23**	79.90	78.83	72.7	**79.23**	78.83
Fungi	56.26	**74.2**	72.40	72.8	54.7	74.20	72.40	72.80	59.8	**75.44**	72.80
VGGFlower	94.45	94.12	96.89	**97.19**	94.2	94.12	97.01	97.19	94.8	96.05	**97.25**
ID Avg	71.29	84.23	84.40	**85.14**	77.81	84.99	85.56	**85.81**	79.47	85.59	**85.88**
TrafficSig	53.7	54.37	56.21	**58.51**	87.3	88.85	89.91	**90.83**	88.1	89.14	90.18
MSCOCO	54.58	57.04	55.75	**57.35**	61.5	62.59	62.15	63.07	62.1	61.71	**63.38**
Cifar10	**85.64**	80.82	84.58	83.95	**92.48**	89.61	91.84	92.08	93.33	91.53	92.46
Cifar100	**76.86**	69.11	70.85	74.85	86.13	82.54	85.88	85.91	**86.17**	85.06	85.88
MNIST	78.57	93.33	94.16	**94.53**	92.54	96.44	96.20	**96.73**	94.98	96.41	96.46
Sketch	47.25	41.10	43.30	**48.91**	56.39	49.65	53.85	56.55	**57.34**	47.59	55.63
Food	91.73	91.37	89.84	**92.31**	92.03	91.73	90.48	**92.31**	92.06	92.01	92.31
Clipart	55.19	53.92	54.83	**59.87**	**67.18**	62.83	65.50	65.76	66.51	60.6	66.07
Pet	62.64	61.89	63.04	**65.59**	65.08	62.97	63.36	67.43	65.06	62.71	**67.77**
Cars	34.58	**38.00**	36.21	36.79	40.98	40.07	41.62	**42.39**	39.49	42.37	40.05
OOD Avg	64.07	64.10	64.87	**67.27**	74.16	72.73	74.08	**75.31**	74.51	72.91	**75.02**

More analysis

The roles of sparsity level τ for SMAT

1. **Sparsity level establishes a trade-off between ID and OOD generalization performance**
2. Appropriate sparsity level encourages expert specialization

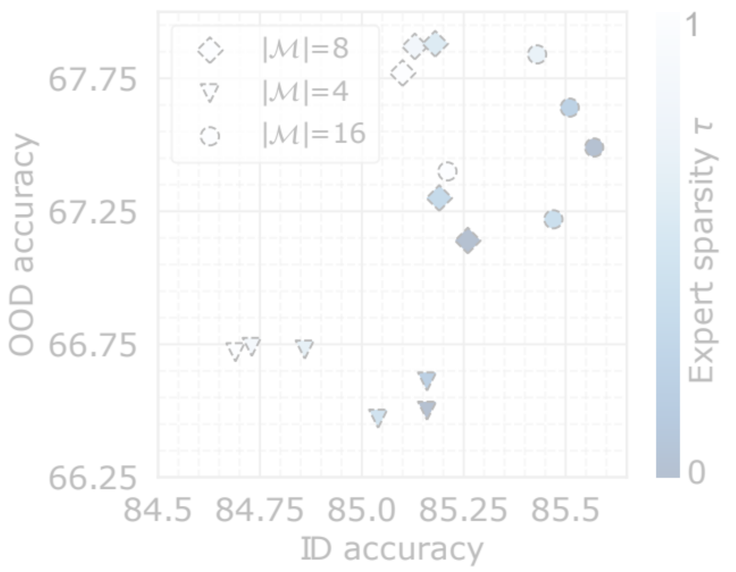


(a) Average performance tradeoff on sampled ID vs OOD tasks as a function of (color) expert sparsity level τ, and (marker) number of experts.

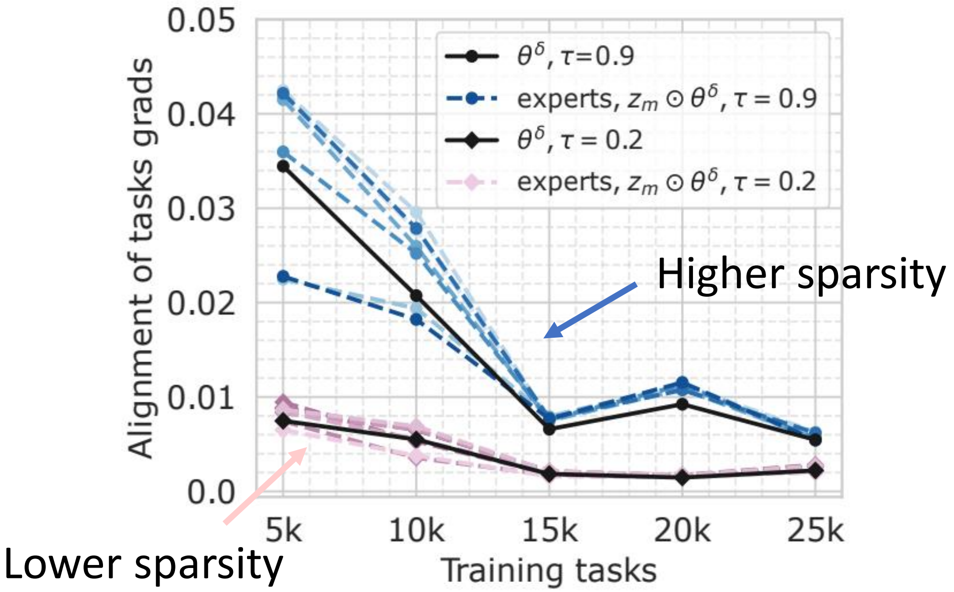
More analysis

The roles of sparsity level τ for SMAT

1. Sparsity level controls the trade-off between ID and OOD performance
2. **Appropriate sparsity level encourages gradient alignment between tasks**



(a) Average performance tradeoff on sampled ID vs OOD tasks as a function of (color) expert sparsity level τ, and (marker) number of experts.



