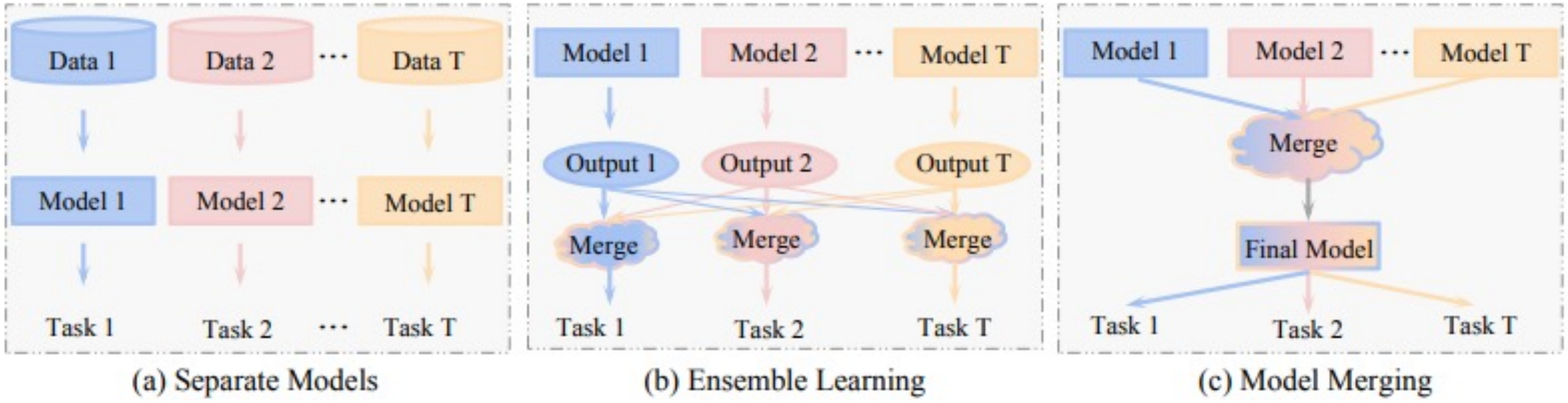
AdaMMS: Model Merging for Heterogeneous Multimodal Large LanguageModels with Unsupervised Coefficient Optimization

Yiyang Du*, Xiaochen Wang*, Chi Chen*, Jiabo Ye, Yiru Wang,
Peng Li, Ming Yan, Ji Zhang, Fei Huang, Zhifang Sui, Maosong Sun, Yang Liu

Tsinghua University, Beijing, China

Model merging

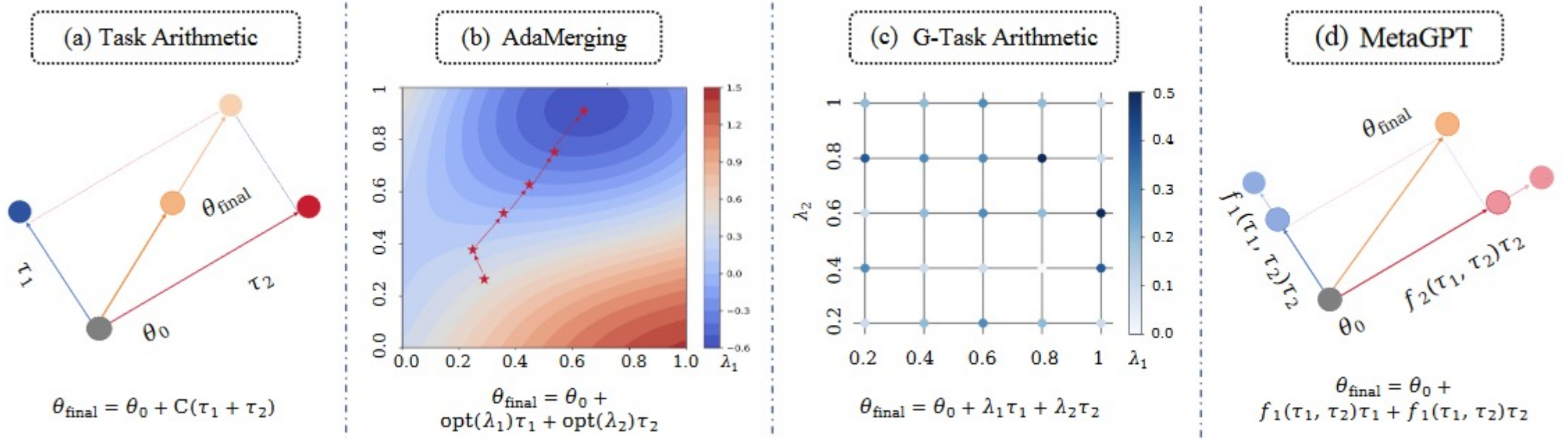
- Integrating capabilities of differently fine-tuned LLMs via operation on models' parameter space.



Yu, et al. Model Merging in LLMs, MLLMs, and Beyond: Methods, Theories, Applications and Opportunities.

Model merging

- Current model merging methods
- Mainly focus on homogeneous LLMs



Zhou, et al. MetaGPT: Merging Large Language Models Using Model Exclusive Task Arithmetic.

