CVPR 2025 HIGHLIGHT

Olympus: A Universal Task Router for Computer Vision Tasks

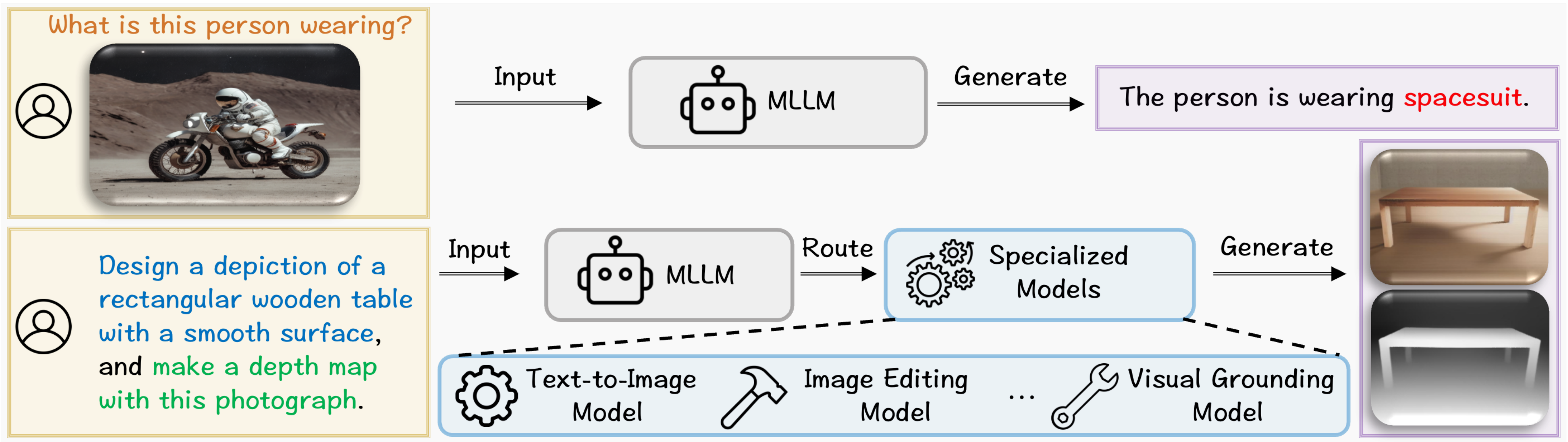
Yuanze Lin[1] Yunsheng Li[2] Dongdong Chen[2] Weijian Xu[2] Ronald Clark[1] Philip H.S. Torr[1]

