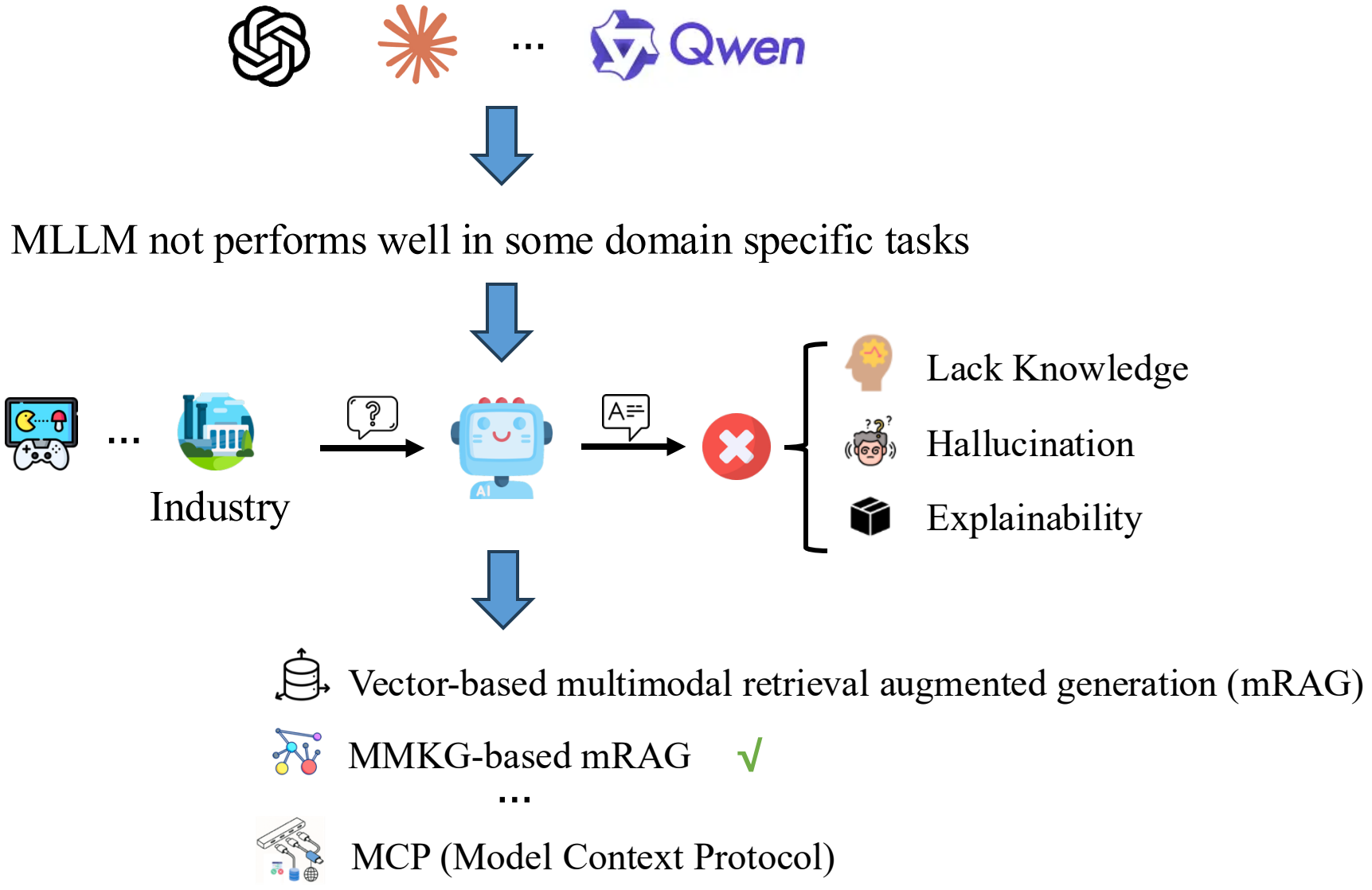
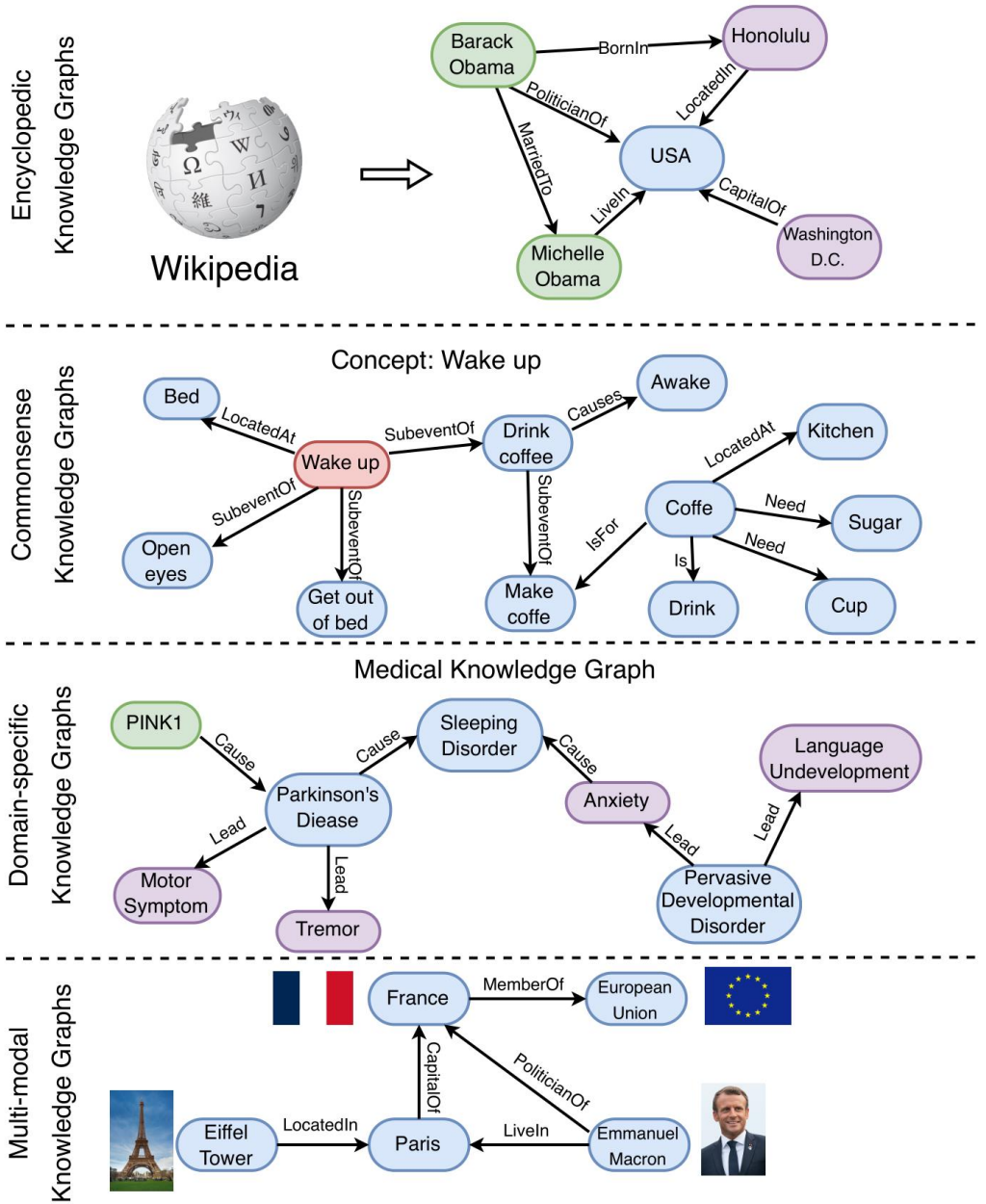
Taming the Untamed: Graph-Based Knowledge Retrieval and Reasoning for MLLMs to Conquer the Unknown

