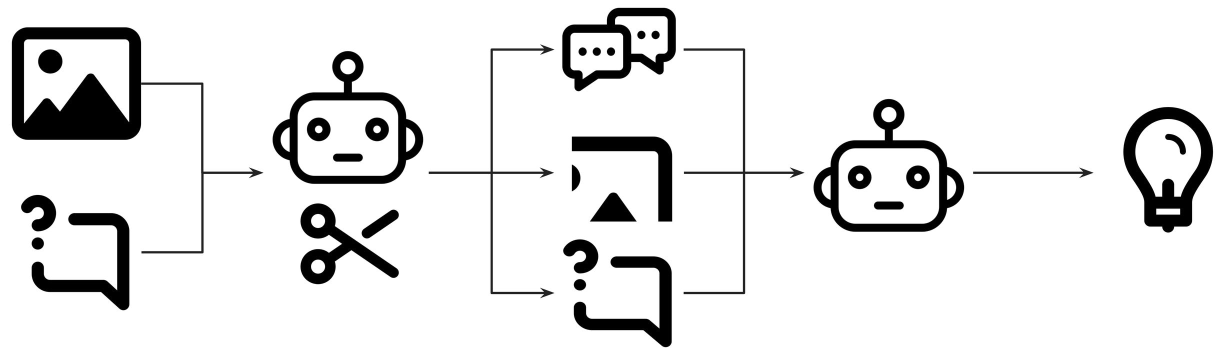
Orchestrating Vision-Language Reasoning: A Dynamic Routing Framework for Multi-Agent Visual Question Understanding

Szu-Ju Chen[1], Yu-Neng Chuang[2], Xia Hu[2]
[1]Department of Computer Science and Information Engineering, National Taiwan University [2]Department of Computer Science, Rice University

Vision-Language Reasoning and Chain-of-Thought

`Chain-of-Spot` enhances visual reasoning performance by cropping images to focus on key areas

