Dynamic Cross-Modal Prompt Generation for Multimodal Continual Instruction Tuning

Tao Hu, Da-Wei Zhou

Nanjing University hut@lamda.nju.edu.cn

NEURAL INFORMATION PROCESSING SYSTEMS

Highlights

- DRAPE synthesizes instance-specific soft prompts for each query-image pair in rehearsal-free MCIT.
- A cross-modal prompt generator summarizes textual instructions and cross-attends to visual patch features.
- Null-space gradient projection preserves the shared visual projector and reduces forgetting.
- DRAPE achieves state-of-the-art performance on CoIN and UCIT with a lightweight, rehearsal-free design.

Multimodal Continual Instruction Tuning



Instruction: What is the capital of Washington? A. Santa Fe B. Lansing C. Olympia D. Seattle. Answer with the option's letter from the given choice.

Answer: C



Instruction: What is the object in the image? Answer the question using a single word or phrase.

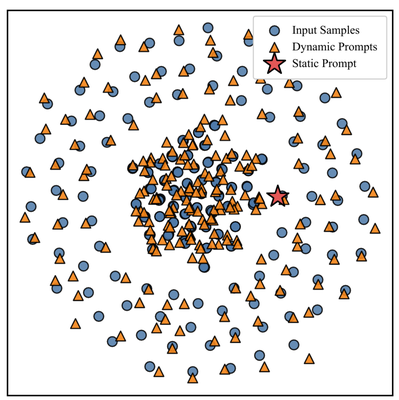
Answer: Castle



Instruction: Please provide the bounding box coordinate of the region this sentence describes: person on right.

Answer: [0.67, 0.32, 0.85, 0.77]

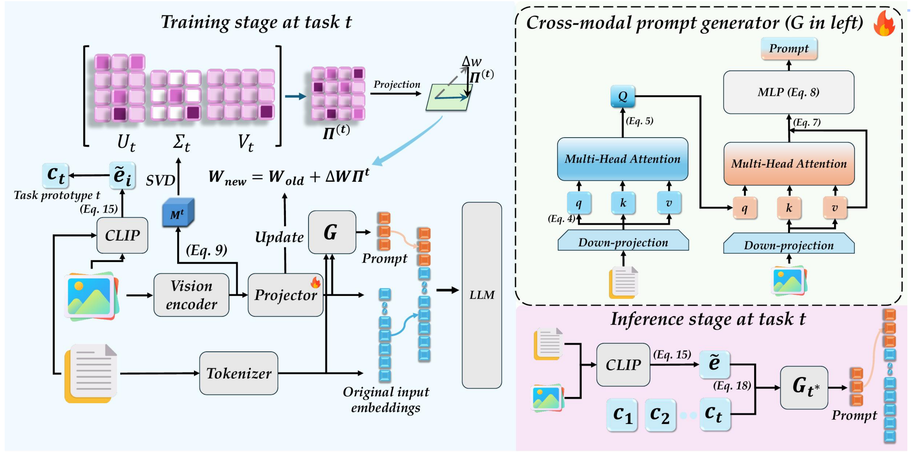
Motivation



Limitation: Existing MCIT methods mainly select or combine pre-learned task-level prompts or LoRA experts. However, instances within the same task may require different visual evidence and reasoning capabilities.

Our insight: Instead of only reusing fixed modules, synthesize a soft prompt conditioned on each image-instruction pair.

DRAPE: Dynamic Cross-Modal Prompt Generation



Key Ideas

Instance-Specific Prompt Generation:
A task-specific generator derives prompt queries from the instruction, then cross-attends to visual patch features to synthesize a unique soft prompt for each image–instruction pair. The prompt is prepended to the frozen LLM input.

Preserve Knowledge via Null-Space Projection:
The shared visual projector is updated across tasks. We accumulate previous-task feature statistics, estimate an approximate null space via SVD, and project projector gradients onto it to reduce interference with earlier mappings.

$$\nabla W' = \nabla W \mathbf{\Pi}^{(t)} = \nabla W V_\perp V_\perp^\top,$$
$$\Delta \mathbf{o}^{\text{old}} = -\eta \nabla W' \tilde{\mathbf{v}}^{\text{old}} = -\eta \nabla W V_\perp \left(V_\perp^\top \tilde{\mathbf{v}}^{\text{old}}\right) \approx \mathbf{0}$$

Route Generators Without Task Labels:
Each learned task is represented by a compact prototype from CLIP image and instruction embeddings. At inference, cosine similarity selects the most relevant generator, without requiring task identities.

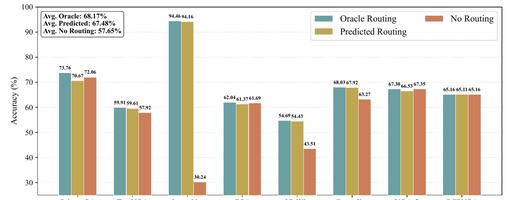
$$t^* = \arg\max_{t \in \{1,\ldots,T\}} \cos(\tilde{\mathbf{e}}, \mathbf{c}_t)$$

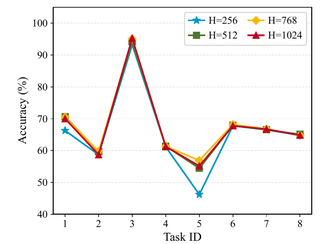
Setting: Multimodal instruction-tuning tasks arrive sequentially. Only current-task data is available during training, while all seen tasks are evaluated without task labels.

State-Of-The-Art Results

Method	ScienceQA	TextVQA	ImageNet	GQA	VizWiz	Grounding	VQAv2	OCR-VQA	Average
Finetune	26.00	25.38	28.51	33.07	26.52	0.10	40.00	52.92	29.06
CODA-Prompt	58.15	50.16	24.04	54.33	48.94	17.83	55.86	54.42	45.46
DualPrompt	56.40	47.12	34.96	42.03	44.14	12.01	54.43	53.36	43.05
L2P	54.42	46.04	30.36	57.09	42.19	9.38	50.45	54.03	42.99
MoELoRA	47.34	32.91	38.73	37.15	42.48	0.97	42.77	57.50	37.48
Continual LLaVA	58.67	49.99	57.66	**62.53**	42.32	16.25	64.33	**74.91**	53.33
ModalPrompt	68.42	<u>56.40</u>	41.13	61.11	50.13	<u>36.69</u>	**66.90**	59.68	55.06
ProgLoRA	**74.84**	51.83	<u>83.90</u>	49.93	<u>53.87</u>	31.19	62.71	64.44	<u>59.09</u>
DRAPE (Ours)	<u>70.67</u>	**59.61**	**94.16**	<u>61.37</u>	**54.43**	**67.92**	<u>66.53</u>	<u>65.11</u>	**67.48**

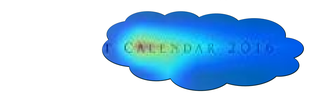
Best in bold, second best underlined







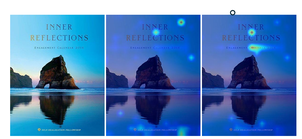
Q: Which year's calendar is this? A: 2016



Q: What type of book is this? A: Calendars



Original Baseline **DRAPE**



Original Baseline **DRAPE**

Conclusion

DRAPE brings instance-specific prompt generation to rehearsal-free MCIT, achieving strong continual instruction tuning on CoIN and UCIT.

Wechat ***Toolbox*** ***Code***