Bowen Wang, Zhouqiang Jiang, Yasuaki Susumu, Shotaro Miwa, Tianwei Chen, Yuta Nakashima

Multimodal Knowledge Graph (MMKG) Based Reasoning

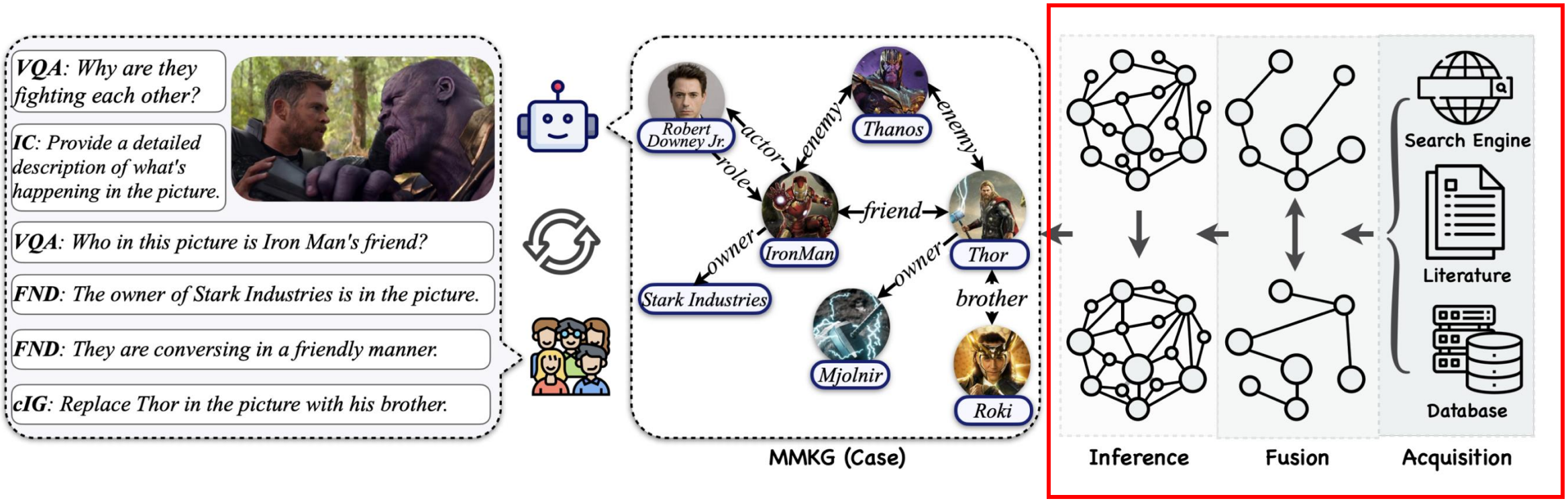


Knowledge Graph



Multimodal Knowledge Graph

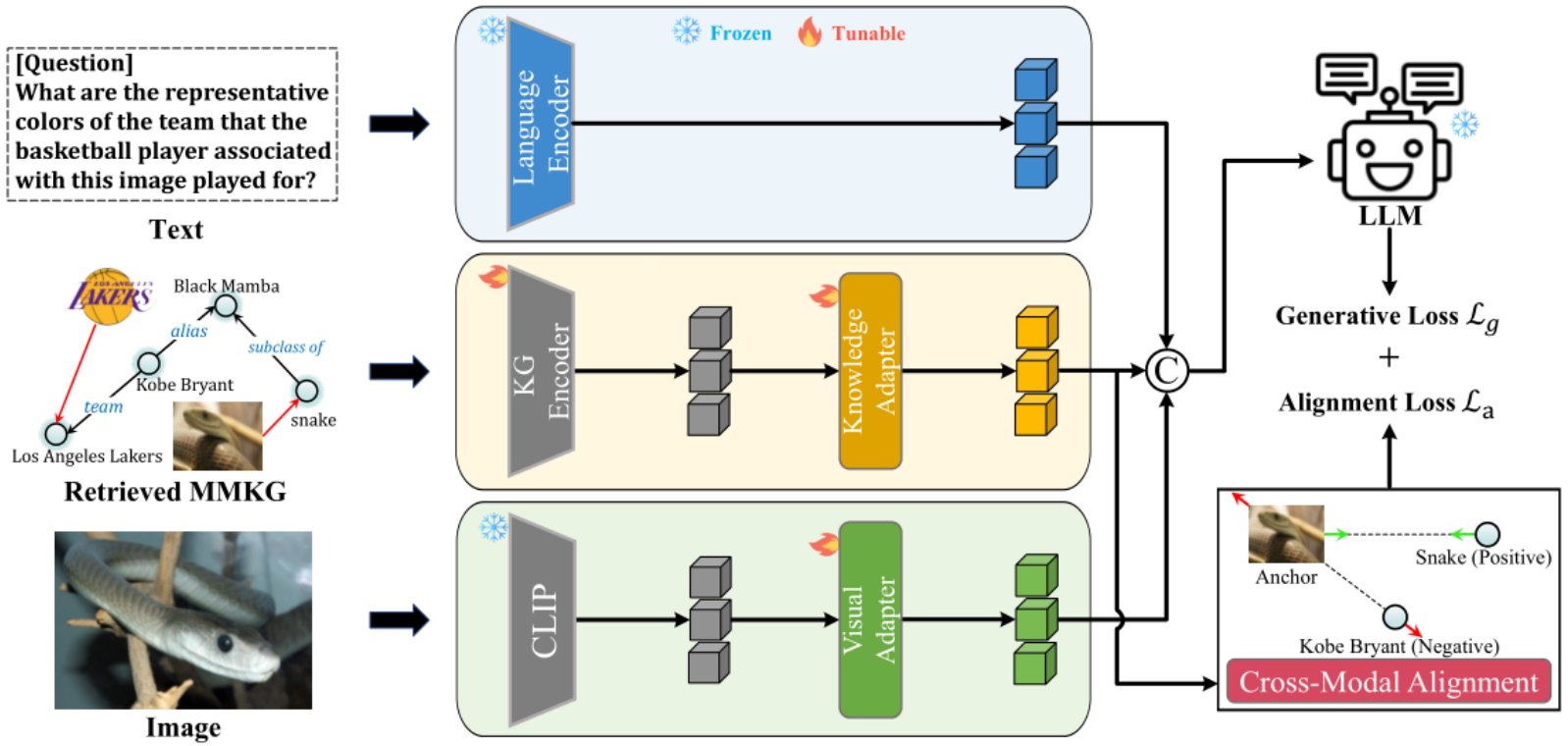
Multimodal Knowledge Graph (MMKG)



A-MMKG (Attribution)

Chen et al. Knowledge graphs meet multi- modal learning: A comprehensive survey, arXiv 2024.

