

## Appendix

In the Appendix, we provide additional details to support the main content of the paper. First, we expand on the specifics of the VTTS-80K dataset, including its composition and annotation types. Next, we detail the training and inference parameters used for the model. In the Case Study section, we present multiple cases of model reasoning, showcasing both successful examples and an analysis of failure cases to provide a comprehensive understanding of the model’s performance. Finally, we include a discussion of the paper’s limitations, broader societal impacts, and safeguards to address potential risks, ensuring a responsible and ethical approach to the development and deployment of the proposed method.

## A Details of VTTS-80K

Table 1: VTTS-80k Data Distribution and Examples.

Data Type	Data Source	Data Example
Video QA	LLaVaVideo [5]	<b>Question:</b> Why does the text 'FROM ACCLAIMED DIRECTORS WERNER HERZOG AND CLIVE OPPENHEIMER' appear? <b>Options:</b> A. To introduce the festive atmosphere, B. To show the location of the observatory domes, C. To highlight the geological patterns of the crater, D. To indicate the directors of the film <b>Think:</b> The text "FROM ACCLAIMED DIRECTORS WERNER HERZOG AND CLIVE OPPENHEIMER" appears in the video to indicate the directors of the film. This is a standard practice in cinematic trailers to highlight the creative team behind the production, especially when the directors are renowned figures like Werner Herzog (a celebrated filmmaker) and Clive Oppenheimer (a producer known for documentaries). Their names serve to attract audiences familiar with their previous works and establish credibility for the film's quality. <b>Answer:</b> D. To indicate the directors of the film
	PerceptionTest [8]	<b>Question:</b> What action not related to making tea did the person do? <b>Options:</b> A. clapping hands B. stacking pots on the stove C. putting a sugar cube into the teabox. <b>Think:</b> In the video, the person prepares tea by opening a tea box, placing a tea bag in a mug, pouring hot water, adding sugar, and stirring. There is no interaction with the stove or pots; the focus is solely on tea preparation. <b>Answer:</b> C. The cup/glass/bottle.
	Star [10]	<b>Question:</b> What happened after the person tidied up the clothes? <b>Options:</b> A. Took the towel. B. Opened the refrigerator. C. Put down the shoe. D. Closed the laptop. <b>Think:</b> After tidying up the clothes, the person picked up the towel from the floor and carried it with them as they exited the room. This action directly follows the tidying process and aligns with the sequence shown in the video. <b>Answer:</b> A. Took the towel.
Temporal Clue	Charades [3]	<b>Question:</b> To accurately pinpoint the event "the person takes a box out" in the video, determine the precise time period of the event. <b>Think:</b> The event "the person takes a box out" occurs after entering the room and bending down to retrieve the box from a low shelf or the floor, then exiting the room while carrying the box. The sequence is: Enters the room. Bends down to pick up the box. Stands up with the box. Exits the room. <b>Clue:</b> [6.0, 11.9]
	NextGQA [11]	<b>Question:</b> Why does the woman in blue swipe her eyes near the end of the video? <b>Options:</b> A. remove makeup, B. wipe tears, C. dance together, D. make face, E. put cream on man's face <b>Think:</b> The woman in blue is seen covering her face with her hands and then wiping her eyes. This action suggests she might be emotional or crying. The options provided do not include 'remove makeup' or 'put cream on man's face,' so we can eliminate those. 'Dance together' and 'make face' do not fit the context of the scene. Therefore, the most likely reason is that she is wiping tears. <b>Clue:</b> [11.6, 19] <b>Answer:</b> B. wipe tears