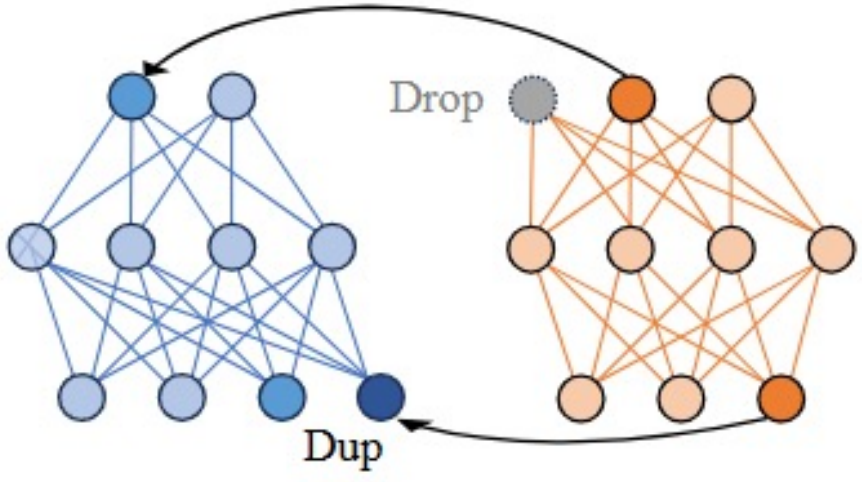
Model merging

- Research question:

  *How to apply model merging on **heterogeneous** models?*

  Different model architecture in MLLMs:

  - Language model architecture
  - Modality encoder choice



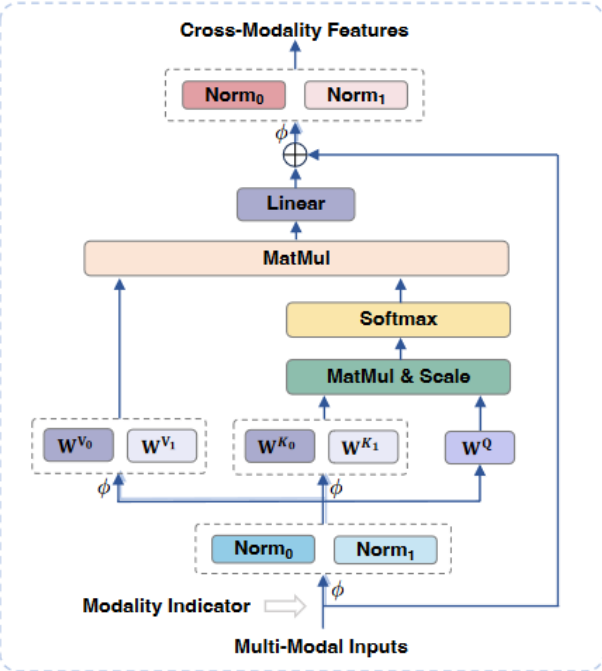
In the design of MLLMs, language model parameters may be duplicated for multimodal capability.

Model merging

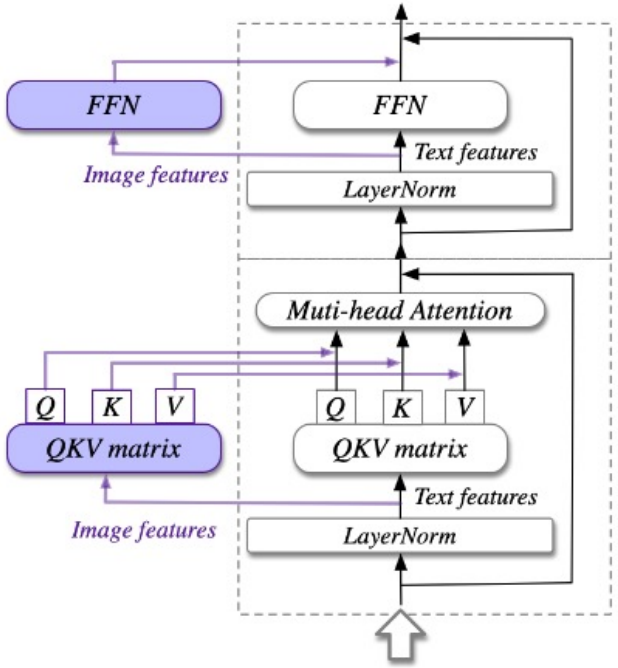
- Research question:

*How to apply model merging on **heterogeneous** models?*

Different language model architecture in MLLMs:



Wang, et al. **mPLUG-Owl2:**
Revolutionizing Multi-modal Large Language Model with Modality Collaboration



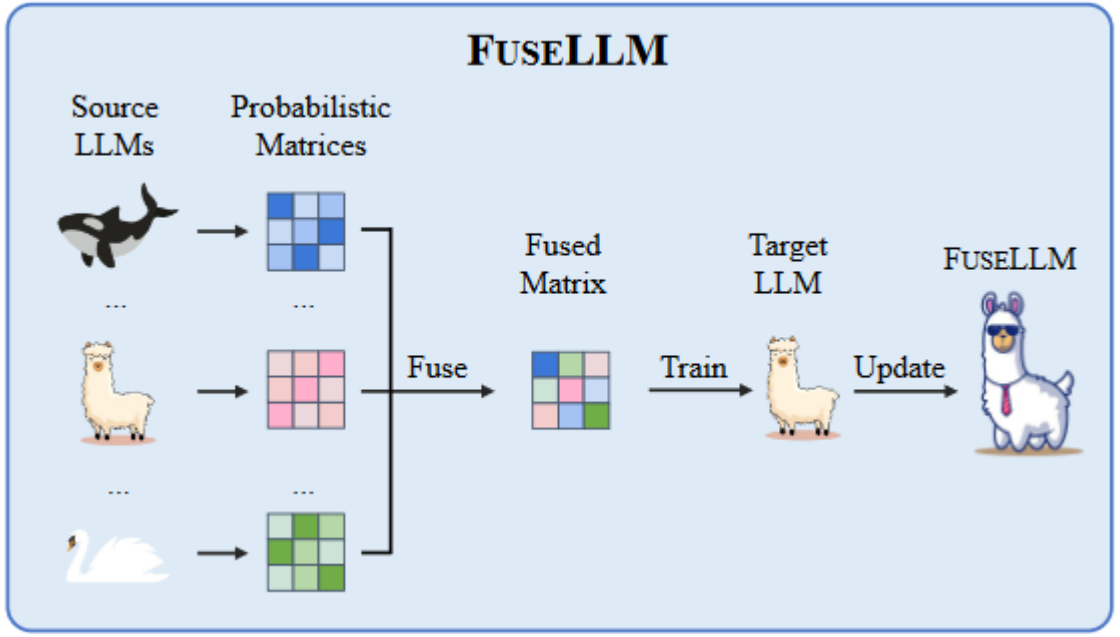
Wang, et al. **CogVLM:**
Visual Expert for Pretrained Language Models