MMKG-Based RAG

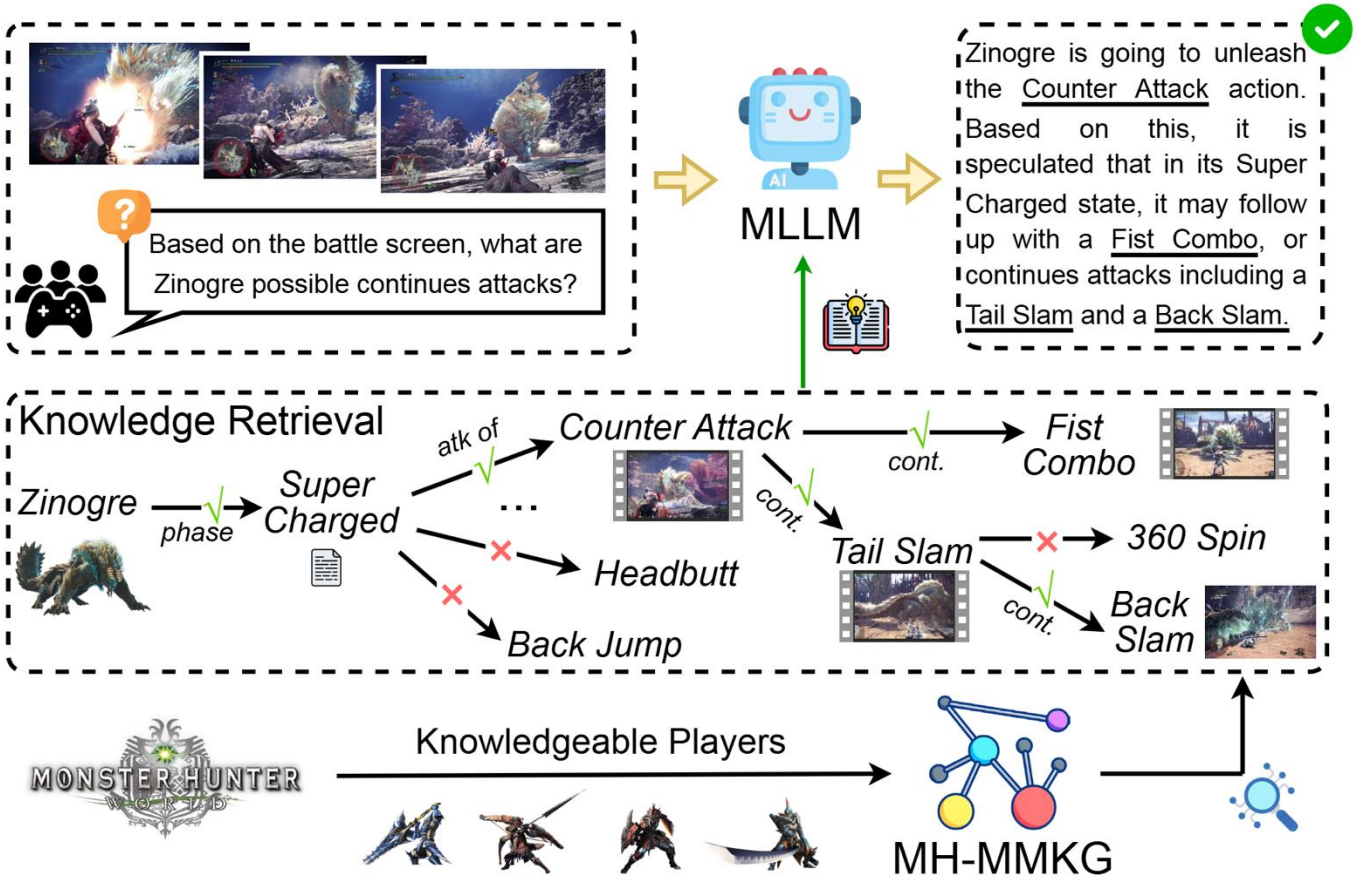


Limitations of existing methods

1. The existing MMKG is built based on Internet data.
2. Training a graph retriever requires a lot of data.
3. Difficulty accessing powerful closed-source models.

Lee et al. Multimodal Reasoning with Multimodal Knowledge Graph, ACL 2024.

MMKG-Based Knowledge Retrieval and Reasoning



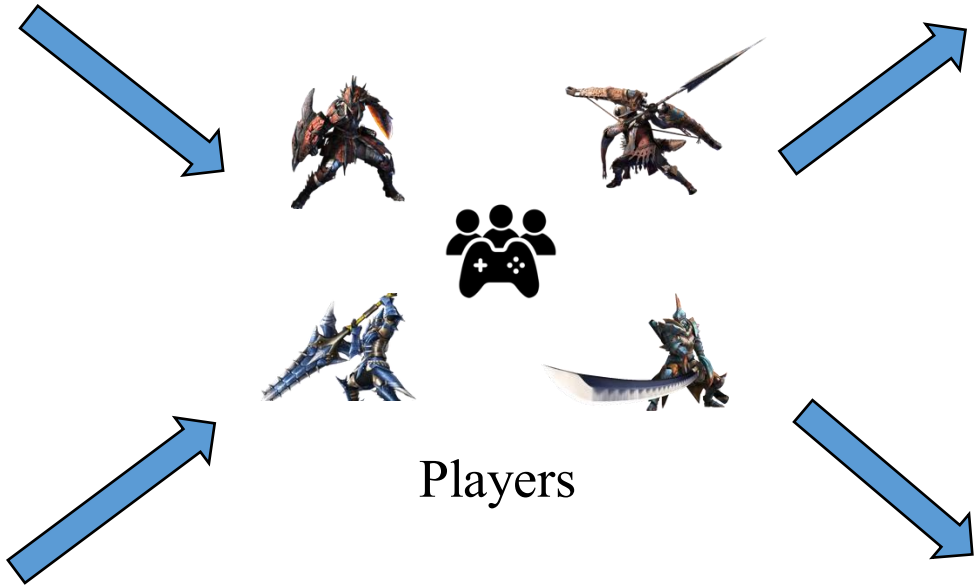
Main Contributions

- Construction of Domain-Specific MH-MMKG and Benchmarks (for uncommon scenarios).
- Training-free Multimodal Knowledge Graph Retrieval Augmentation via Multi-Agents (enabling autonomous reasoning and compatibility with various models).