[1]University of Oxford [2]Microsoft

Source code

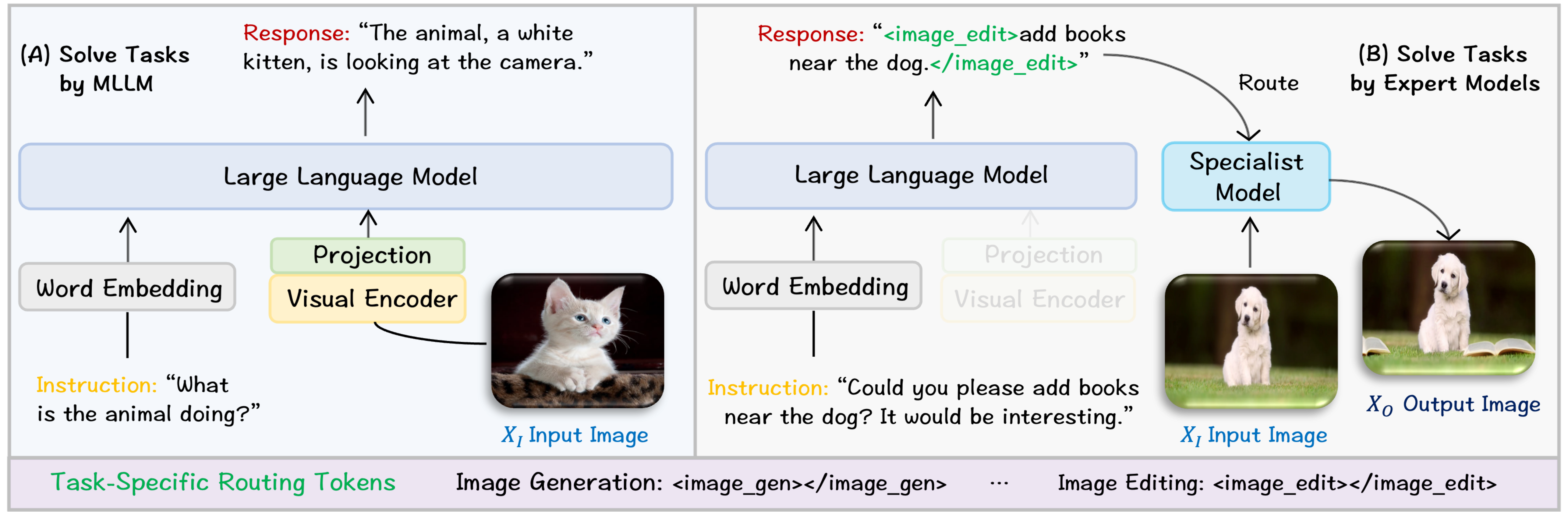
CVPR Nashville JUNE 11-15, 2025

Motivation The MLLM can function as a task router delegating to specialized models!


What is this person wearing?
Input
MLLM
Generate
The person is wearing spacesuit.
Design a depiction of a rectangular wooden table with a smooth surface, and make a depth map with this photograph.
Input
MLLM
Route
Specialized Models
Generate
Text-to-Image Model
Image Editing Model
...
Visual Grounding Model

- Leverage MLLMs to address various tasks via allocating specialized models.
- Develop task-specific routing tokens and enhance MLLMs with chain-of-action capabilities.
- Curate OlympusInstruct (446.3K) & OlympusBench (49.6K) across **20** computer vision tasks.

Method


(A) Solve Tasks by MLLM
Response: "The animal, a white kitten, is looking at the camera."
Large Language Model
Word Embedding
Projection
Visual Encoder
Instruction: "What is the animal doing?"
X_I Input Image
Response: "<image_edit>add books near the dog.</image_edit>"
Route
(B) Solve Tasks by Expert Models
Large Language Model
Specialist Model
Word Embedding
Projection
Visual Encoder
Instruction: "Could you please add books near the dog? It would be interesting."
X_I Input Image
X_O Output Image
Task-Specific Routing Tokens
Image Generation: <image_gen></image_gen>
···
Image Editing: <image_edit></image_edit>

- Solve VQA directly, while allocate specific models for other tasks.
- The MLLM predicts the refined task-specific response together with its routing tokens.
- Train the MLLM via next-token prediction paradigm using $P(Y_a|\mathcal{F}_v,\mathcal{F}_t)=\prod_{i=1}^{L} P_\theta(y_i|\mathcal{F}_v,\mathcal{F}_t,Y_{a,<i})$.