Model merging

- Research question:

  *How to apply model merging on **heterogeneous** models?*

FuseLLM: fusing and continue training

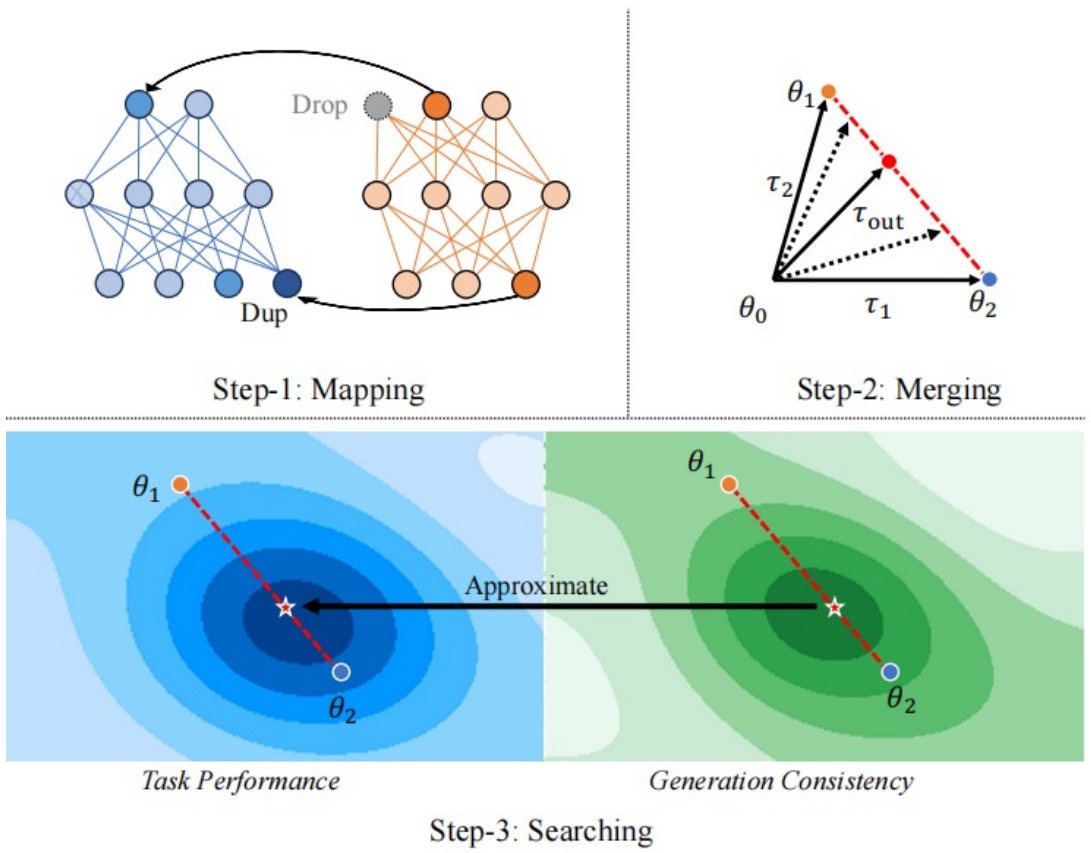


Requiring labeled data and computation resources for continue training.

