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

Supplementary Material

Contents

1. Detail of MH Benchmark Construction **1**
- 1.1. MH-MMKG 1
- 1.2. MH Benchmark 2

2. Experiment Details **2**
- 2.1. Prompt Template for Our Method 2
- 2.2. Knowledge Consistency Calculation 4
- 2.3. Human Evaluation of GPT-4o as a Judge . . 5
- 2.4. Additional Experiments for MH Benchmark 5
- 2.5. More Result Samples 6

1. Detail of MH Benchmark Construction

In this section, we detailed construction of our MH-MMKG and MH benchmark.

1.1. MH-MMKG

A total of 22 monsters are incorporated into the graph construction, with each represented as a subgraph connected through various relationships, such as species relation and elemental weaknesses. Each subgraph contains rich information crucial for successful conquests, particularly regarding attack strategies, combos, attack phases, and launch conditions. The monsters are: Anjanath, Azure Rathalos, Barroth, Bazelgeuse, Brachydios, Diablos, Frostfang Barioth, Glavenus, Kushala Daora, Legiana, Nergigante, Rathalos, Rathian, Teostra, Tigrex, Uragaan, Zinogre, Pink Rathian, Yian Garuga, Stygian Zinogre, and Radobaan from Monster Hunter World.

To ensure the quality, we hired three experienced Monster Hunter World players, each with over 200 hours of game play experience. They were tasked with gathering relevant monster information from sources such as Wiki, YouTube, and Bilibili to construct the graph. Additionally, since each monster has unique characteristics within the game, the structure of each subgraph is tailored accordingly. The entities are classified into 7 types as show in Table 1. Most of them are attack actions, making MH-MMKG more focused on battles with monsters. We also plan to explore more game elements in the future. Some entities are attached with text, image or video as its attribution. Note that all video or images are captured from Arena field. While for queries in MH Benchmark all visual media are captured from the Wild field. We also show the length statistic of video clips in Figure 1. It can be observed that most videos are around 1s to 5s.

Table 1. Types of entity in MH-MMKG.

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

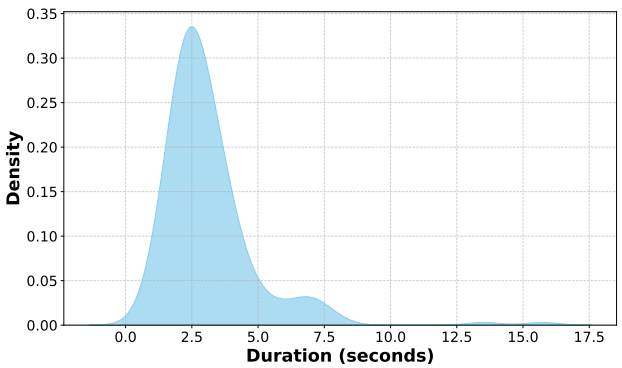

Figure 1. Video clip length statistic.

There are also 158 kinds of edges and 16 of them are base edges: "has attack action of", "continues with attack action of", "has attack variant of", "has attack phase of", "change attack phase to", "is mostly weaken with", "is weaken with", "is resistant with", "provide materials for", "can be stopped by", "has attack variants of", "generates", "cause", "turns to", "mated pair with", "has subspecies of". Some base edges (mostly the first two of them) are further combined with specific constrain mechanism to form 142 variants. The samples of constrains are: "is angry", "hunter step into", "is close to", "stick on the wall", "is knocked by", etc. (We do not show all of them as they too many).

We present some sub-graphs to illustrate structural di-

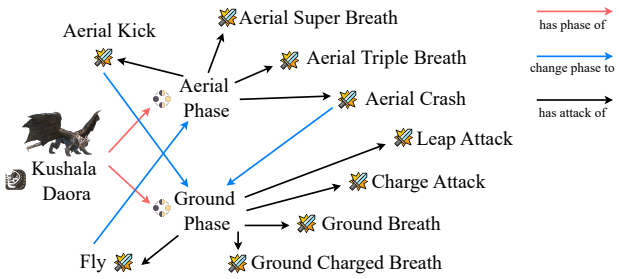

Figure 2. Subgraph structure for Kushala Daora.