Experiments

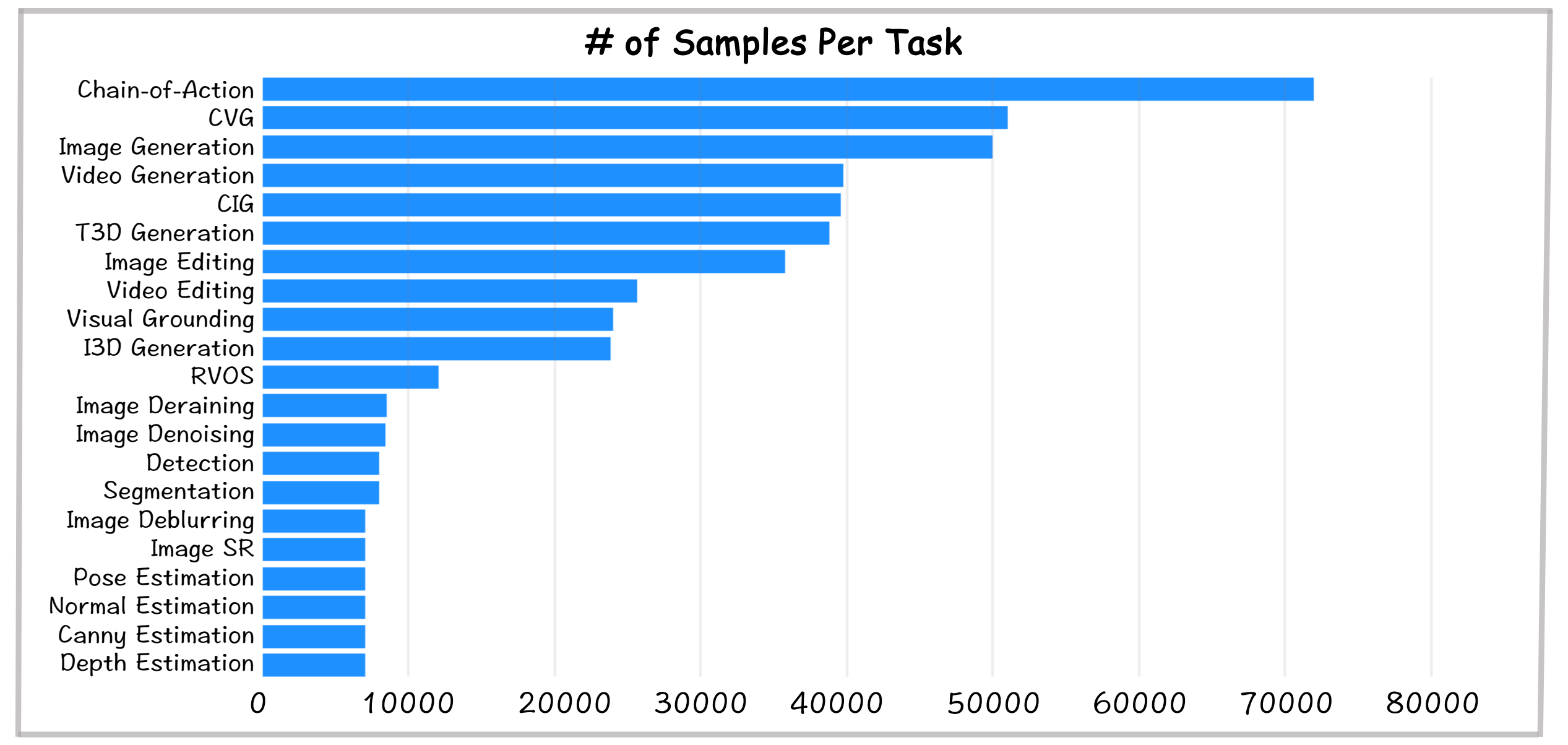
The statistic of the collected dataset

Additional Covered Tasks
Image Generation, Image Editing, Controllable Image Generation (Canny, Pose, Segmentation, Depth, Normal, Scribble), Video Generation, Text-to-3D Generation, Image-to-3D Generation, Image Deblurring, Image Super-Resolution, Image Deraining, Image Denoising, Pose Estimation, Normal Estimation, Canny Estimation, Depth Estimation, Visual Grounding, Object Detection, Object Segmentation, Referring Video Object Segmentation, Controllable Video Generation (Canny, Pose, Segmentation, Depth, Normal, Scribble), Video Editing