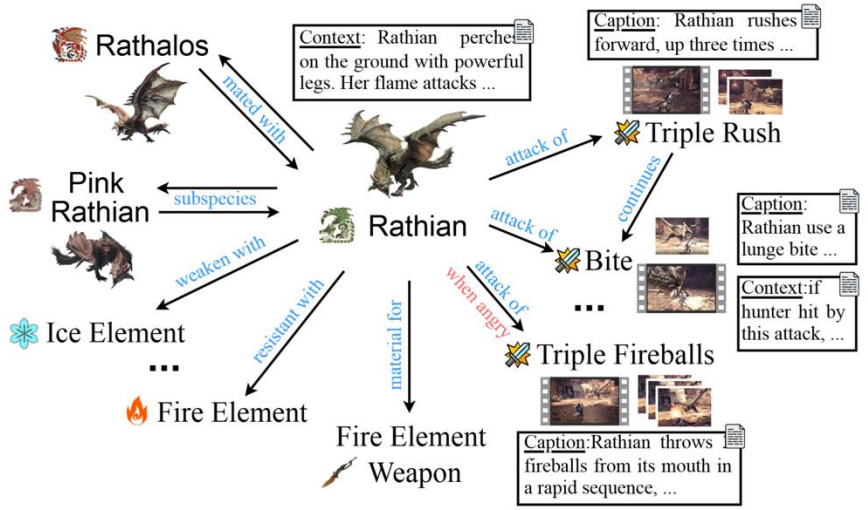
MH-Benchmark

Monster Hunter (MH)

Game Video Recording （4K 60FPS）



MH-MMKG



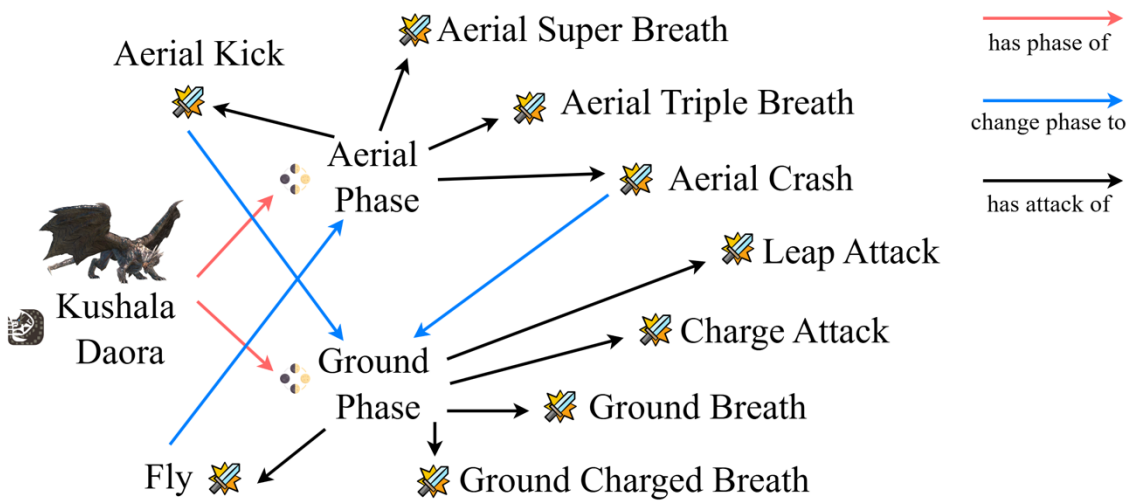
Attributes: text, image, video, keyframe
Edges: node relations and achievement conditions

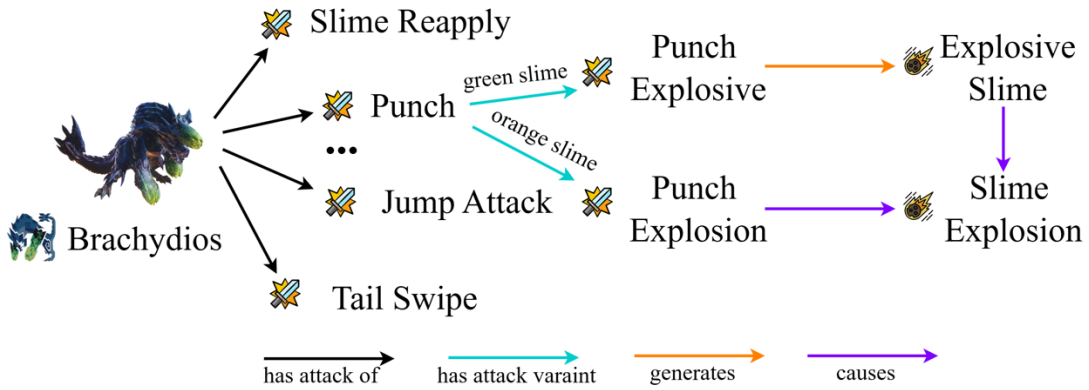
MH-Benchmark (238)

Sub-task	Description	# samples
I: Individual Information	Retrieve the textual information of a monster, e.g., nick name, habitat, and skill mechanics.	24
II: Attack Recognition	Recognize the about to, ongoing or finished attack action of a monster.	109
III: Combo Premonition	Predict upcoming attack sequences based on the monster's current action or previous action track.	28
IV: Condition Awareness	Detect status on monsters or surrounding environment, e.g., whether angry, terrain or phase changes, to anticipate future battle.	29
V: Proc Effect Insight	Analyze the effects such as environment and attack on monster status and movement patterns.	35
VI: Cross Monster Analysis	Compare attack patterns or behaviors across different monsters to optimize hunting strategies.	13

MH-Benchmark的构建

MH-MMKG consists of 22 subgraphs, 7 types of nodes, and 158 types of edges.





Node Type

Type	Description	# number
Topic Entity	Names of monsters that can serve as root entities for knowledge retrieval. Each entity is accompanied by an image of monster as its attribute.	22
Attack Action	Possible attack movements of a monster, each accompanied by text, images (key frames for video), or a video as its attribute. Each of them also attached with human written-caption for the video.	265
Attack Phase	In different phases, a monster will have varying attack patterns, damage, combos, and other attributes. Only some monsters have unique phase settings. Textual context is attached as attribution.	20
Element	The element indicates a monster's weakened resistance to a specific type of attack.	9
Weapon	Types of damage for weapons crafted from monster materials.	10
Props	Various types of game props for interacting with monsters.	6
Attack Effects	The effects of monster attacks or skills during battle, including generated ice patches on ground, scratches, and explosions. Textual context is attached as attribution.	9

Visual Capability Verification

The Zinogre is positioned in a crouched stance, with its body low to the ground. Its forelimbs are spread wide, and its claws are firmly planted on the terrain. The monster's spiked fur is glowing with a bright, electrified aura, indicating an active charge state. Its head is slightly tilted downward, and its mouth is open, revealing sharp teeth, as if growling or preparing to attack. The tail is raised slightly, and the spikes along its back are prominently flared. The surrounding area features coral-like structures and rocky terrain, with a clear blue sky above. A hunter is crouched nearby, seemingly evading or preparing for the Zinogre's next move.

Zinogre crouches on the ground, its massive forelimbs firmly planted on the surface, as if it has just pounced on a hunter (the hunter can be seen lying face-down in front of Zinogre). Its sharp claws are exposed. Zinogre's spine crackles with electricity, glowing brilliantly, while its back and shoulders are charged with electric energy, indicating it is in its charged phase. Its head is lowered and facing forward, mouth wide open, revealing sharp fangs, seemingly ready to eat the hunter.

Visual Capability Verification






In the first frame, Zinogre is facing the hunter, with its body low to the ground and its forelimbs spread apart. Its fur and spikes are glowing with electricity, and sparks are visible around its body. The surrounding area is illuminated by pink coral-like structures and rocky terrain.

In the second frame, Zinogre's head is slightly raised, and its body appears to be shifting forward. The glowing electricity around its body intensifies, with more sparks visible.