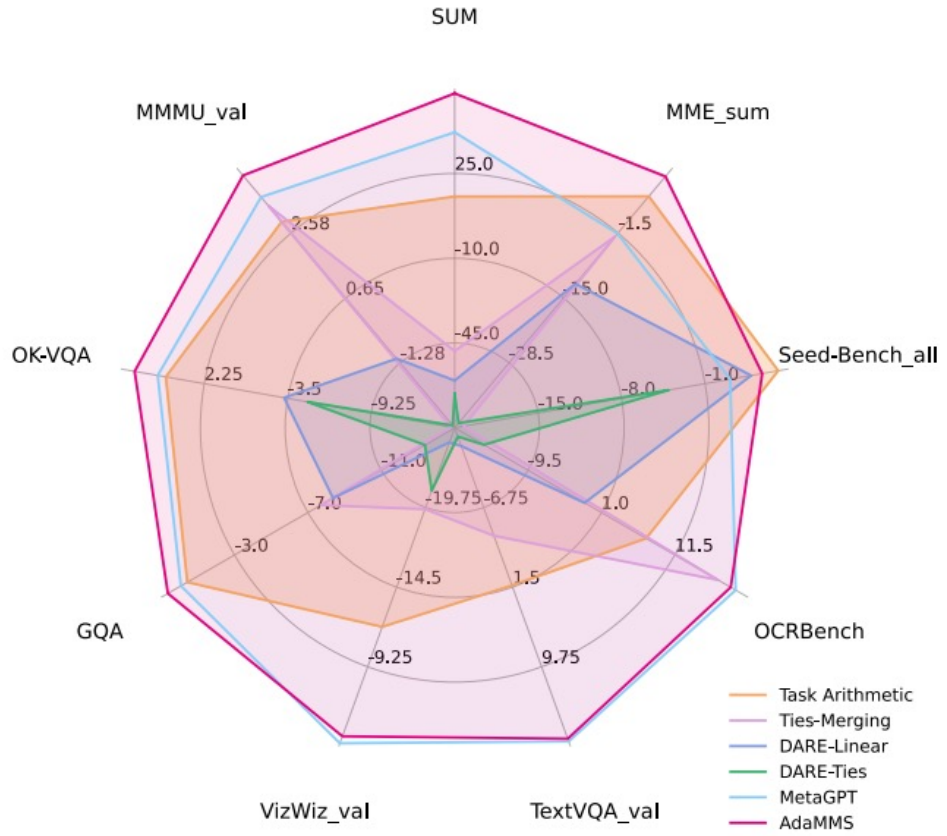
Wan, et al. Knowledge Fusion of Large Language Models

Model merging for heterogeneous MLLMs

- AdaMMS: a model merging technique for heterogeneous MLLMs



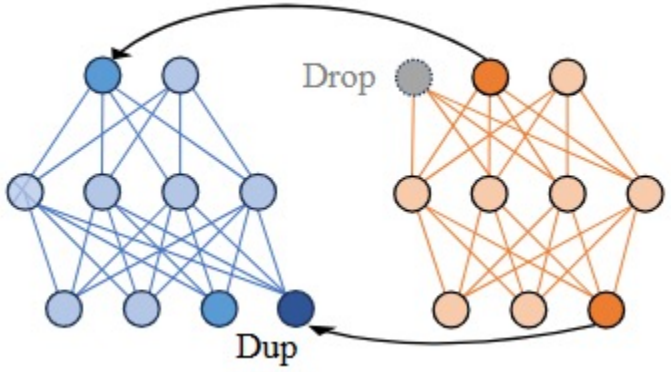
(a) AdaMMS overview.

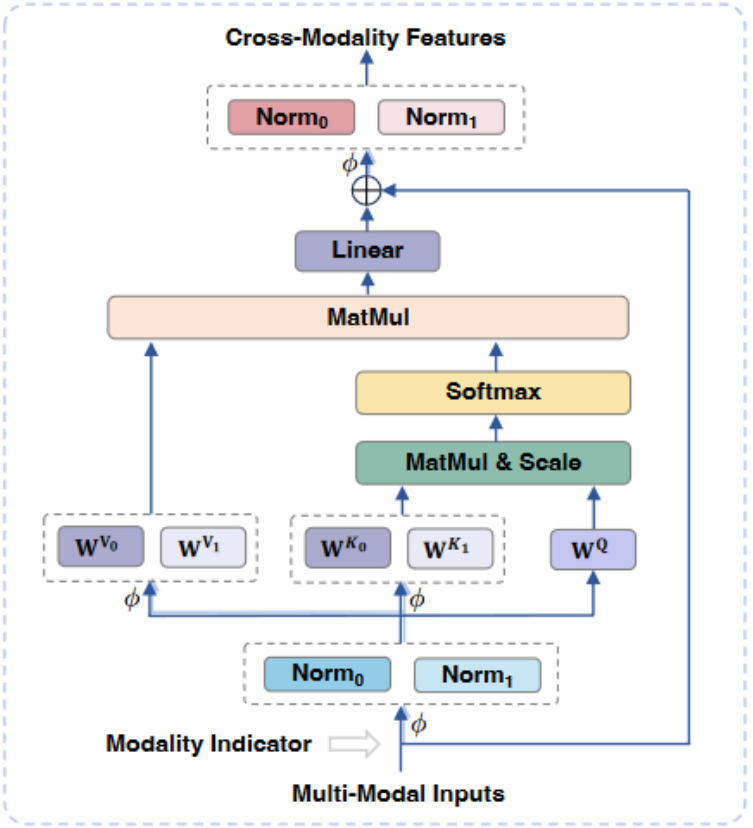


(b) Gains from different merging methods.

