F4MwCons | NeurIPS 2026

Small Model Portfolios for Many Deployment Profiles: Submodular Coverage under Bundled Constraints

Ji Cheng · City University of Hong Kong

CIFAR-10 / ONE-MODEL BUDGET

+53.6

percentage points

Greedy covers 97.7% of coverable requests, compared with 44.1% for fixed FewForMany.

FULL CONDITIONAL COVERAGE

2 models

chosen by Greedy per run

The pair served every coverable test request in all 90 runs (18 families × 5 seeds).

DEPLOYMENT PROBLEM

Which models should a team keep?

A request can set a minimum accuracy and a maximum response time on each of two devices. One stored model must meet all those limits.

Four models and four requests

Model	Request 1	Request 2	Request 3	Request 4
x1 • selected	**1**	**1**	–	–
x2	–	1	–	–
x3 • selected	–	–	**1**	**1**
x4	–	–	–	1

1 = all limits met. Selecting x1 and x3 covers every request in this example.

SELECTION RULE

Greedy adds complementary models

1 Build a feasibility matrix
Mark each request that each measured model can serve.

2 Choose the next model
Maximize the total weight of requests newly served by this choice.

3 Update the covered requests
Repeat until the portfolio contains K models.

Why the rule works

Weighted coverage is monotone submodular. For a fixed set of profiles, Greedy earns at least 1 − 1/e of the optimal coverage at the same budget.

EVALUATION

Tests on a measured model archive

15,625 archived models

18 profile families

5 seeds per family

HW-NAS-Bench and NAS-Bench-201 on CIFAR-10: three device pairs × three threshold schedules × energy on or off.

Each run uses 100 training and 100 test requests. The reported means weight all 18 families equally.

When some requests matter more

Mean weighted conditional coverage at K = 1

Request weights	Greedy	Fixed FewForMany
Business priority	**96.3%**	42.0%
Hard profiles	**96.4%**	33.2%
Zipf demand	**97.9%**	44.0%

Fixed FewForMany scores each model by its worst relative constraint violation for a request.

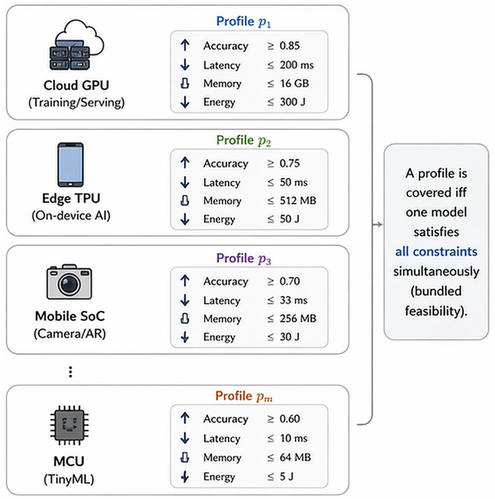
DEPLOYMENT LANDSCAPE / PAPER FIGURE 1

How one portfolio serves different devices

The illustration compares choices from the same archive. Colored regions show which profiles one selected model can serve.

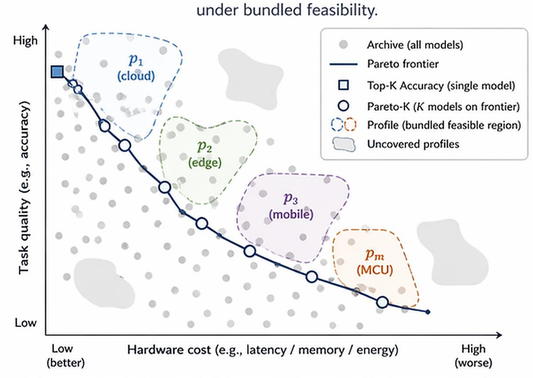
(a) Heterogeneous deployment landscape

Each deployment profile imposes a bundle of coupled constraints.



(b) Why marginal views fall short

Trade-off (marginal) perspectives miss many profiles under bundled feasibility.