In the third frame, Zinogre's right forelimb is raised, and its body is twisting slightly to the left. The electricity around its body continues to glow brightly, and its spikes are illuminated.

In the fourth frame, Zinogre is crouched low to the ground, with its body coiled and its head turned slightly to the left. The glowing electricity remains prominent, and the sparks around its body are more concentrated.

In the fifth frame, Zinogre's body is arched, and its head is lowered. The electricity around its body is surging, with bright flashes visible near its forelimbs.

In the sixth frame, Zinogre is in a similar position, with its body glowing intensely. A bright flash of electricity is visible in front of its body, indicating a surge of energy.

Zinogre raises its right claw, swiftly moves it to the left side of its body, and presses it firmly against the ground. Its head lowers slightly, as if charging up energy. Then, with sudden force, the right claw sweeps across the ground in a wide arc, generating a surge of lightning along its path. Simultaneously, Zinogre jerks its head upward.

Experimental Setup

Methods	Query		MMKG			
	Vision	H-Cap.	Path	H-Cap.	Vis.-Off	Vis.-On
Vanilla	✔					
Vanilla^{+}		✔				
Know.		✔	✔	✔		
Perceptive		✔		✔		
Unaided-Offline	✔				✔	
Unaided-Online	✔					✔

Vanilla: Relies solely on the model's own visual capability (Vision), without additional knowledge.

Vanilla+: Query uses human-annotated visual descriptions (H-Cap.), without additional knowledge.

Know.: Query uses H-Cap. and directly provides the relevant knowledge paths (Path) from the graph along with their corresponding H-Cap.

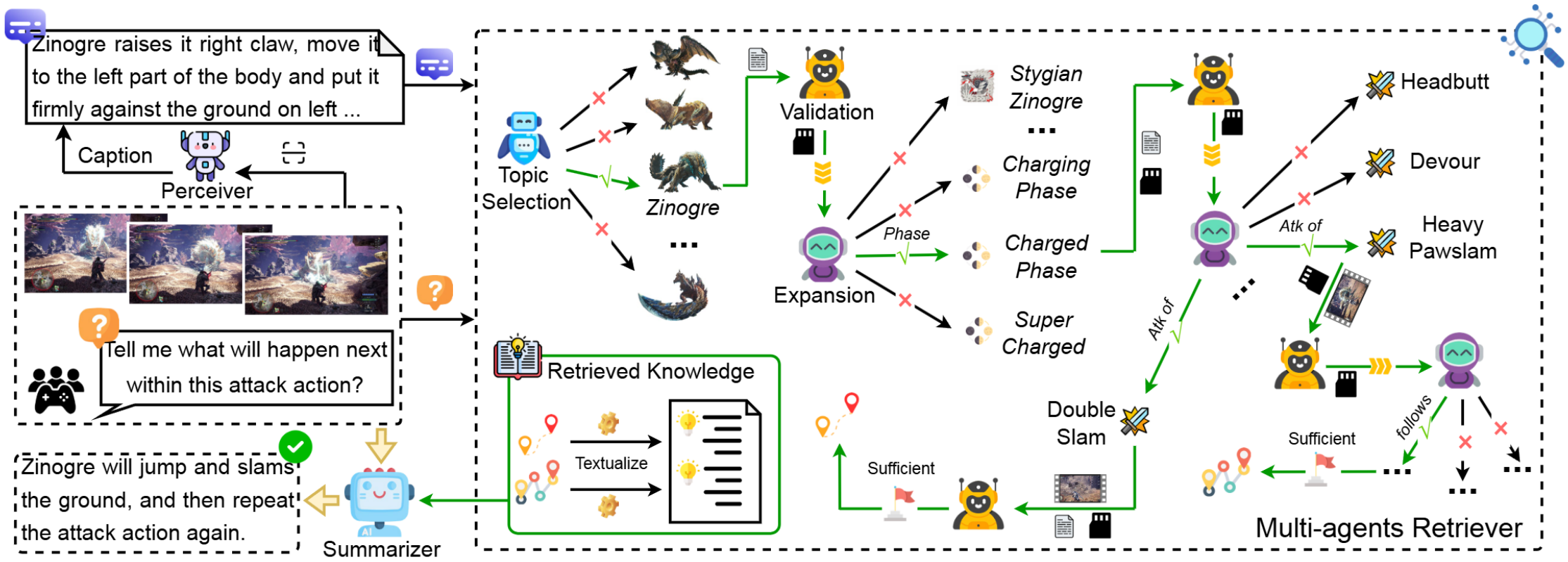
Perceptive: Query uses H-Cap.; the model retrieves knowledge by itself and provides the graph's H-Cap.

Unaided-Offline: Relies on its own Vision, generates visual descriptions offline, and retrieves knowledge by itself.

Unaided-Online: Relies on its own Vision, retrieves knowledge by itself, and analyzes visual information online.

Multi-Agents Retriever

For example, under the Unaided-Online setting, using the Top-5 knowledge.



Perceiver: Generates textual descriptions of the input.

Topic Selection: Selecting the initial nodes for retrieval.

Validation: Assessing the sufficiency of knowledge.

Expansion: Plan the next retrieval node.

Summarizer: Retrieval-augmented.

Experimental Results

The same model serves as all agents, and GPT-4o evaluates whether the answers are consistent (accuracy).

Models Accuracy	Vanilla	Vanilla+	Know.	Perceptive			Unaided-Offline			Unaided-Online		
				Acc.	*Pre.*	*Rec.*	*Acc.*	*Pre.*	*Rec.*	*Acc.*	*Pre.*	*Rec.*
GPT-4o [1]	.3122	.3924	.8565	**.7383**	.5061	**.7046**	.4050	.2595	.4416	**.5105**	.2756	**.5625**
GPT-4o mini [1]	**.3218**	**.4135**	.8481	.6877	.2963	.5450	**.4514**	.2059	**.5028**	.3544	.1626	.3009
Claude 3.7 Sonnet [2]	.2827	.3375	**.8987**	.7004	.5817	.6322	.3628	.2775	.3270	.4388	**.2911**	.4029
Claude 3.5 Sonnet [2]	.2869	.3755	.8776	.7215	**.5922**	.6800	.3966	**.3215**	.4008	.3966	.2330	.3270
Claude 3.5 Haiku [2]	.2356	.3206	.8823	.6455	.3739	.5007	.3544	.2002	.3164	.3670	.1735	.3361
Gemini 2.0 Flash [47]	.1983	.2995	.8438	.6919	.3507	.6146	.3839	.1703	.4092	.3713	.1515	.3663
Gemini 1.5 Pro [47]	.2194	.2700	.8438	.6962	.4761	.6033	.3164	.1615	.2194	.4050	.2122	.4585
Step-1o [43]	.2436	.2815	.8235	.5747	.4372	.5095	.3025	.1831	.2483	.3403	.2204	.2987
InternVL2.5-78B-MPO [12]	.1603	.2616	.8649	.5991	.4198	.5428	.2700	.1729	.2250	.3080	.1556	.2378
Qwen2.5-VL-72B [3]	.1476	.2616	.8734	.6244	.4602	.4908	.3206	.2139	.2383	.3164	.1615	.1814
Ovis2-16B [36]	.1645	.2573	.8902	.7046	.5407	.5949	.2869	.1963	.2383	.3459	.1853	.2878
MiniCPM-o-2.6 [53]	.1139	.2405	.8312	.4683	.3183	.4001	.2194	.1311	.2376	.1687	.0189	.0210
DeepSeek-VL2-Small [52]	.1139	.2362	.6455	.3586	.1419	.1708	.1814	.0400	.0759	.1181	.0042	.0042
Human (Knowledgeable)	.5252	—	—	—	—	—	—	—	—	.9033	.9207	.8535
Human (Random)	.0336	—	—	—	—	—	—	—	—	.6092	.7113	.6457

A lack of visual ability can cause the model to give incorrect answers, but the impact of insufficient knowledge is even more significant.