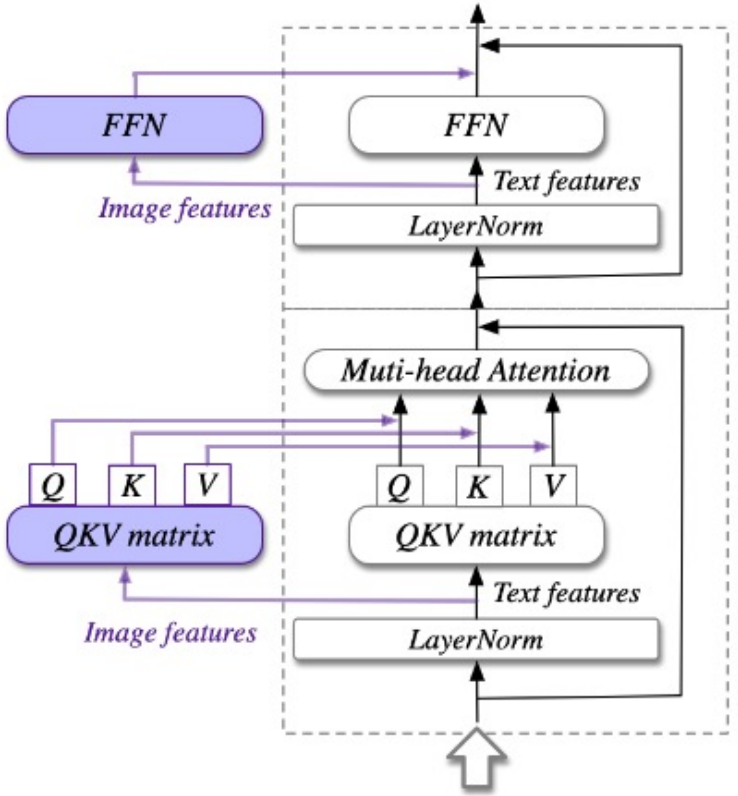
Model merging for heterogeneous MLLMs

- Mapping: match LLM weights in different architecture





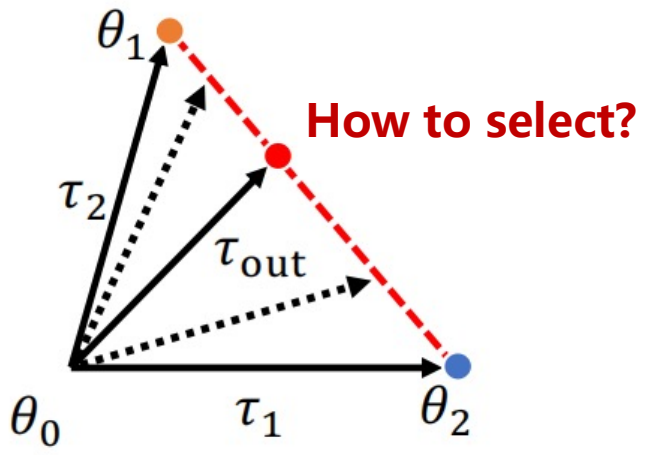
mPLUG-Owl2



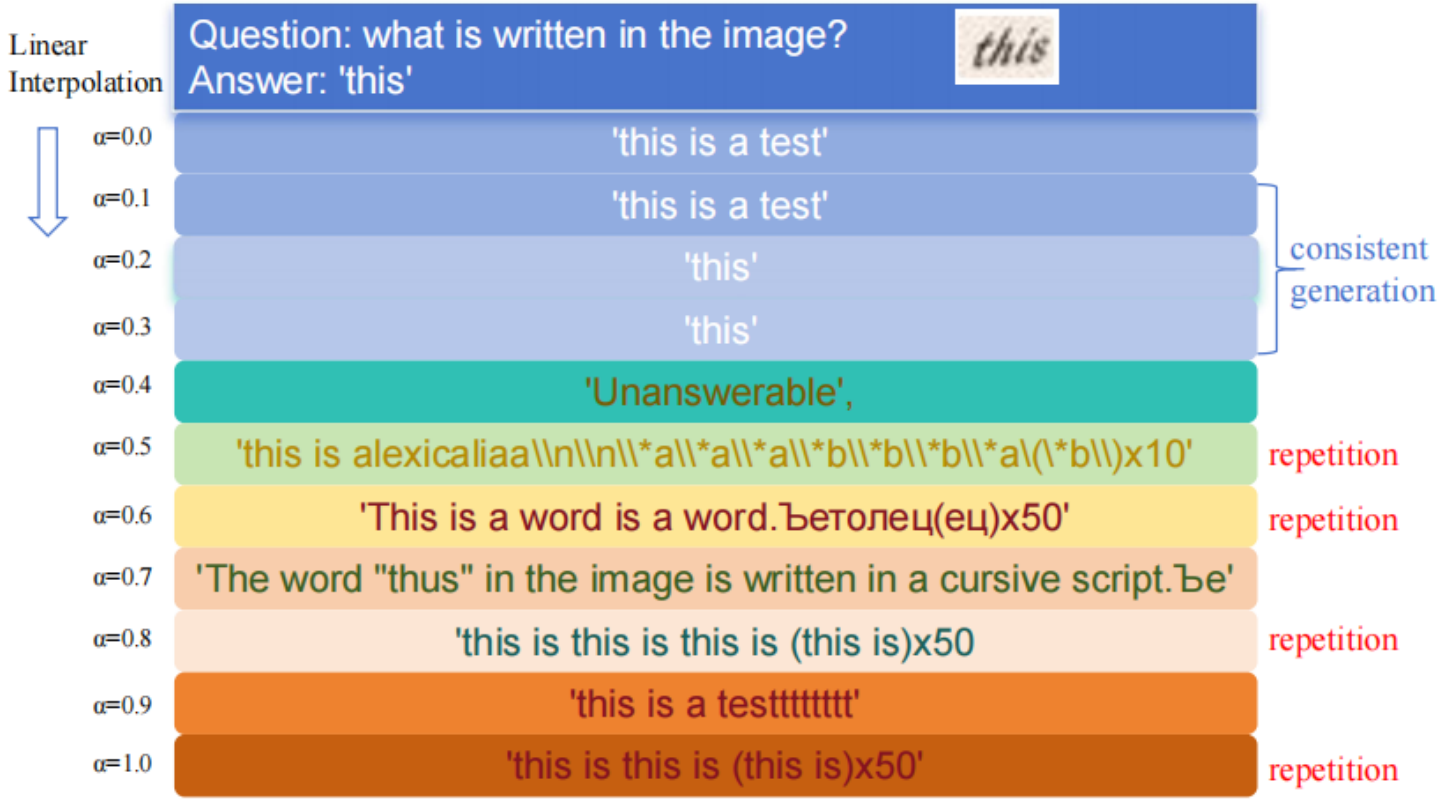
CogVLM

Model merging for heterogeneous MLLMs

- Merging: adaptive linear interpolation
- Why adaptive: asymmetry in the parameter space of two MLLMs