Image Reasoning	Multimodal-Open-R1-8k [1]	<b>Question:</b> Based on the image, calculate the ratio of the base of the largest triangle to the base of the smallest triangle. <b>Options:</b> A. 3 to 2. B. 5 to 3. C. 2 to 1. D. 9 to 5 <b>Think:</b> To determine the ratio of the base of the largest triangle to the smallest triangle, we analyze the similar triangles formed in the image. The largest triangle is $\triangle ACE$ with base $AE=30$ , and the smallest triangle is $\triangle CBD$ with base $BD$ . The horizontal lines divide the sides proportionally: on the left side $CA$ , the segments are 3 and 3 (total 6), and on the right side $CE$ , the segments are 5 and 5 (total 10). These divisions simplify to a ratio of 1:1, indicating similar triangles with a scale factor of 2:1. Since the triangles are similar, the base of the smallest triangle is half the base of the largest triangle: $BD=30 \times 1/2=15$ . Therefore, the ratio of the largest base to the smallest base is $30:15=2:1$ . <b>Answer:</b> C. 2 to 1.
	Geo [2]	<b>Question:</b> What is the degree measure of angle PEF in the quadrilateral ABCD figure, where AD is equal to BC, and point P serves as the midpoint of diagonal BD, while E and F represent the midpoints of AB and CD, if angle EPF equals 130 degrees? <b>Options:</b> A. $25^\circ$ B. $30^\circ$ C. $35^\circ$ D. $50^\circ$ <b>Think:</b> Since P and F are the midpoints of BD and CD, respectively, we have $PF = 0.5BC$ . Similarly, we can obtain $PE = 0.5AD$ . Since $AD = BC$ , we have $PF = PE$ . Since angle $EPF = 130^\circ$ , we have angle $PEF = \text{angle } PFE = 0.5 \times (180^\circ - 130^\circ) = 25^\circ$ . <b>Answer:</b> A. $25^\circ$
	ScienceQA [6]	<b>Question:</b> Which better describes the Tibetan Plateau ecosystem? <b>Options:</b> A. It has warm summers. It also has cool winters. B. It has long, cold winters. It also has mostly small plants. <b>Think:</b> A tundra is a type of ecosystem. Tundras have the following features: long, cold winters and short, cold summers, soil that is frozen year-round, and mostly small plants. So, the Tibetan Plateau has long, cold winters. It also has mostly small plants. <b>Answer:</b> B. It has long, cold winters. It also has mostly small plants.
Spatial Clue	GOT [4]	<b>Question:</b> Please track the motorcycle wheel in the video, its initial coordinates are [107, 54, 159, 82] <b>Think:</b> The video depicts a scene of a motorcycle passing by. The tire of the motorcycle first moves to the upper left and then rotates nearby. <b>Clue:</b> [107, 54, 159, 82], [122, 59, 161, 86], [126, 61, 169, 86], [136, 69, 179, 92]
	RefCOCO [12, 7]	<b>Question:</b> Detect the giant doughnut with white icing and red , white , and blue sprinkles. <b>Think:</b> The giant doughnut with white icing and red, white, and blue sprinkles is clearly visible on the right side of the tray held by the man. It is distinct from the chocolate-frosted doughnut on the left. <b>Clue:</b> [334.72, 298.08, 522.88, 450.23999999999995]
	VisualCoT [9]	<b>Question:</b> Can you tell me about the hairstyles of the individuals in the image? <b>Options:</b> A. They have shaggy hair, B. They have perfectly straight hair. C.They have short cropped hair. D. They have very short buzz cuts. <b>Think:</b> The image depicts a casual garden scene, making shaggy hair (Option A) the most fitting choice due to its relaxed, low-maintenance style. Perfectly straight or short, structured cuts (Options B-D) are less suited to such informal outdoor settings. <b>Clue:</b> [98,114,240,146] <b>Answer:</b> A. They have shaggy hair.