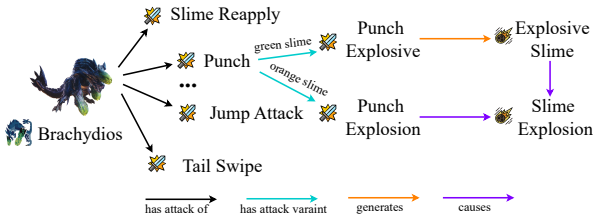

Figure 3. Subgraph structure for Brachydios.

versity, focusing only on attack actions and their related entities, as the other components resemble Figure 2 in the main paper. The main paper showcases the graph structure of Zinogre, known for its extensive combo attacks. Here, we provide two additional examples: Kushala Daora and Brachydios. Kushala Daora exhibits distinct attack patterns in its Aerial Phase (attacking from the air) and Ground Phase (attacking on the ground), as shown in Figure 2. Certain attacks can transition between these phases, making this information crucial for an MLLM to accurately answer related questions. Brachydios, on the other hand, has attack variations that depend on the color of the slime on its fists or head, as illustrated in Figure 3. The color change alters both the attack variant and its effect, adding another layer of complexity to its combat behavior. MLLMs have to comprehend such complex information to correctly answer the question in MH Benchmark.

1.2. MH Benchmark

To differentiate from MH-MMKG, all visual media for queries are captured from the Wild field. Additionally, we present statistics on the average number of entities and depth of knowledge associated with each query in the MH Benchmark, as shown in Table 2. It can be observed that sub-task I is relatively simple, as it relies solely on the topic node. In contrast, sub-tasks II and VI involve a greater number of steps and deeper analysis, as they pertain to combo recognition and cross-monster comparison, both of which require more complex reasoning.

Table 2. Average number of entities and depth of knowledge for each query in MH Benchmark.

Sub-tasks	I	II	III	IV	V	VI
Number_{avg}	1	2.339	3.535	2.4137	3.028	4.076
Depth_{avg}	1	2.278	3.250	2.4137	2.900	3.038

Table 3. Prompt for *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':

Table 4. Prompt for *topic entity selection* 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} All possible monster names 'Options' are structured in a list format as follows:{topic entity} Note that your 'Response' is to directly output the name of the monster you are looking for. Note that you should not output any information other than your 'Response'. Now, start to complete your task. Your 'Response':

2. Experiment Details

In this section, we show the detailed settings of our baseline method, including prompt for each agent, additional experiments, and more samples.

2.1. Prompt Template for Our Method

We first present the prompt templates for all agents in the retrieval pipeline. Table 3 shows the prompt for the **perceiver** agent, which translates input images into text based on the given question. Table 4 provides the prompt for the **topic selection** agent, responsible for selecting the starting entity for knowledge retrieval from the graph. Table 5 contains the prompt for the **expansion** agent, which plans the next neighboring entity for search. Table 6 presents the prompt for the **validation** agent, designed to assess the efficiency of knowledge transfer from the starting entity to the current entity. Finally, Table 7 includes the prompt for the **summarizer** agent, which synthesizes the retrieved knowledge for final answer generation. Among these, the {monster name},

Table 5. Prompt for *expansion* 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. The following contents are the knowledge you found so far (up to current entity {entity}): ****** {memory} ****** You need to select the relevant 'Neighbor Entity' that may provide knowledge to answer the question. The relation and condition from current entity 'entity' to all 'Neighbor Entity' are: ****** {neighbor entity} ****** Your 'Response' is directly output the name of all relevant 'Neighbor Entity' and separate them directly by ';'. If there is no relevant 'Neighbor Entity', directly output 'None'. Note that if the 'Neighbor Entity' is an attack action, always choose it (if it is not highly irrelevant). Note that if the 'Neighbor Entity' is a phase, you can only choose one. Note that you should not output any information other than your 'Response'. Now, start to complete your task. Your 'Response':

displayed in blue text, represents additional information as a part of question. The {entity} represents the name of current entity during search. The {question} refers to the input query, while {topic entities} denote the names of all topic entities. {entity infp} is the visual irrelavent additional information for an entity. The {caption} is the generated description by the **perceiver** agent.