Step-2: Merging



Generated response of the merged model with different linear interpolation coefficients

Model merging for heterogeneous MLLMs

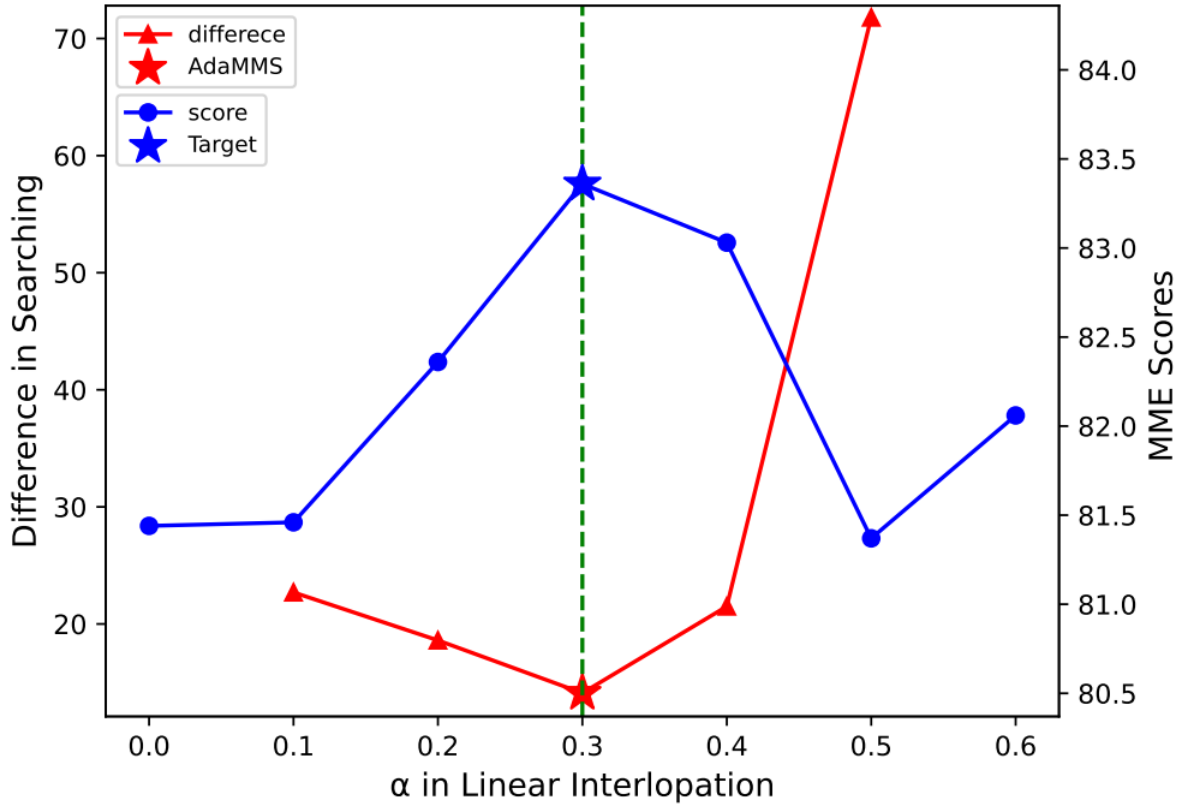
- Searching: **unsupervised** hyper-parameter selection

- Previous model merging methods generally relies on *supervised* searching on validation sets to determine hyper-parameters
  - Requires labeled data (which can be hard to obtain in some cases)
  - Suffers from the distribution gap between validation and test set

- Formally: $\hat{\alpha} = \underset{\alpha}{\operatorname{argmax}}(S_{\hat{\mathbf{t}}_\mathbf{i}}(\Theta_1, \Theta_2, \alpha)) \approx \underset{\alpha}{\operatorname{argmax}}(S_{\mathbf{t}_\mathbf{i}}(\Theta_1, \Theta_2, \alpha))$