### A.1 Dataset Annotation

As shown in Tab. 1, the VTTS-80K dataset is composed of four key components: VideoQA, Temporal Clue, Image Reasoning, and Spatial Clue, which are carefully curated from existing datasets to support iterative perception. The dataset includes five types of annotations: question, option, think,

clue, and answer. Among these, the clue annotation plays a critical role, as it encapsulates the essential information required to answer the question. This annotation is further divided into two subcategories: temporal clue, which captures time-related information, and spatial clue, which focuses on location or spatial relationships within the input data.

Notably, not all datasets contain every type of annotation. To address this, our training process dynamically selects the corresponding reward function based on the annotations available in each dataset. This ensures that the model can effectively leverage the diverse information provided by different subsets of the dataset while maintaining robustness across tasks with varying levels of annotation completeness. By aligning the reward mechanism with the available annotations, our approach maximizes the utility of the data and enables efficient learning even when certain types of annotations are missing.

## A.2 Dataset Statistics

The VTTS-80K dataset is designed to support diverse multimodal reasoning tasks, with varying annotation types across its data sources. In the dataset, all entries include Question and Think annotations, which form the foundational components for reasoning and iterative perception. However, the specific annotation types differ depending on the source of the data.

Data from three sources—GOT [4], Charades [3], and RefCOCO [12, 7]—are primarily of the grounding or tracking type. In these datasets, the task involves directly identifying the relevant clue (either temporal or spatial) within the input, and as such, they only provide clue annotations without accompanying QA pairs. On the other hand, the remaining datasets follow a QA format, where the primary focus is on answering questions based on visual and textual inputs. Among these QA datasets, NextGQA [11] and VisualCoT [9] provide both QA annotations and clue annotations, enabling joint reasoning over questions and supporting evidence. In contrast, the rest of the datasets contain only QA annotations, lacking explicit clue information.

In total, the VTTS-80K dataset comprises 15K temporal clues, 30K spatial clues, 80K Think annotations, and 50K QA pairs. This diverse composition ensures that the dataset supports a wide range of tasks, from fine-grained spatiotemporal localization to complex reasoning over multimodal inputs. By incorporating data with varying levels of annotation richness, VTTS-80K not only facilitates training models to handle incomplete or heterogeneous data but also reflects real-world scenarios where annotation availability may vary significantly.

## B Training and Eval Details

Table 2: Settings for VTTS RL training.

<i>Shared</i>	<b>Learning Rate</b>	$2 \times 10^{-6}$
	<b>Epoch</b>	1
	<b>Optimizer</b>	AdamW
	<b>Warmup Ratio</b>	0
<i>Video</i>	<b>LR Schedule</b>	Linear
	<b>Totao Batch Size</b>	16
	<b>Frames</b>	4~768
	<b>FPS</b>	2
<i>Image</i>	<b>Video Max Pixels</b>	$768 \times 28 \times 28$
	<b>Video Min Pixels</b>	$128 \times 28 \times 28$
	<b>Max Ratio</b>	200
	<b>Max Pixels</b>	$768 \times 28 \times 28$
<i>Grounding</i>	<b>Min Pixels</b>	$4 \times 28 \times 28$

Table 3: **Settings for VTTS Eval.** Key Ratio indicates the ratio of video frames selected from the time clues.

<i>Video</i>	<b>Frames</b>	4~2048
	<b>FPS</b>	2
	<b>Video Max Pixels</b>	$768 \times 28 \times 28$
	<b>Video Min Pixels</b>	$128 \times 28 \times 28$
<i>Image</i>	<b>Key Ratio</b>	0.5
	<b>Image Factor</b>	28
	<b>Max Ratio</b>	200
	<b>Max Pixels</b>	$768 \times 28 \times 28$
<i>Grounding</i>	<b>Min Pixels</b>	$4 \times 28 \times 28$
	<b>Spatial Output</b>	Pixel
	<b>Temporal output</b>	Time
<i>Tracking</i>	<b>Tracking Output</b>	Pixel Sequence

The VTTS RL training is configured with the following parameters. We use an AdamW optimizer with a learning rate of  $2 \times 10^{-6}$ , zero weight decay, and a linear learning rate schedule without warmup. The total batch size is set to 16. For video inputs, the number of frames ranges from 4 to 768 at a fixed frame rate of 2 FPS, with video resolution constrained by a maximum pixel count of

50  $768 \times 28 \times 28$  and a minimum of  $128 \times 28 \times 28$ . In terms of image settings, the image factor is set  
51 to 28, with a maximum scaling ratio of 200. The image resolution follows the same pixel constraints  
52 as the video input, ranging from a minimum of  $4 \times 28 \times 28$  to a maximum of  $768 \times 28 \times 28$ .

53 Our VTTS inference is configured with the following parameters. For video inputs, the number  
54 of frames ranges from 4 to 2048 at a fixed frame rate of 2 FPS, with a maximum pixel count of  
55  $768 \times 28 \times 28$  and a minimum of  $128 \times 28 \times 28$ . A key ratio of 0.5 is applied to select video frames  
56 based on time clues. In image settings, the image factor is set to 28, with a maximum scaling ratio  
57 of 200. The image resolution follows the same pixel constraints as the video input, ranging from  
58  $4 \times 28 \times 28$  to  $768 \times 28 \times 28$ . For grounding tasks, the model outputs spatial predictions in pixels,  
59 temporal predictions in time units, and tracking predictions as sequences of pixel coordinates.