(b) Meta-gradients alignment between tasks throughout for SMAT with low and high sparsity levels. Meta-gradients are calculated w.r.t. the parameters shown in the legend.

Qualitative visualization

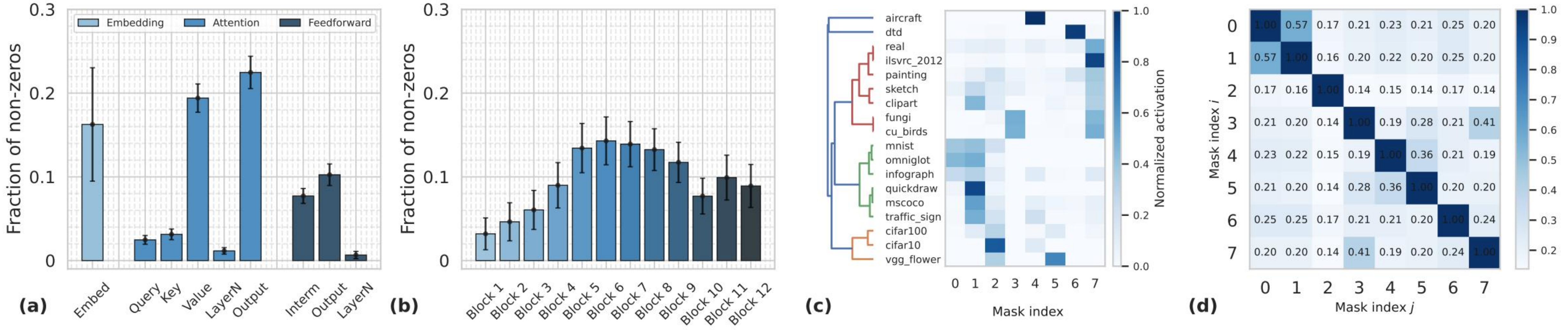
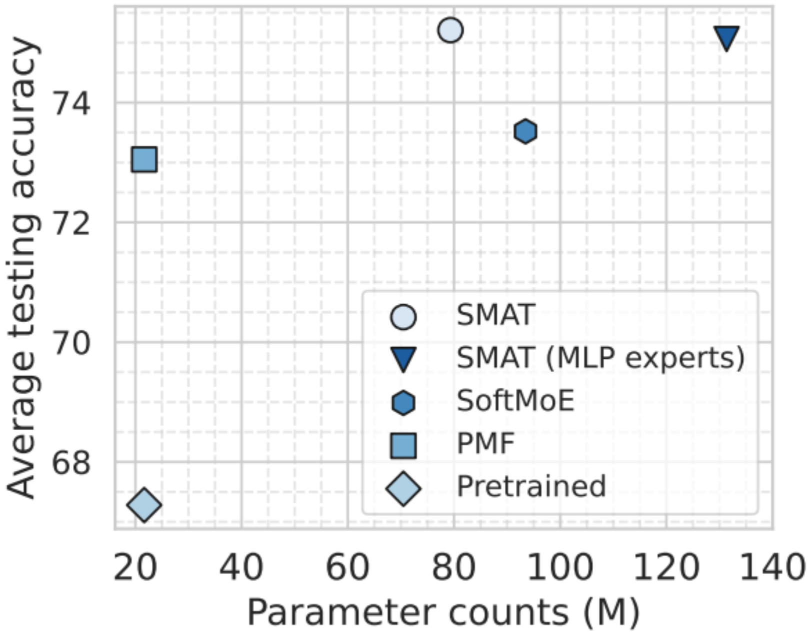


Figure 5. (a-b) Model capacity (i.e., number of non-zero parameters) grouped by ***a)***: layer types, and ***b)***: layer depth. (c-d) Expert specialisation. ***c)*** Dendrogram of task similarity computed on different tasks. ***d)*** Overlap between masks.

More ablation study

Performance vs scale for different models



Ablation experiments

Table 2. Ablation studies on different components of SMAT. **MLS** meta-learned sparsity, **Meta**: Meta-training using support and query splits (otherwise no split), **DT**: dense teachers. **IE**: interpolated experts

ID	MODEL	MLS	META	DT	IE	ID	OOD	AVG
1	SMAT	✓	✓	✓	✓	85.14	**67.27**	**75.21**
2		✓	✓	✗	✓	85.07	66.44	74.74
3		✓	✓	✓	✗	84.77	67.02	74.90
4		✓	✗	✓	✓	82.35	63.64	71.95
5		✗	✓	✓	✗	**85.21**	66.21	74.75
6	PMF	✗	✗	✗	✗	84.23	64.09	73.05