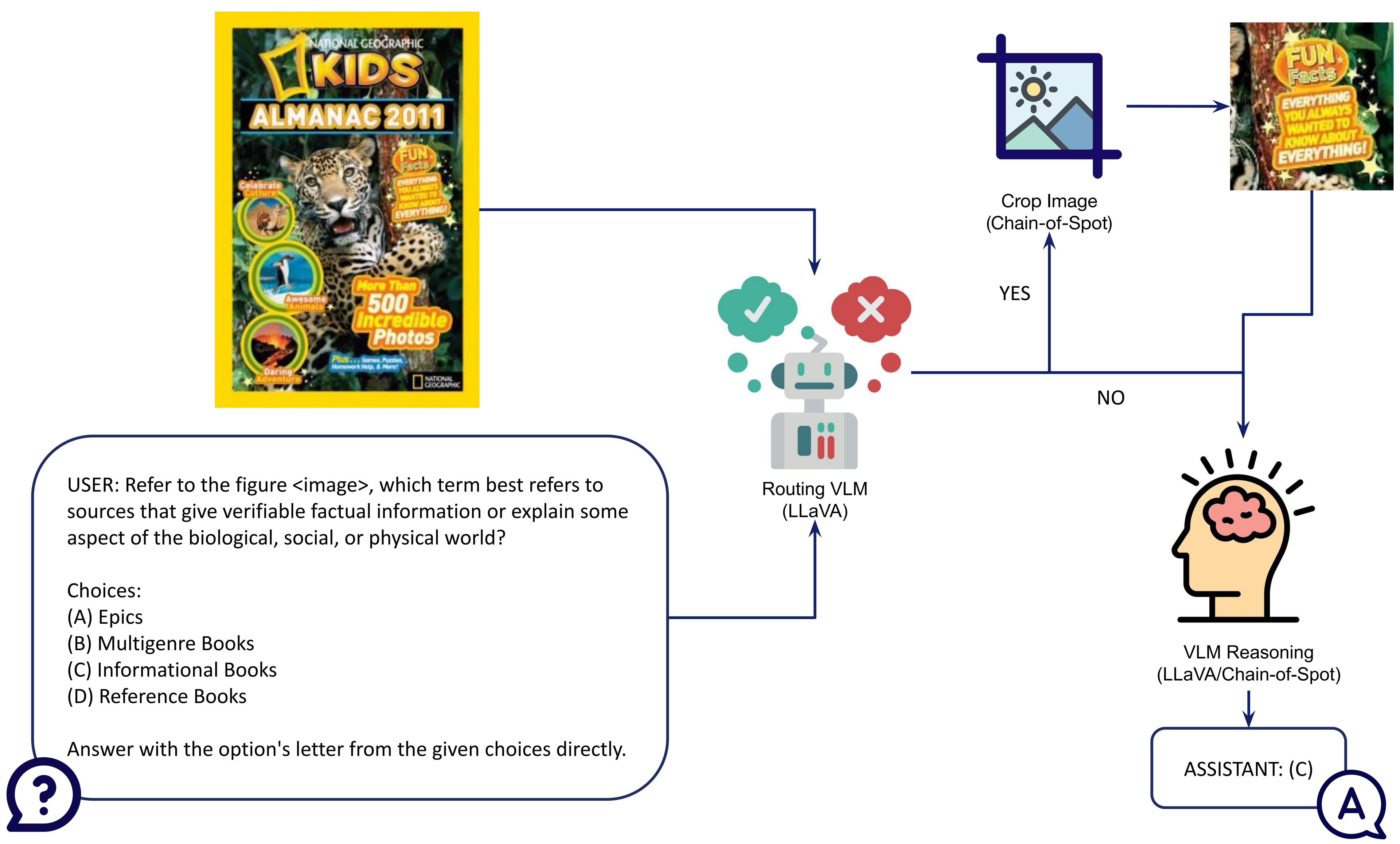


Challenges

- Longer Processing: concatenated image embeddings and extended conversation history
- Unnecessary Cropping: applied even when the input image already contains enough information
- Answer Distortion: misinterpretation caused by incorrect cropping results

Framework

- **Introduce a routing agent** before the cropping and reasoning stages
- Query the routing agent `LLaVA` to determine if the image is clear and suitable for answering the question
  - YES: skip cropping process and proceed directly to the reasoning stage
  - NO: use the cropping agent `Chain-of-Spot` to identify the region of interest and crop the image
- Feed the updated image, user question, and cropping history into reasoning agent `LLaVA/Chain-of-Spot` to enhance the accuracy and informativeness of the final answer



Dataset

Massive Multi-discipline Multimodal Understanding and Reasoning (MMMU) benchmark

- college-level problems across six disciplines and 30 college subjects
- 900 validation questions: using 857 single-image questions as experiment dataset, including 805 multiple choice questions and 52 open questions

Experiments

Routing Prompts

USER: <image>
You are a supervisor. Answer "YES" if the given image needs cropping to get accurate answer for the question: "{question}" Otherwise, answer "NO".

USER: <image>
You are a supervisor. Answer "YES" if the given image contains too many noises and information for the question: "{question}" Otherwise, answer "NO".

Routing Conditions

- Directly Decode: Check whether the generated output from VLM models is YES
- Token Threshold: Compare the probability of the `"YES"` token against a predefined threshold
- Yes/No Token Odds: Compute the ratio of the `"YES"` token probability to the `"NO"` token probability

Share image embedding

Save the original image embedding when the routing condition is not met

	Prompt 1			Prompt 2		
	Route Rate	Time	Accuracy	Route Rate	Time	Accuracy
LLaVA	0%	0.2386	34.54%	0%	0.2386	34.54%
Chain-of-Thought(Llava)	100%	1.4254	31.51%	100%	1.4254	31.51%
Chain-of-Thought(CoS)	100%	1.6299	36.64%	100%	1.6299	36.64%
Route (decode)	95%	1.7017 (0.9x)	36.52%	73%	1.4476 (1.1x)	36.52%
Route (threshold=0.1)	99%	1.7572 (0.9x)	36.64%	97%	1.7268 (0.9x)	36.52%
Route (threshold=0.15)	75%	1.4714 (1.1x)	36.87%	37%	1.0414 (1.5x)	**36.76%**
Route (threshold=0.2)	14%	0.7833 (**2.1x**)	35.71%	3%	0.6590 (**2.4x**)	35.24%
Route (odd)	53%	1.2306 (1.3x)	**37.11%**	19%	0.8310 (1.9x)	35.94%
Route	53%	1.2865 (1.2x)	37.11%	19%	0.8869 (1.8x)	35.94%
Route (share embedding)	53%	1.2306 (**1.3x**)	37.11%	19%	0.8310 (**1.9x**)	35.94%

Results

- Achieve approximately **2x** improvement in processing speed
- Sharing image embeddings skips repeated processes and reduces operation time
- **Maintain or exceed the accuracy** performance compared to the Chain-of-Spot baselines
- Utilizing generated token odds ensures high-quality routing performance and reduces significant time
- Dynamic routing mechanism works effectively with both reasoning agents, `LLaVA` and `Chain-of-Spot`
- Directly applicable to the inference process without requiring additional training