Experimental Results

Models	Vanilla	Vanilla^{+}	Know.	Perceptive			Unaided-Offline			Unaided-Online		
				Acc.	*Pre.*	*Rec.*	*Acc.*	*Pre.*	*Rec.*	*Acc.*	*Pre.*	*Rec.*
GPT-4o [1]	.3122	.3924	.8565	**.7383**	.5061	**.7046**	.4050	.2595	.4416	**.5105**	.2756	**.5625**
GPT-4o mini [1]	**.3218**	**.4135**	.8481	.6877	.2963	.5450	**.4514**	.2059	**.5028**	.3544	.1626	.3009
Claude 3.7 Sonnet [2]	.2827	.3375	**.8987**	.7004	.5817	.6322	.3628	.2775	.3270	.4388	**.2911**	.4029
Claude 3.5 Sonnet [2]	.2869	.3755	.8776	.7215	**.5922**	.6800	.3966	**.3215**	.4008	.3966	.2330	.3270
Claude 3.5 Haiku [2]	.2356	.3206	.8823	.6455	.3739	.5007	.3544	.2002	.3164	.3670	.1735	.3361
Gemini 2.0 Flash [47]	.1983	.2995	.8438	.6919	.3507	.6146	.3839	.1703	.4092	.3713	.1515	.3663
Gemini 1.5 Pro [47]	.2194	.2700	.8438	.6962	.4761	.6033	.3164	.1615	.2194	.4050	.2122	.4585
Step-1o [43]	.2436	.2815	.8235	.5747	.4372	.5095	.3025	.1831	.2483	.3403	.2204	.2987
InternVL2.5-78B-MPO [12]	.1603	.2616	.8649	.5991	.4198	.5428	.2700	.1729	.2250	.3080	.1556	.2378
Qwen2.5-VL-72B [3]	.1476	.2616	.8734	.6244	.4602	.4908	.3206	.2139	.2383	.3164	.1615	.1814
Ovis2-16B [36]	.1645	.2573	.8902	.7046	.5407	.5949	.2869	.1963	.2383	.3459	.1853	.2878
MiniCPM-o-2.6 [53]	.1139	.2405	.8312	.4683	.3183	.4001	.2194	.1311	.2376	.1687	.0189	.0210
DeepSeek-VL2-Small [52]	.1139	.2362	.6455	.3586	.1419	.1708	.1814	.0400	.0759	.1181	.0042	.0042
Human (Knowledgeable)	.5252	—	—	—	—	—	—	—	—	.9033	.9207	.8535
Human (Random)	.0336	—	—	—	—	—	—	—	—	.6092	.7113	.6457

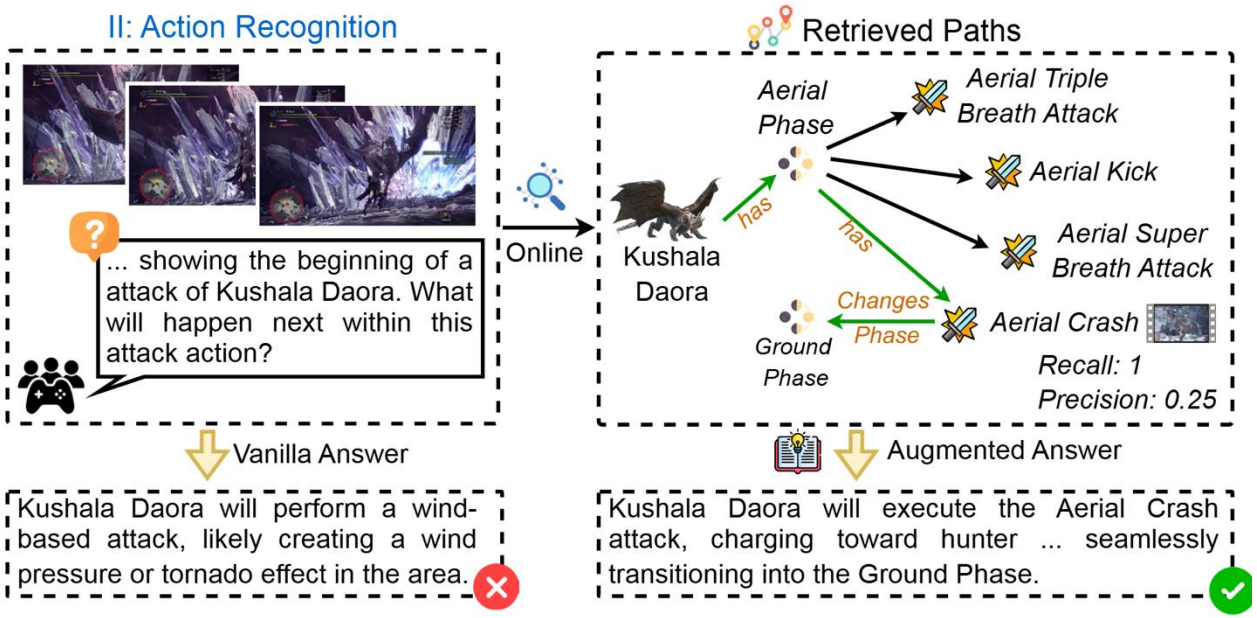
In plain-text settings, most models tend to exhibit strong retrieval-augmented capabilities.