Model merging for heterogeneous MLLMs

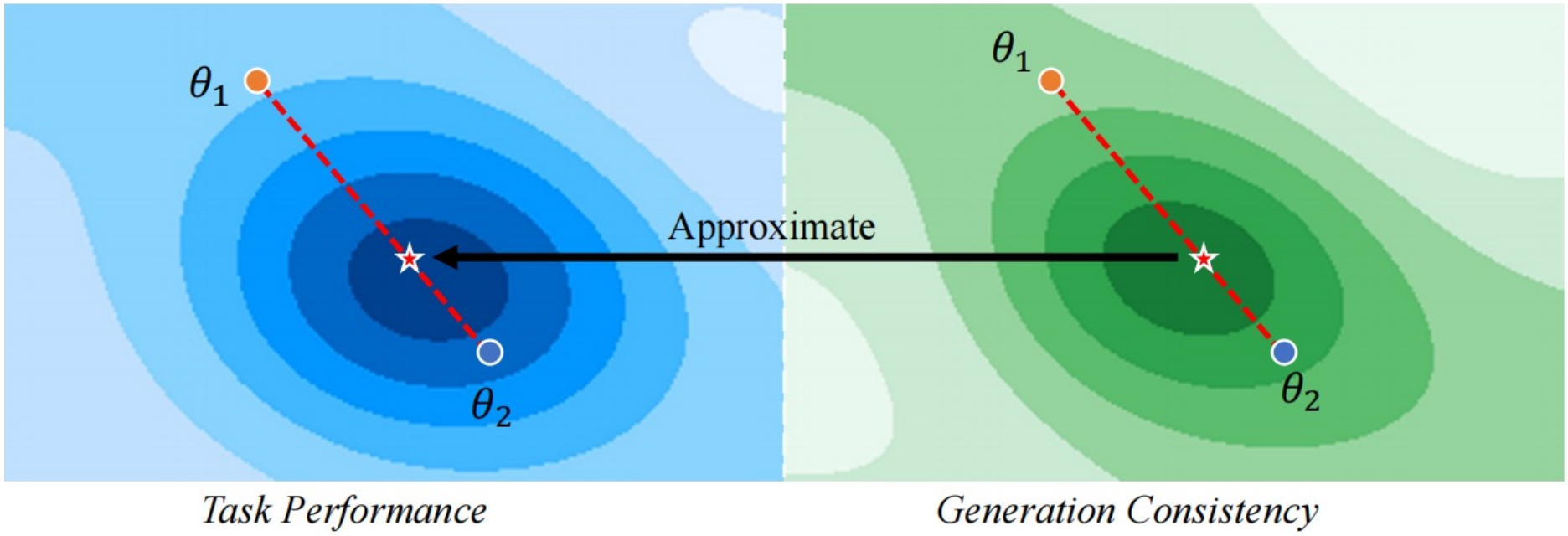
- Searching: **unsupervised** hyper-parameter selection
- Approximate task performance via **generation consistency**
- Observation: higher performance, higher consistency



Generation Consistency:
difference in generated responses compared with adjacent candidates

Model merging for heterogeneous MLLMs

- Searching: **unsupervised** hyper-parameter selection
- Approximate task performance via **generation consistency**
- Measure consistency between the adjacent parameter candidates
- Formally: $\bar{\alpha} = \underset{\alpha}{\operatorname{argmin}}(D_{t_i}(\alpha; \alpha^-, \alpha^+)) \approx \underset{\alpha}{\operatorname{argmax}}(S_{t_i}(\Theta_1, \Theta_2, \alpha))$



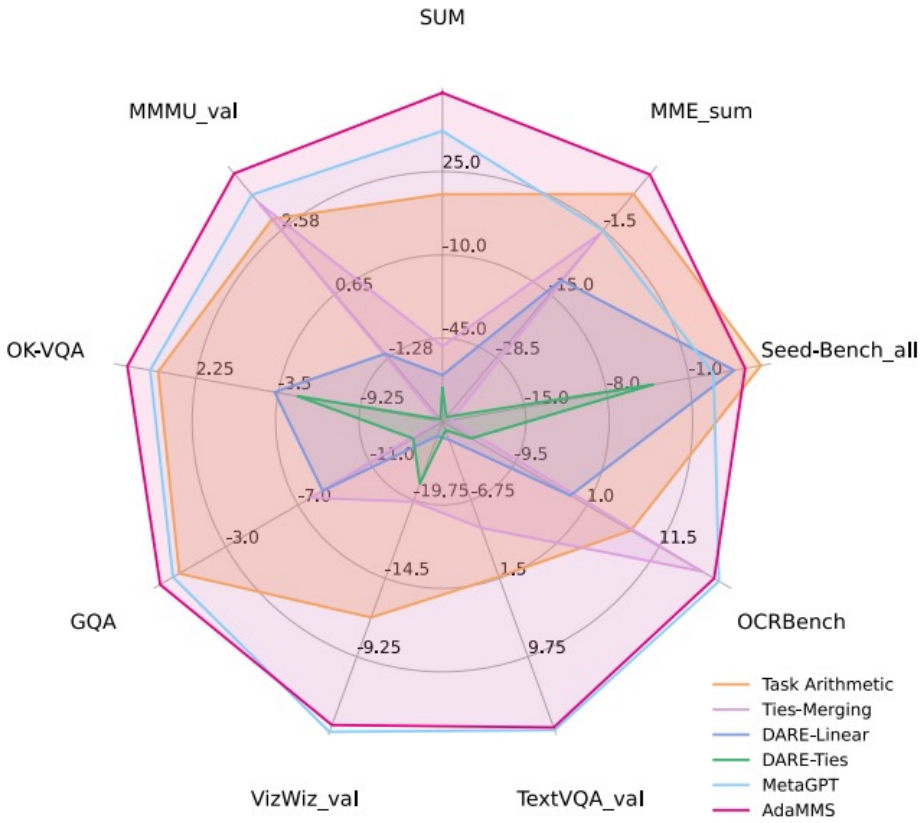
Step3: Searching

Model merging for heterogeneous MLLMs

- Searching: **unsupervised** hyper-parameter selection
- Approximate task performance via **generation consistency**
- Theoretical Proof:
  - If model performance is **convex** near the optimal point,
  - Then the highest performance point has high generation consistency
- Experimental Evidence:
  - **62.75%** of experiment settings show this **convex property**

Experiment: Setup

- We conduct experiments on various pairs of MLLMs and benchmarks
- Model pairs:
  - Qwen2-VL ← LLaVA-OneVision
  - CogVLM ← LLaVA-v1.5
  - LLaVA-v1.5 ← mPLUG-Owl2
  - mPLUG-Owl2 ← LLaVA-v1.5
  - mPLUG-Owl2 ← CogVLM
- Benchmarks:
  - MMMU, MME, SeedBench, OCRBench, TextVQA, OKVQA, GQA, VizWiz



(b) Gains from different merging methods.