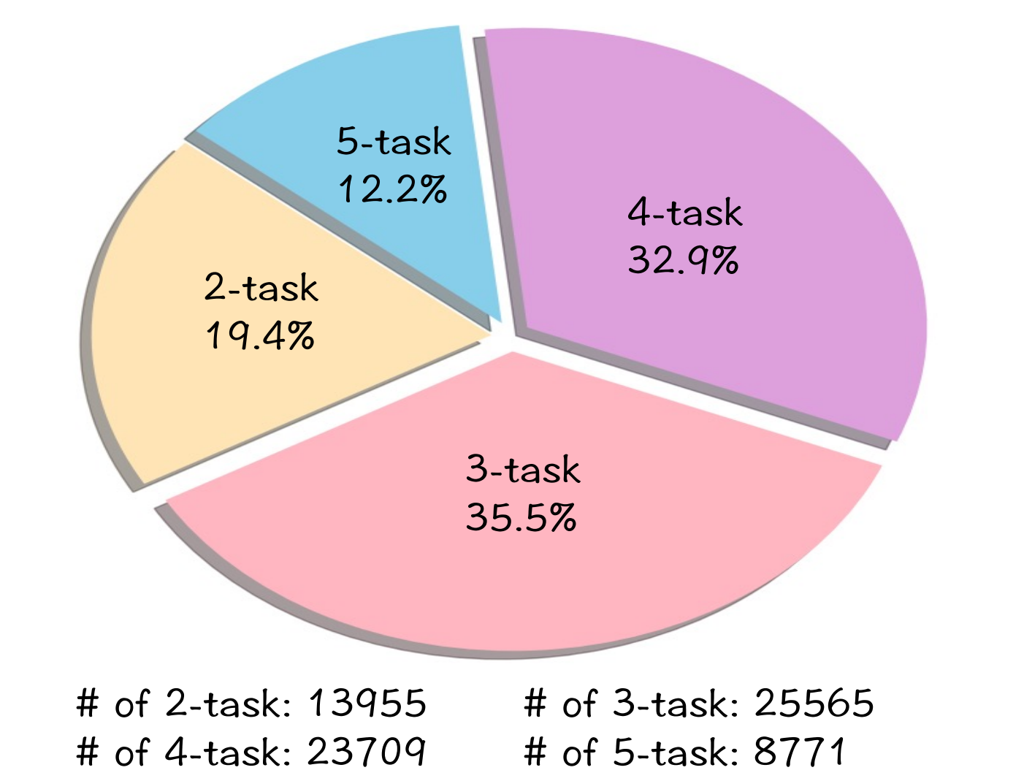
(a) 20 covered tasks in Olympus framework.

Dataset Statistic	
# of Training Instructions (Single Task)	381.5K
# of Training Instructions (Chain-of-Action)	64.8K
# of Evaluation Instructions (Single Task)	42.4K
# of Evaluation Instructions (Chain-of-Action)	7.2K
Max Instruction Word Length	372
Ave Instruction Word Length	20.2
Ave Response Word Length	10.7
Ave # of COA Tasks	3.4

(c) Statistic of the collected dataset.