## 60 C Case Study

### 61 C.1 Good Cases

62 As shown in Fig. 1 and Fig. 2, our Iterative Perception (ITP) approach demonstrates strong perfor-  
63 mance on both video QA and image reasoning tasks. In the two cases presented, the model initially  
64 fails to provide the correct answer during the first round of reasoning. However, through iterative  
65 perception, the model successfully identifies the critical clues required to answer the questions.  
66 By leveraging the multi-step "think" process, which integrates insights from multiple perception  
67 iterations, the model ultimately arrives at the correct answers. This highlights the effectiveness of  
68 iterative refinement in enhancing the model's ability to locate relevant information and reason more  
69 accurately over complex multimodal inputs.

70 As shown in Fig 3, Fig 4, Fig. 5, Fig. 6 and Fig. 7, our model demonstrates strong performance  
71 across a variety of visual tasks, including grounded video QA, temporal grounding, grounded image  
72 QA, spatial grounding, and tracking. In these fine-grained spatiotemporal perception tasks, the  
73 model not only accurately provides the perceived temporal and spatial results but also generates  
74 well-reasoned explanations for its decisions. This dual capability highlights the model's robust  
75 reasoning and spatiotemporal perception abilities, showcasing its proficiency in handling complex,  
76 multimodal inputs while maintaining interpretability and precision.

### 77 C.2 Failure Cases

78 Fig. 8 and Fig. 9 presents two failure cases of the model, corresponding to scenarios where the  
79 model either consistently provides incorrect answers across multiple perception steps or initially  
80 answers incorrectly but later corrects itself. Through analysis of these cases, we observe that the  
81 model's failures are primarily due to inaccurate localization of critical clues during the identification  
82 process. For knowledge-based questions, the model struggles to pinpoint precise supporting evidence,  
83 which can lead to errors in reasoning. Additionally, the model may lose some global contextual  
84 information while focusing on localized clues, further contributing to its inability to arrive at the  
85 correct answer. These findings underscore the challenges of balancing fine-grained localization with  
86 holistic understanding in complex multimodal tasks.

## 87 D Discussions

### 88 D.1 Limitations

89 Despite the effectiveness of our proposed method, several limitations remain that warrant further  
90 investigation. First, the iterative nature of our approach may lead to increased inference time, as each  
91 additional perception step requires further computation. While this enhances performance, it could  
92 pose challenges for real-time applications or scenarios with strict latency constraints. Second, the  
93 "thinking" process inherent in iterative perception introduces potential safety concerns. Specifically,  
94 intermediate reasoning steps may generate unintended or inappropriate content, which could propagate  
95 into the final output. Such issues highlight the need for careful design and safeguards to ensure the  
96 reliability and safety of the model in practical deployments.