The {neighbor entity} are options of neighbor for current entity. It is presented in a text format consisting of a combination of entity-edge triplets and corresponding constraints or conditions (if any). Here is a neighbor sample for monster "Frostfang Barioth" attack action entity "Straight Ice Breath":

- "Straight Ice Breath" continues with attack action of "Super Fang Slam" (Condition: When hunter hitted by the breath...)
- "Straight Ice Breath" continues with attack action of "Tail Spin" (Condition: When Frostfang Barioth already released two...)

In our prompt, we instruct the model to select relevant neighboring entities while placing greater emphasis on attack action entities, as most tasks are designed around them. For phase entities, we allow the model to select only one, in accordance with the game mechanics.

The {memory} records the search path from the starting entity to the current entity, including entity names and all relevant information at each step. Below is an example illustrating this transition from a knowledge path:

Table 6. Prompt for *validation* agent. Content in [] is used solely for unaided-online experiments.

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':

$$\textit{Zinogre} \xrightarrow{\text{phase of}} \textit{Charged Phase} \xrightarrow{\text{attack of}} \textit{Double Slam} \tag{1}$$

will be transferred into:

- "Zinogre": Additional Information: Zinogre has the appearance of a wild wolf and lives in the mountains full of dense trees ...
- "Zinogre" has attack phase of "Charged Phase".
- "Charged Phase": Additional Information: Zinogre is charged, the body will be surrounded by electric ...
- "Charged Phase" has attack action of "Double Slam".
- "Double Slam": Action Description: Zinogre lowers his head and rubs the ground with...

Note that Additional Information is the attribution of an entity (if exist). Action Description is given as human-made caption in Knowledgeable experiments, pre-extracted from visual attribution (if exist) in unaided-offline, and dynamic generated for visual attribution in unaided-online (if exist). Especially, the [] highlights the content for unaided-online that requires the model to comprehend visual references

Table 7. Prompt for *summarizer* 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}. Based on the battle screen, here is the 'Question' you need to answer: {question}. Here is the 'Knowledge' you retrieved from a knowledge graph for this 'Question': ****** {knowledge} ****** Your 'Response' is to provide the answer for this 'Question' based on the retrieved Knowledge. Note that you should not give any analysis. Note that you should not output any information other than your 'Response'. Now, start to complete your task. Your 'Response':

during validation and output corresponding description as the temporal visual attribution for the current entity.

As shown in Table 7, the final agent **summary** will treat all retrieved paths as {knowledge} using the same strategy as {memory}. Each path will be converted into a text description and attached to the query as input.

Table 8 shown the prompt template for unaided-offline experiments. It is used to pre-extract the visual reference (images or video for MLLMs or Video models, respectively in our experiments) into text description. This transition is not related to query or search memory.

Note that, the prompts for the agent pipeline were developed using InternVL2.5-78B [3], with the expectation that even open-source models, by their instruction-following capabilities, can understand these prompts and generate responses in the required format. This ensures a fair comparison for all close-source models in the main paper. We further conducted a preliminary prompt robustness analyses for GPT-4o and Claude 3.7 (unaided-online). Our observations show that Claude generally exhibited robust performance across prompt variations, particularly for agents with straightforward instructions such as *Perceiver*, *Topic Selection*, and *Summarizer*. However, GPT-4o exhibited sensitivity to lexical choice. For instance, in the *Validation* agent, the use of the term "sufficient" to determine whether the retrieved knowledge is enough and the retrieval should be stopped. When we replaced it with "necessary," GPT-4o tended to more cautious during retrieval. This minor change led to a .0546 and .0871 drops on *Acc.* and *Rec.*, respectively, though with a .0194 improvement in *Pre.* These findings suggest that prompt robustness is both model-specific and agent-specific.