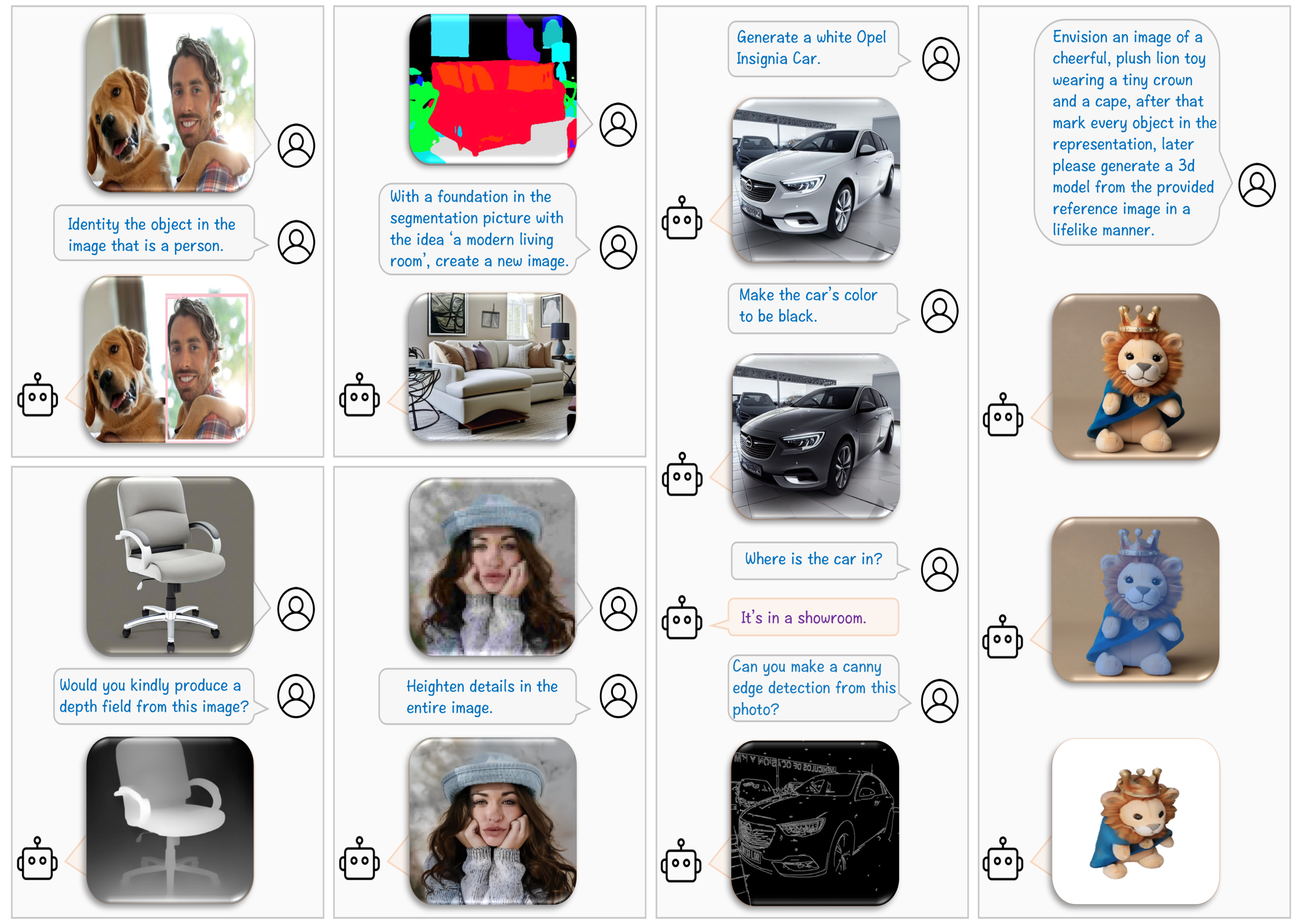


(b) Number of instructions for different tasks.



(d) Distribution of chain-of-action instructions.

Diverse applications of Olympus


Identity the object in the image that is a person.
With a foundation in the segmentation picture with the idea 'a modern living room', create a new image.
Generate a white Opel Insignia Car.
Make the car's color to be black.
Where is the car in?
It's in a showroom.
Can you make a canny edge detection from this photo?
Envision an image of a cheerful, plush lion toy wearing a tiny crown and a cape, after that mark every object in the representation, later please generate a 3d model from the provided reference image in a lifelike manner.
Would you kindly produce a depth field from this image?
Heighten details in the entire image.