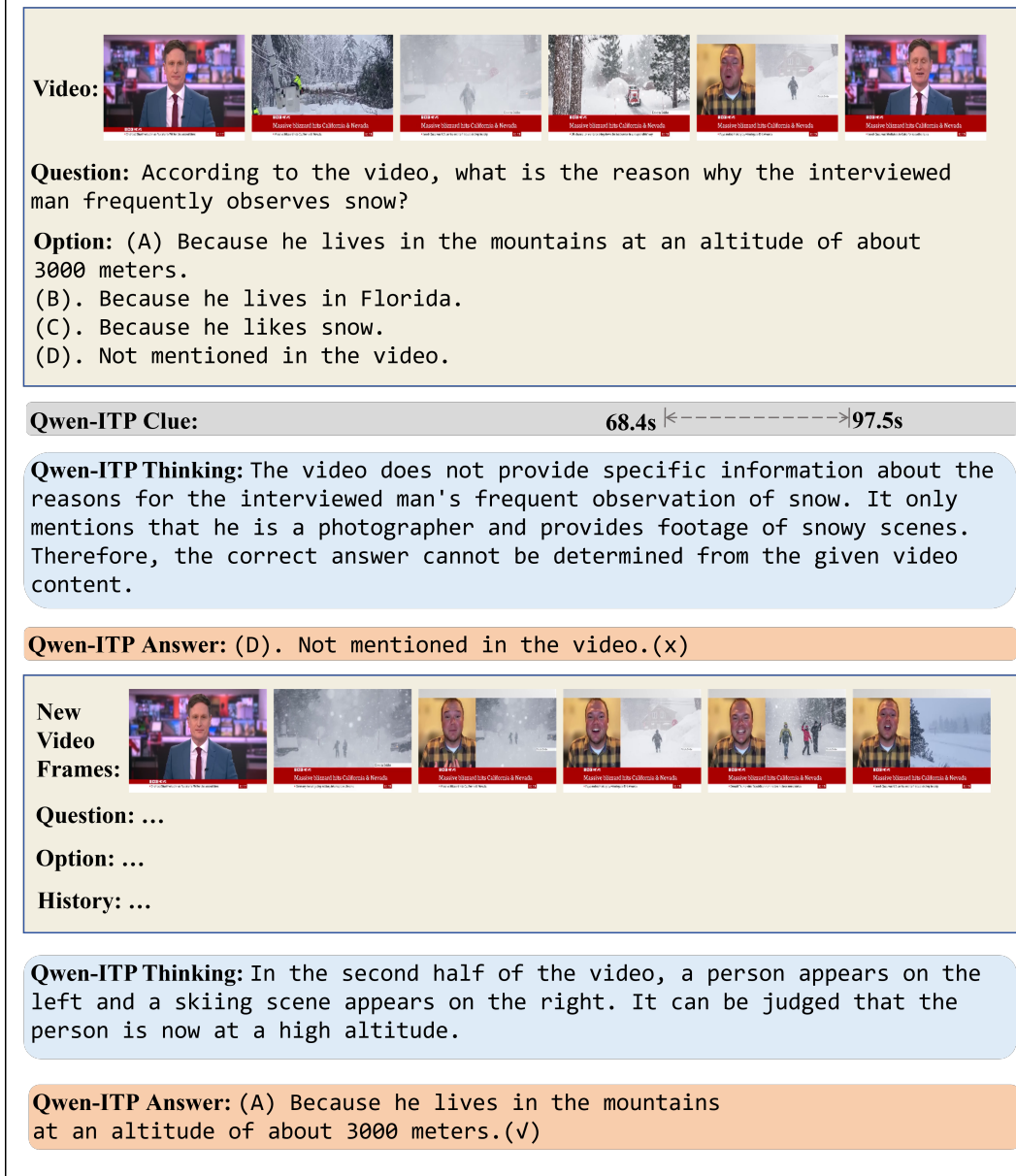


Figure 1: Iteration Perception For Video.

## 97 D.2 Broader Impacts

98 Our VTTS method for MLLMs introduces significant advancements in multimodal reasoning but  
 99 also raises important societal considerations, as outlined under the NeurIPS Code of Ethics.

100 VTTS significantly enhances multimodal reasoning by improving iterative visual perception, enabling  
 101 AI systems to interact with the world more accurately and in a human-like manner. This advancement  
 102 translates into superior performance across diverse tasks, including video conversation, image rea-  
 103 soning, and spatio-temporal perception, offering benefits in accessibility (e.g., assistive technologies  
 104 for visually impaired individuals), education (e.g., intelligent tools for complex subjects), safety  
 105 (e.g., surveillance and autonomous systems), and content creation (e.g., video summarization and  
 106 moderation). Furthermore, by introducing a novel test-time scaling approach and the VTTS-80K  
 107 dataset, this work lays a strong foundation for future research, paving the way for the development of  
 108 adaptive and efficient multimodal AI systems.