2.2. Knowledge Consistency Calculation

As defined in the main paper, the model's final output is a retrieved subgraph, denoted as $\hat{\mathcal{I}}$. We consider each path from the root entity to a leaf entity as a unique knowledge instance and represent the set of such paths as $\hat{\mathcal{L}}$. The knowledge consistency is computed between $\hat{\mathcal{L}}$ and the ground-truth knowledge paths $\mathcal{L}$ using a one-to-one matching approach.

The recall and precision of retrieved knowledge paths are defined as follows:

Table 8. Prompt for *offline* caption pre-extraction.

Input Prompt
You are a professional Monster Hunter player. You are playing 'Monster Hunter: World'. You will receive consecutive video frames showing the battle screen as visual information for {entity}. Make a 'Description' for the battle screen as your 'Response'. 'Description' should include monster's limb and body movements, mouth, surrounding and others details. Note that you should not output any information other than your 'Response'. Now, start to complete your task. Your 'Response':

Table 9. Prompt for accuracy calculation using GPT-4o as a judge.

Input Prompt
You are a professional Monster Hunter player. You are playing 'Monster Hunter: World'. Here is a 'Question' need to be answered: {question}. There are also two answers for this 'Question': Answers 1: {answer gt}. Answers 2: {answer pred}. Your 'Response' is to decide whether the content of these two answers are similar. If similar directly output 'Yes'. If not similar directly output 'No'. Note that you may ignore the format difference. Ignore the difference of monster name before word, e.g., Zinogre Leap Attack and Leap Attack are with same meaning. Here are some samples for decide similarity: Sample 1: 'Question': Tell me what is the specific name of attack action that Zinogre is performing? "Answer 1": Static Charge "Answer 2": Thunder Charge B "Response": No Sample 2: 'Question': Start with counterattack, Zinogre released the attack action shown in the input battle screen. Tell me what is the next attack action? "Answer 1": Zinogre Back Slam "Answer 2": Back Slam "Response": Yes Sample 3: 'Question': What attack action Brachydios is unleashing? "Answer 1": Brachydios is unleashing the Brachydios Ground Slime Explosion attack "Answer 2": Ground Slime Explosion "Response": Yes Note that you should not output any information other than your 'Response'. Now, start to complete your task. Your 'Response':

Table 10. Prompt for similarity calculation between generated and human-made caption using GPT-4o as a judge.

Input Prompt
You are a professional Monster Hunter player. You are playing 'Monster Hunter: World'. Here are two text description of a monster attack action. Your 'Response' is to decide whether the content of these two text descriptions are similar. Your should focus on the details of movement and some key information that can help you to discriminate the action. If similar directly output 'Yes'. If not similar directly output 'No'. The First description is {truth}. The Second description is {generated}. Note that you should not output any information other than your 'Response'. Now, start to complete your task. Your 'Response':

$$\text{Recall} = \frac{|\hat{\mathcal{L}} \cap \mathcal{L}|}{|\mathcal{L}|} \tag{2}$$

$$\text{Precision} = \frac{|\hat{\mathcal{L}} \cap \mathcal{L}|}{|\hat{\mathcal{L}}|} \tag{3}$$

where $\hat{\mathcal{L}} \cap \mathcal{L}$ represents the set of correctly retrieved knowledge paths. Recall measures the proportion of ground-truth knowledge paths successfully retrieved by the model, while precision measures the proportion of retrieved paths that are correct.

2.3. Human Evaluation of GPT-4o as a Judge

Tables 9 and 10 present the templates for using GPT-4o as a judge [5] to assess result accuracy (*Acc.*) and caption similarity (*Sim.*). For accuracy evaluation, we prompt GPT-4o to compare the similarity between the ground-truth answer {answer gt} and the generated answer {answer pred}. Additionally, we provide three few-shot examples as references for the model.