Experiment: Result

- AdaMMS merges heterogeneous MLLMs with higher performance

Model	Unsupervised	$MMMU_{val}$	MME_{sum}	$SeedBench_{all}$	OCRBench	$TextVQA_{val}$	OKVQA	GQA	$VizWiz_{val}$	SUM	Top2
Original Models											
Qwen2-VL(base)		50.11	81.44	75.85	*86.00*	*84.12*	51.43	61.80	68.32	559.07	2
LLaVA-OneVision		43.44	77.04	75.44	69.60	78.47	49.57	59.84	60.97	514.37	0
Baselines											
Task Arithmetic	×	48.44(+1.67)	82.33(+3.09)	75.81(+0.17)	77.90(+0.10)	76.22(-5.08)	50.60(+0.10)	**62.26(+1.44)**	62.76(-1.89)	536.32(-0.40)	1
Ties-Merging	×	**51.11(+4.34)**	82.65(+3.41)	76.29(+0.64)	84.40(+6.60)	79.56(-1.74)	52.56(+2.06)	61.84(+1.02)	66.34(+1.69)	554.75(+18.03)	4
DARE-Linear	×	43.78(-3.00)	66.06(-13.18)	74.32(-1.33)	72.40(-5.40)	64.65(-16.65)	43.41 (-7.09)	55.13(-5.69)	50.18(-14.47)	469.93(-66.79)	0
DARE-Ties	×	45.00(-1.78)	54.43(-24.81)	74.07(-1.58)	75.20(-2.60)	78.54(-2.76)	49.61(-0.89)	58.51(-2.31)	58.05(-6.60)	493.41 (-43.31)	0
MetaGPT	✓	50.67(+3.90)	81.21(+1.97)	**76.35(+0.70)**	**85.50(+7.70)**	83.63(+2.33)	52.24(+1.74)	61.99(+1.17)	**69.16(+4.51)**	560.75(+24.03)	5
Our Method											
AdaMMS	✓	**51.11(+4.34)**	**83.36(+4.12)**	76.20(+0.55)	**85.50(+7.70)**	83.41(+2.11)	**53.56(+3.06)**	62.02(+1.20)	68.40(+3.75)	**563.56(+26.84)**	8

Results on merging LLaVA-OneVision-7B into Qwen2-VL-7B.

Experiment: Result

- AdaMMS merges heterogeneous MLLMs with higher performance

Model	Unsupervised	$MMMU_{val}$	MME_{sum}	$SeedBench_{all}$	OCRBench	$TextVQA_{val}$	OKVQA	GQA	$VizWiz_{val}$	SUM	Top2
Original Models											
CogVLM(base)		34.80	59.23	61.22	*56.50*	*77.57*	60.82	59.43	37.09	446.66	2
LLaVA		35.10	66.68	60.52	31.30	46.04	53.42	*61.94*	54.29	409.29	0
Baselines											
Task Arithmetic	×	36.20 (+1.25)	65.99 (+3.03)	**65.85 (+4.98)**	51.20 (+7.30)	68.21 (+6.40)	**61.92 (+4.80)**	58.82 (-1.87)	35.70 (-9.99)	443.89 (+15.91)	4
Ties-Merging	×	34.00 (-0.95)	57.29 (-5.67)	38.97 (-21.90)	55.00 (+11.10)	59.73 (-2.08)	40.31 (-16.81)	51.97 (-8.72)	24.36 (-21.33)	361.63 (-66.35)	0
DARE-Linear	×	**36.80 (+1.85)**	64.08 (+1.12)	65.07 (+4.20)	47.90 (+4.00)	65.35 (+3.54)	60.96 (+3.84)	58.01 (-2.68)	36.12 (-9.57)	434.29 (+6.31)	2
DARE-Ties	×	33.60 (-1.35)	46.75 (-16.21)	58.41 (-2.46)	26.50 (-17.40)	50.48 (-11.33)	53.15 (-3.97)	49.62 (-11.07)	31.43 (-14.26)	349.94 (-78.04)	0
MetaGPT	✓	34.70 (-0.25)	59.37 (-3.59)	61.29 (+0.42)	**56.40 (+12.50)**	**76.96 (+15.15)**	60.84 (+3.72)	59.44 (-1.25)	36.97 (-8.72)	445.97 (+17.99)	5
Our Method											
AdaMMS	✓	34.90 (-0.05)	**69.09 (+6.13)**	64.12 (+3.25)	55.70 (+11.80)	76.90 (+15.09)	61.11 (+3.99)	**60.12 (-0.57)**	**37.27 (-8.42)**	**459.21 (+31.23)**	7

Results on merging LLaVA-v1.5-7B into CogVLM-chat-7B.

Conclusion

- We introduce AdaMMS for merging heterogeneous MLLMs:
  - **A simple yet effective method** for applying model merging techniques on heterogeneous MLLMs.
  - **Unsupervised hyper-parameter selection method** based on our investigation on parameter space.
  - Our method **outperforms previous baselines** on existing benchmarks.

清華大學
Tsinghua University

Thanks!