Experimental Results

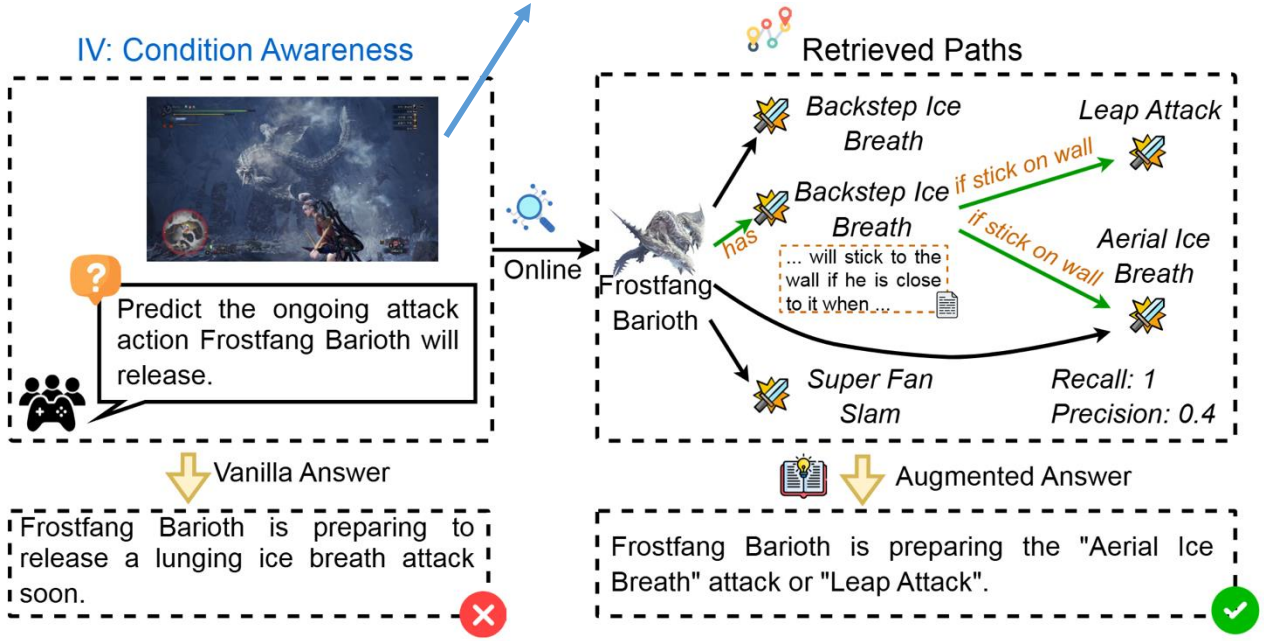
Models	Vanilla	Vanilla^{+}	Know.	Perceptive			Unaided-Offline			Unaided-Online		
				Acc.	*Pre.*	*Rec.*	*Acc.*	*Pre.*	*Rec.*	*Acc.*	*Pre.*	*Rec.*
GPT-4o [1]	.3122	.3924	.8565	**.7383**	.5061	**.7046**	.4050	.2595	.4416	**.5105**	.2756	**.5625**
GPT-4o mini [1]	**.3218**	**.4135**	.8481	.6877	.2963	.5450	**.4514**	.2059	**.5028**	.3544	.1626	.3009
Claude 3.7 Sonnet [2]	.2827	.3375	**.8987**	.7004	.5817	.6322	.3628	.2775	.3270	.4388	**.2911**	.4029
Claude 3.5 Sonnet [2]	.2869	.3755	.8776	.7215	**.5922**	.6800	.3966	**.3215**	.4008	.3966	.2330	.3270
Claude 3.5 Haiku [2]	.2356	.3206	.8823	.6455	.3739	.5007	.3544	.2002	.3164	.3670	.1735	.3361
Gemini 2.0 Flash [47]	.1983	.2995	.8438	.6919	.3507	.6146	.3839	.1703	.4092	.3713	.1515	.3663
Gemini 1.5 Pro [47]	.2194	.2700	.8438	.6962	.4761	.6033	.3164	.1615	.2194	.4050	.2122	.4585
Step-1o [43]	.2436	.2815	.8235	.5747	.4372	.5095	.3025	.1831	.2483	.3403	.2204	.2987
InternVL2.5-78B-MPO [12]	.1603	.2616	.8649	.5991	.4198	.5428	.2700	.1729	.2250	.3080	.1556	.2378
Qwen2.5-VL-72B [3]	.1476	.2616	.8734	.6244	.4602	.4908	.3206	.2139	.2383	.3164	.1615	.1814
Ovis2-16B [36]	.1645	.2573	.8902	.7046	.5407	.5949	.2869	.1963	.2383	.3459	.1853	.2878
MiniCPM-o-2.6 [53]	.1139	.2405	.8312	.4683	.3183	.4001	.2194	.1311	.2376	.1687	.0189	.0210
DeepSeek-VL2-Small [52]	.1139	.2362	.6455	.3586	.1419	.1708	.1814	.0400	.0759	.1181	.0042	.0042
Human (Knowledgeable)	.5252	—	—	—	—	—	—	—	—	.9033	.9207	.8535
Human (Random)	.0336	—	—	—	—	—	—	—	—	.6092	.7113	.6457

Some models achieve better performance when using online retrieval analysis (vision).

Retrieval-Augmented Example (Unaided-Online)



Cognition of Action Details

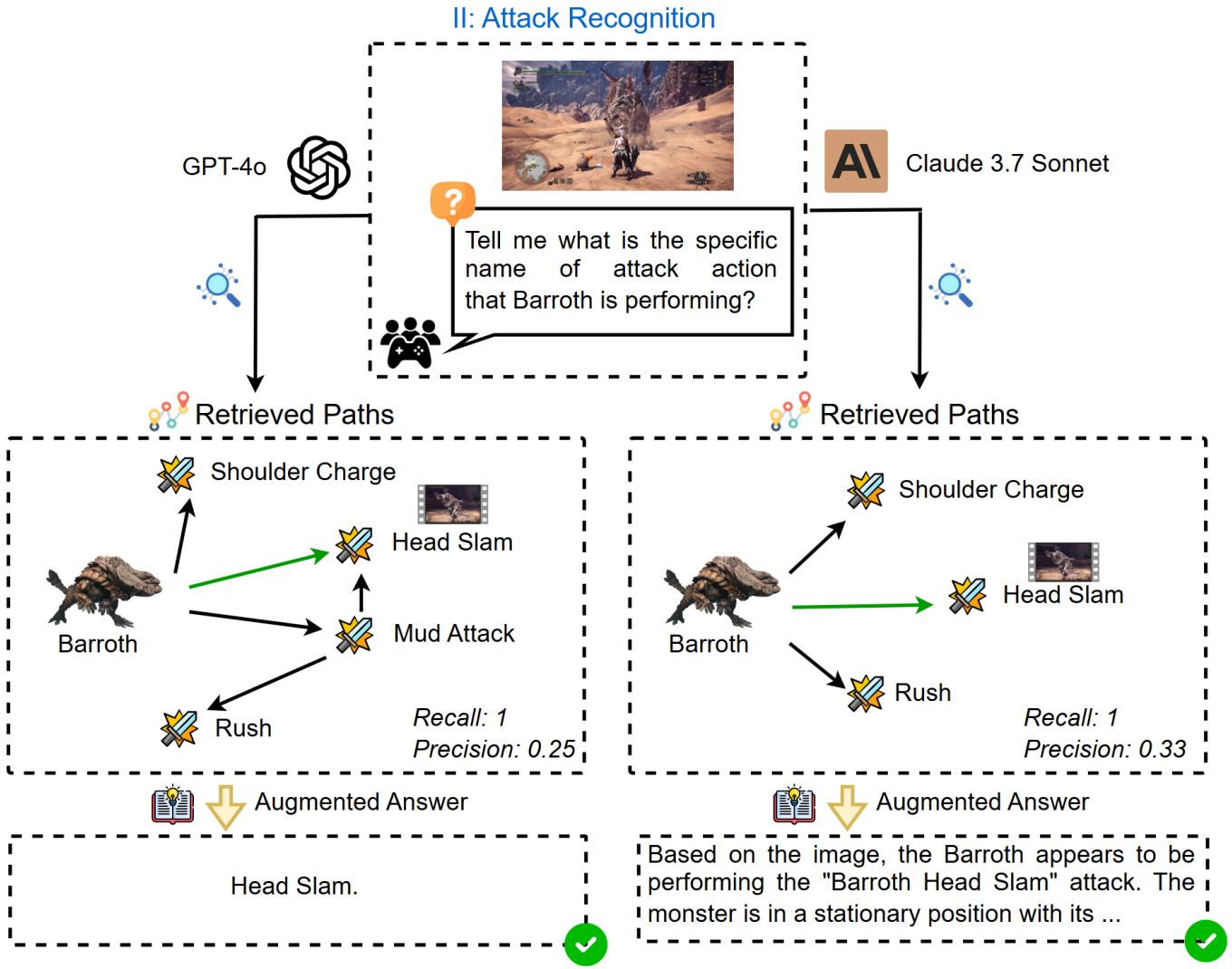


State Recognition

Model Comparison Example (Unaided-Online)

GPT-4o tends to search in a broader scope.