For caption similarity assessment, GPT-4o directly compares the human-written caption {truth} with the model-generated caption {generated}. To further evaluate GPT-4o's judging performance, we conducted a human experiment. As shown in Table 11, two knowledgeable players independently evaluated 200 randomly selected samples from GPT-4o's judgments across all experiments for each model. A judgment was considered correct if both evaluators agreed. Our findings indicate that while there are some variations across models, GPT-4o demonstrates a high overall accuracy in judgment (0.926). Although caption similarity scoring is lower, it remains sufficiently high for such a subjective task. Overall, the results show that using GPT-4o as a judge is with high feasibility.

Table 11. Human evaluation for GPT-4o judgment accuracy. Each model's generation for answer and caption is evaluated by 200 randomly select samples through two knowledgeable players.

	Answer Accuracy	Caption Similarity
GPT-4o [1]	0.925	0.865
Claude 3.7 Sonnet [2]	0.900	0.840
Ovis2-16B [4]	0.955	0.810
average	0.926	0.838

Table 12. Impact of having monster name and Extra information in question. ✔ means having such information.

		Unaided-Online			
Name	Extra	*Acc.*	*Pre.*	*Rec.*	*Top.*
		.2731	.1251	.2413	.5210
	✔	.3781	.2080	.4434	.7365
✔		.4075	.2120	.4636	1
✔	✔	.5105	.2756	.5625	1

Table 13. Computational cost per sample in average.

Models	Retrieval Time (s)	Agent Call (n)	Response Time (s)
GPT-4o	92.46 ± 68.93	7.20 ± 5.01	10.92 ± 9.70
Gemeni 2.0 Flash	17.15 ± 10.92	11.04 ± 9.42	1.12 ± 0.91
InternVL	57.06 ± 41.78	9.32 ± 3.58	7.33 ± 6.95

2.4. Additional Experiments for MH Benchmark

In Table 12, we present the impact of incorporating the monster's name (Name) and additional information (Extra) as part of the input question q. The metric *Top.* represents the accuracy of the model in selecting the correct topic entity as the retrieval root. We observe that removing the monster's name leads to a significant performance drop due to incorrect root entity selection (low *Top.*).

Additional information refers to contextual hints, such as a monster being angry, which players can infer from the game's text. These details are generally too subtle to be captured from images by MLLMs. Removing only the additional information also results in an obvious performance drop, indicating that such visually independent cues are essential for the model to generate the correct answer. One interesting observation is that with additional information *Top.* can be improved than no Name and Extra setting.

Table 13 reports average retrieval time (in seconds), number of agent calls (n rounds), and per-call response time in the format of mean ± std. Experiments were conducted using GPT-4o and Gemini 2.0 Flash via API, and InternVL2.5 on a local GPU server with 2 A6000. The results reveal efficiency as a limitation of the current agent pipeline. More results will be included.

We also perform ablation studies to assess the impact of using cross-models for two key agents: *Summarizer* (knowledge utilization) and *Validation* (knowledge

Table 14. Ablation for cross-models agent pipeline.

Models	Replace *Summarizer* With			Replace *Validation* With		
	GPT-4o	Claude	InternVL	GPT-4o	Claude	InternVL
GPT-4o	.5105	.4994	.4864	.5105	.4716	.4128
Claude	.4510	.4338	.4086	.5052	.4338	.3676
InternVL	.3876	.3624	.3080	.3413	.3225	.3080

retrieval), keeping other agents fixed. Table 14 shows *Acc.* results across GPT-4o, Claude 3.7, and InternVL2.5, with diagonal values representing the results of original single-model pipeline. *Summarizer* replacement yields little change between GPT-4o and Claude, indicating that performance gains stem more from retrieved knowledge quality than summarization strength (InternVL's column with better knowledge, the improvement is more evident than that in the rows, showing in green). In contrast, using a weaker model (InternVL) for *Validation* causes a sharp performance drop (in red), underscoring the importance of this role. Yet, upgrading only the *Validation* agent in InternVL brings limited benefit, suggesting other retrieval-stage agents affect a lot.

