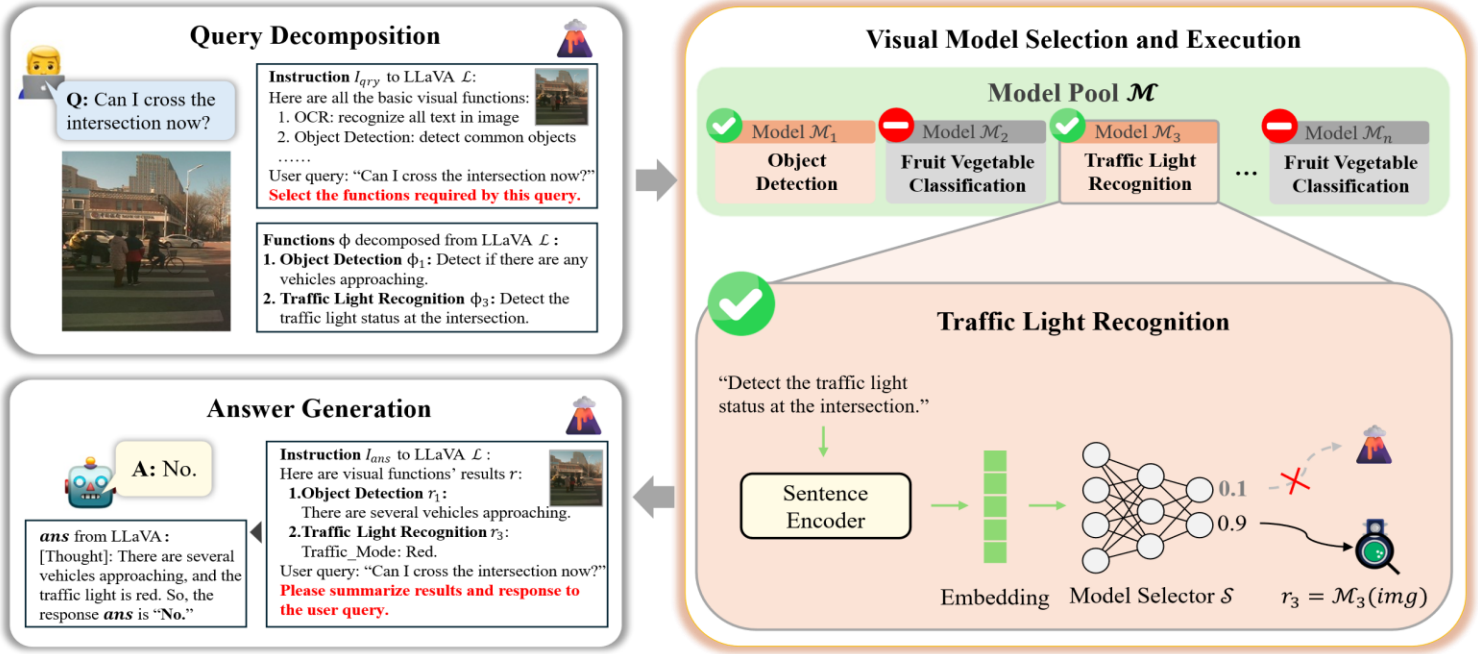
Patching the Visual Ability of Large Multimodal Models by Collaborating with Small Models

Hao LIANG, Xiaolong ZHANG, Meina KAN, Shiguang SHAN, Xilin CHEN

Frontiers of Computer Science, DOI: 10.1007/s11704-025-41126-5

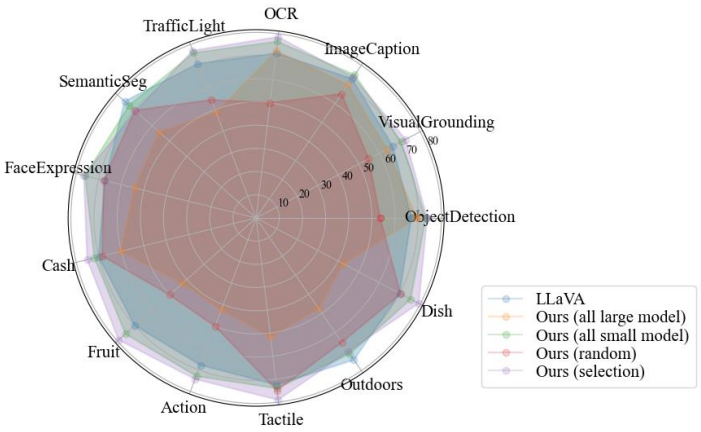
Problems & Ideas

- Problems with Existing LMMs:
  - LMMs struggle with certain visual tasks like object localization, accurate counting, and text recognition.
  - These limitations hinder their use in precision-demanding applications like robotics and assistive technologies.
- Ideas
  - **Collaboration with Smaller Models**: Enhance LMMs by integrating smaller, task-specific models.



Main Contributions

- Contributions:
  - We propose a method that collaborates LMM with smaller models to improve visual perception.
  - Introduce a reinforcement learning-based model selector to dynamically choose between LMM and small models for specific visual tasks.
  - Results show that this collaboration mechanism improves visual performance of LMM.



Success rates of LLaVA, our method, and other model selection strategies across various visual tasks on VizWiz.

	LLaVA-Bench				MM-VET						
	Conv.	Detail	Reas.	All	Rec	OCR	Know	Gen	Spat	Math	All
LLaVA	73.6	66.1	76.3	72.5	30.4	13.3	19.2	20.1	18.7	8.1	24.1
LLaVA-Plus	-	-	-	-	30.5	23.6	**20.5**	**22.5**	**28.5**	7.7	27.5
CLLM	71.4	66.5	**83.9**	74.4	20.3	10.2	6.2	7.4	19.2	0.0	16.7
VisGPT*	61.5	40.9	59.3	54.5	21.2	9.0	6.9	6.1	12.8	3.5	17.1
GPT4Tools*	72.6	42.9	82.4	67.0	20.9	8.6	7.2	7.3	15.1	0.0	16.4
Ours($\mathcal{L}$)	76.3	**72.4**	67.0	71.9	28.2	20.2	15.6	14.9	20.5	7.3	25.8
Ours($\mathcal{M}$)	72.2	67.1	83.3	74.7	30.2	20.5	11.8	14.4	22.9	7.7	26.7
Ours	**77.2**	71.5	83.0	**76.4**	**32.1**	**29.3**	20.1	22.4	27.3	**15.4**	**31.1**

Comparative performance of various LMMs on LLaVA-Bench and MM-VET.