Multimodal evaluation across 11 benchmarks

Method	LM	Res.	VQAv2	GQA	VisWiz	SQA^I	VQA^T	MME-P	MME-C	MMB	MM-Vet	POPE	MMMU
Shikra [8]	V-13B	224	77.4	-	-	-	-	-	-	58.8	-	-	-
IDEFICS-9B [33]	L-7B	224	50.9	38.4	35.5	-	25.9	-	-	48.2	-	-	-
IDEFICS-80B [33]	L-65B	224	60.0	45.2	36.0	-	30.9	-	-	54.5	-	-	-
Qwen-VL-Chat [5]	Q-7B	448	78.2	57.5	38.9	68.2	61.5	1487.5	360.7	60.6	-	-	32.9
mPLUG-Owl2 [88]	L-7B	448	79.4	56.1	54.5	68.7	58.2	1450.2	313.2	64.5	36.2	85.8	32.1
LLaVA-1.5 [45]	V-7B	336	78.5	62.0	50.0	66.8	58.2	1510.7	316.1	64.3	30.5	85.9	32.0
MobileVLM-3B [12]	M-2.7B	336	-	59.0	-	61.2	47.5	1288.9	-	59.6	-	84.9	-
MobileVLM-v2-3B [13]	M-2.7B	336	-	61.1	-	70.0	57.5	1440.5	-	63.2	-	84.7	-
LLaVA-Phi [100]	P-2.7B	336	71.4	-	35.9	68.4	48.6	1335.1	-	59.8	28.9	85.0	-
Imp-v1 [67]	P-2.7B	384	79.5	58.6	-	70.0	59.4	1434.0	-	66.5	33.1	88.0	-
MoE-LLaVA-3.6B [39]	P-2.7B	384	79.9	62.6	43.7	70.3	57.0	1431.3	-	68.0	35.9	85.7	-
TinyLLaVA [95]	P-2.7B	384	79.9	62.0	-	69.1	59.1	1464.9	-	66.9	32.0	86.4	-
Bunny-3B [25]	P-2.7B	384	79.8	62.5	-	70.9	-	1488.8	289.3	68.6	-	86.8	33.0
Mipha-3B [99]	P-2.7B	384	81.3	63.9	45.7	70.9	56.6	1488.9	295.0	69.7	32.1	86.7	32.5
Olympus (Ours)	P-2.7B	384	80.5	63.9	48.2	70.7	53.4	1520.7	283.2	71.2	33.8	86.6	32.8

Routing performance (single-task)

Method	ED↓	Pre↑	Recall↑	F1↑
HuggingGPT (GPT-4o mini)	0.45	65.14	48.51	53.14
HuggingGPT (GPT-4o)	0.35	75.03	60.23	61.25
Olympus (Ours)	**0.18**	**91.82**	**92.75**	**91.98**

Routing performance (chain-of-action)

Method	Acc↑	Pre↑	Recall↑	F1↑
HuggingGPT (GPT-4o mini)	70.14	76.51	72.14	75.46
HuggingGPT (GPT-4o)	81.35	85.54	81.55	83.56
Olympus (Ours)	**94.75**	**95.80**	**94.75**	**95.77**

Human evaluation

Method	Success Rate↑
HuggingGPT (GPT-4o mini)	65.8
HuggingGPT (GPT-4o)	75.2
Olympus (Ours)	**86.5**

Ablation of varying tasks on multimodal benchmarks

# of Tasks	VQAv2	GQA	VisWiz	SQA^I	VQA^T	MME-P	MME-C	MMB	MM-Vet	POPE	MMMU
0	81.0	64.0	46.2	70.8	55.3	1498.3	293.2	70.1	32.6	86.6	32.4
5	80.5	64.2	45.6	70.9	53.5	1468.3	310.4	70.5	34.9	86.5	32.5
10	80.4	64.1	46.1	71.2	53.0	1546.7	333.9	70.2	33.8	86.2	32.9
20	80.5	63.9	48.2	70.7	53.4	1520.7	283.2	71.2	33.8	86.6	32.8

Ablation of varying tasks for single-task routing

# of tasks	Acc↑	Pre↑	Recall↑	F1↑
5	96.38	96.36	96.45	97.61
10	96.15	95.85	96.23	97.07
15	95.84	95.78	95.84	96.79
20	94.75	95.80	94.75	95.77

Ablation of varying tasks for chain-of-action routing

# of tasks	ED↓	Pre↑	Recall↑	F1↑
5	0.12	93.23	94.32	93.35
10	0.14	92.23	93.45	92.28
15	0.17	91.97	92.89	92.01
20	0.18	91.82	92.75	91.98