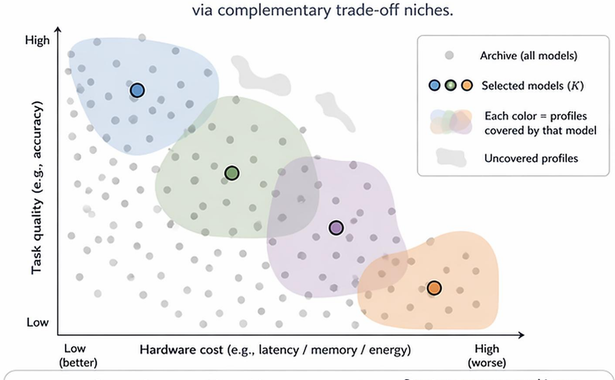
Top-K Accuracy (picks one model) — Excels in one niche, but leaves most profiles uncovered.

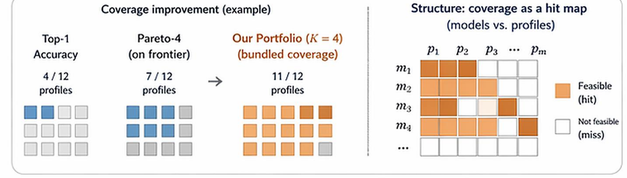
Pareto-K (models on frontier) — Models cluster in similar trade-off regions and are still blocked by some dimensions (e.g., memory).

Uncovered profiles (gray regions) — No model satisfies the bundled constraints.

(c) Bundled portfolio selection (ours)

Select K models to maximize coverage of profiles via complementary trade-off niches.

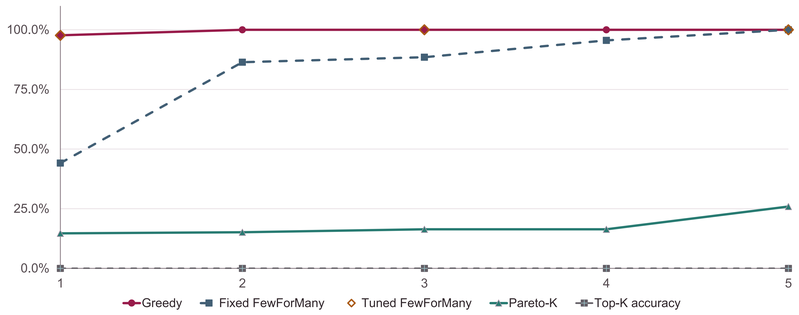




PORTFOLIO SIZE / CIFAR-10

Coverage rises with portfolio size

Mean conditional test coverage (%)



K is the number of models kept. Diamonds mark validation-tuned settings.

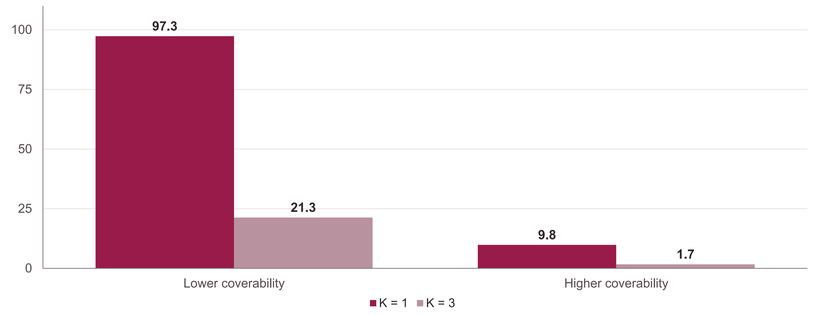
What changes with extra validation?

The tuned baseline selects its scoring rule on 100 extra validation requests per run. At K = 1, it reaches 97.67% against 97.70% for Greedy.

FAMILY COMPARISON / CIFAR-10

Gains are largest in harder families

Greedy minus fixed FewForMany (percentage points)



Nine families per group; median archive-coverable share = 62.3%.

Which requests count?

Conditional coverage counts only test requests at least one archive model can serve. By K = 5, both Greedy and fixed FewForMany cover all such requests.

F4MwCons

Paper and reproducibility materials

Benchmark setup, experiment scripts, and the records behind the figures.

Paper
NeurIPS poster 155994

Code
jicheng9617/F4MwCons