	Unaided-Online		
	Acc.	*Pre.*	*Rec.*
GPT-4o [1]	**.5105**	.2756	**.5625**
Claude 3.7 Sonnet [2]	.4388	**.2911**	.4029



The Advantages of Online Settings

Vis.-Off: offline generation

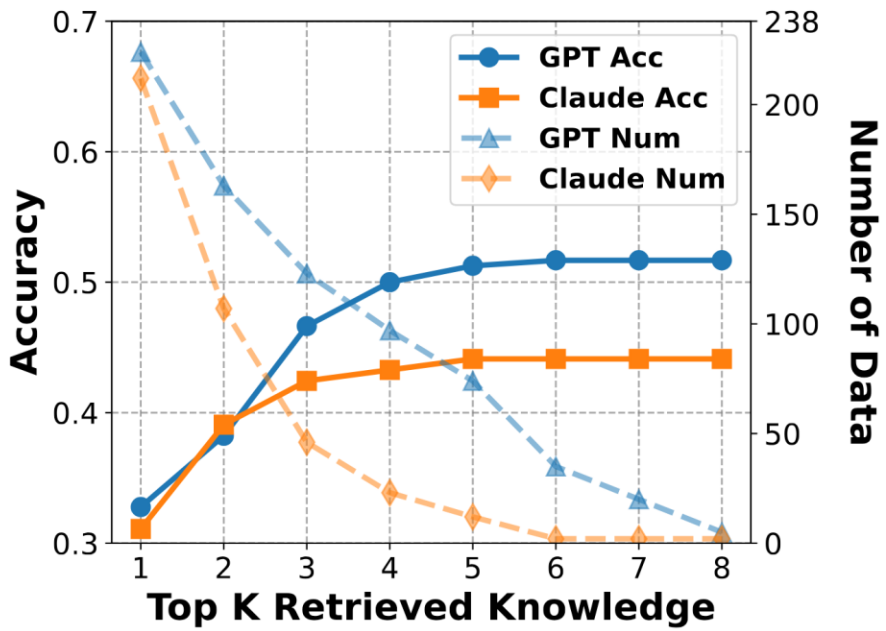
Vis.-On: online generation

Sim: Similarity of text descriptions (GPT-4o)

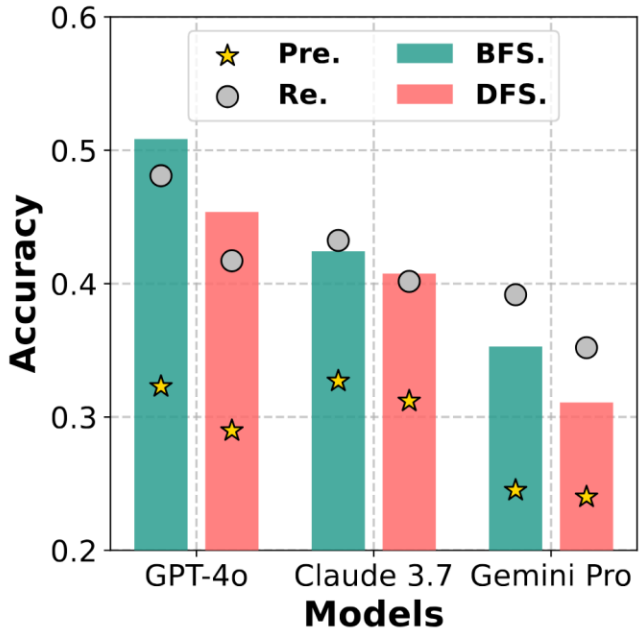
Model	Vis.-Off	Vis.-On	*Acc.*	*Pre.*	*Rec.*	*Sim.*
GPT-4o [1]	✔		.4050	.2595	.4416	.2806
GPT-4o [1]		✔	.5105	.2756	.5625	.2948
Claude 3.7 Sonnet [2]	✔		.3628	.2775	.3270	.2776
Claude 3.7 Sonnet [2]		✔	.4388	.2911	.4029	.3208
Gemini 1.5 Pro [47]	✔		.3164	.1615	.2194	.1608
Gemini 1.5 Pro [47]		✔	.4050	.2122	.4585	.1746
InternVideo2.5 [50]	✔		.3697	.1960	.2959	.0525
VideoChat-Flash [32]	✔		.3445	.2135	.2863	.0644

Knowing what it is doing enables the model to better parse visual information.

Ablation Studies



(a) Using K paths as knowledge.



(b) BFS vs DFS

Most questions will not require retrieving more than five pieces of knowledge.

Prompt Settings

Perceiver Agent

Input Prompt

You are a professional Monster Hunter player. You are playing 'Monster Hunter: World'.
You will receive consecutive video frames displaying the battle screen with the monster {monster name}.
The given 'Question' regarding the battle screen is: {question}
Generate a 'Description' of the battle scene as your 'Response', detailing the monster's limb and body movements, mouth actions, surroundings, and other relevant details.
Note that you should not give any assumptions for the 'Description'.
Note that you should directly output your 'Response' and do not output any information other than your 'Response'.
Now, start to complete your task.
Your 'Response':

Validation Agent

Input Prompt

You are a professional Monster Hunter player. You are playing 'Monster Hunter: World'.
The text description of the battle screen is: {caption}.
Based on the battle screen, here is the 'Question' you need to answer: {question}.
To answer the above question, you are now searching a knowledge graph to find the route towards relevant knowledge.
You are a professional Monster Hunter player. You are playing 'Monster Hunter: World'.
To answer the above question, you are now searching a knowledge graph to find the route towards relevant knowledge. The following contents are the knowledge you found so far (up to current entity {entity}):
******
{memory}
******
And here is some information of current entity: {entity info}.
[You will also receive consecutive video frames showing the battle screen with the monster {monster name} as visual information for current entity {entity}.
Make a 'Description' (do not affected by previous text description of the battle screen for the 'Question') for the battle screen as a part of your 'Response'. 'Description' should include monster's limb and body movements, mouth, surrounding and others details.
Note that you should not give any assumptions for the 'Description'.]
You have to decide whether visual and text information of this entity together with previous found knowledge is sufficient for answering this 'Question'.
For sufficient analysis, your 'Answer' is 'Yes' or 'No'.
[Directly output your 'Response' as the combination of 'Answer' and 'Description', separating them directly by ';'.]
Note that you should not output any information other than your 'Response'.
Now, start to complete your task.
Your 'Response':

Thanks!

MH-Bench

wang@im.sanken.osaka-u.ac.jp