2.5. More Result Samples

This section presents some randomly selected examples of generated answers via various models.

Figure 4 shows a sample for "Glavenus" continues attack action recognition. Both GPT-4o and Claude 3.7 output wrong answer, although GPT-4o catch the path towards true knowledge. Show models lack the ability to comprehend the knowledge.

Figure 5 shows a sample for "Bazelgeuse" attack action recognition. Although some difference in response, both GPT-4o and Gemini 1.5 Pro generate correct answer. GPT-4o find more paths as its knowledge augmentation.

Figure 6 shows a sample for "Barroth" attack action recognition. Both GPT-4o and Claude 3.7 generate the correct answer, however, GPT-4o's answer is more clear, showing better instruct following ability.

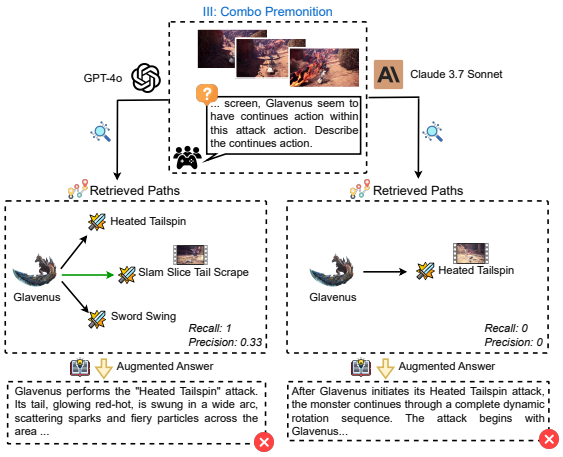

Figure 4. A sample for "Glavenus" continues attack recognition.

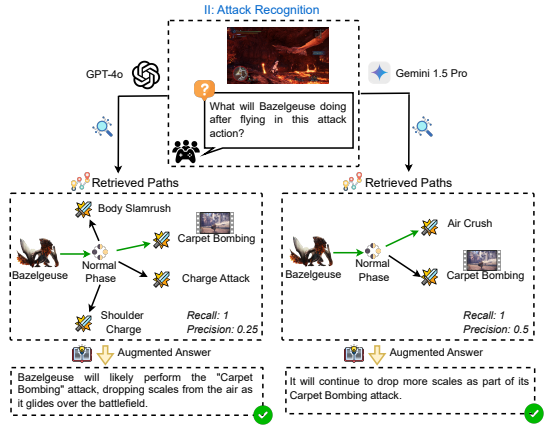

Figure 5. A sample for "Bazelgeuse" attack action recognition.

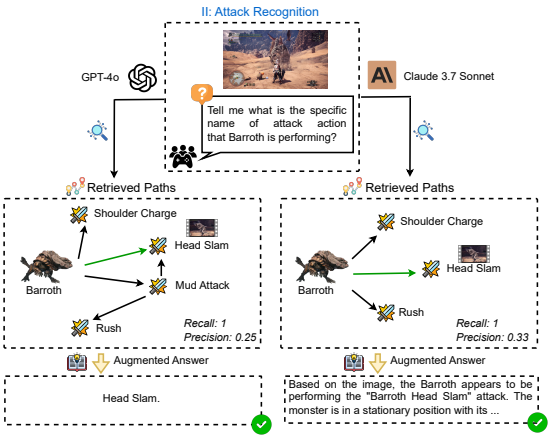

Figure 6. A sample for "Barroth" attack action recognition.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 5

[2] Anthropic. The claude 3 model family: Opus, sonnet, haiku. 2024. 5

[3] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 4

[4] Shiyin Lu, Yang Li, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Han-Jia Ye. Ovis: Structural embedding alignment for multimodal large language model. *arXiv:2405.20797*, 2024. 5

[5] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li,

Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *NeurIPS*, 36:46595–46623, 2023. 5