109 VTTS also introduces several potential societal risks, reflecting broader challenges associated with  
110 advanced AI systems. A significant concern is the heightened potential for misinformation and  
111 deepfakes, as the enhanced ability to generate and manipulate visual and textual content could be  
112 exploited to create deceptive or harmful media, such as disinformation campaigns or fabricated  
113 evidence. Additionally, the risk of hallucination—where models generate plausible but incorrect  
114 or unsupported outputs—poses further challenges, potentially leading to misleading conclusions in  
115 critical applications like news reporting, scientific analysis, or legal contexts. Privacy concerns also  
116 arise due to the model’s capacity to process detailed visual information, which could be misused  
117 in surveillance, unauthorized data collection, or intrusive monitoring scenarios. Moreover, biases  
118 inherent in training data may be amplified by VTTS’s improved reasoning capabilities, resulting  
119 in unfair or discriminatory outcomes in real-world applications, such as biased decision-making in  
120 hiring, law enforcement, or healthcare.

### 121 **D.3 Safeguards**

122 To address these risks, we adhere to the NeurIPS Code of Ethics, emphasizing responsible research  
123 practices and acknowledging the need to consider and mitigate potential harms. Although no specific  
124 technical safeguards against deepfake generation are outlined for VTTS, the broader AI community’s  
125 efforts in detection tools and responsible release strategies provide relevant mitigation pathways.  
126 Privacy concerns are addressed through a commitment to ethical guidelines and the integration of  
127 privacy-preserving techniques in future deployments. To tackle bias and fairness issues, rigorous  
128 dataset auditing and bias mitigation practices are essential.

Image:



**Question:** Which of the following is NOT true regarding the divided Germany shown on the map?

- Option:** (A): The capital of West Germany was located in East Germany  
(B): West Germany experienced a postwar economic miracle.  
(C): Both East Germany and West Germany were allowed to rearm  
(D): France had occupied West Germany.

**Qwen-ITP Clue:** [156, 209, 264, 437]

**Qwen-ITP Thinking:** A. The capital of West Germany was located in East Germany - This is not true. Bonn, the capital of West Germany, is clearly shown on the map as being west of the division line.  
B. West Germany experienced a postwar economic miracle - This is true. The economic recovery of West Germany after World War II is well-documented.  
C. Both East Germany and West Germany were allowed to rearm - This is not true. The map shows East Germany as part of the Soviet zone, which would have been under Soviet control and not allowed to rearm independently.  
D. France had occupied West Germany - This is not true. The map does not show any French occupation of West Germany; it is clearly marked as West Germany.

**Qwen-ITP Answer:** (C): Both East Germany and West Germany were allowed to rearm(x)

Image:



**Question:** ...

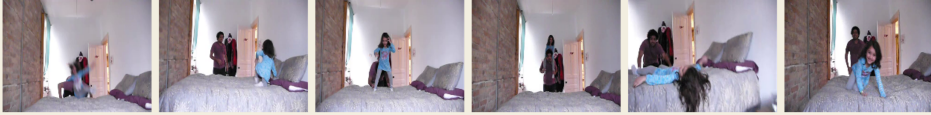
**Option:** ...

**History:** ...

**Qwen-ITP Thinking:** (A): The capital of West Germany was located in East Germany. This is false. The capital of West Germany was Bonn (marked on the map within West Germany), while East Germany's capital was East Berlin. (B): West Germany experienced a postwar economic miracle True. West Germany's rapid economic recovery in the 1950s (the Wirtschaftswunder ) is well-documented. (C): Both East Germany and West Germany were allowed to rearm True. West Germany joined NATO and rearmed under U.S. oversight, while East Germany formed the National People's Army under Soviet influence. (D): France had occupied West Germany True. France occupied the southwestern zone of West Germany as part of the Allied post-WWII occupation.

**Qwen-ITP Answer:** (A): The capital of West Germany was located in East Germany(✓)

Figure 2: Iteration Perception For Image.

**Video:** 

**Question:** Why did the man bend and lower his head to the bed?

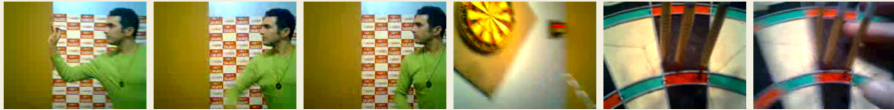
**Option:** (A) laughing  
(B) to pick the girl up  
(C) pick up a toy at the side  
(D) microphone too short  
(E) put phone

**Qwen-ITP Clue:** 3.8s <-----> 8.2s

**Qwen-ITP Thinking:** The man bends and lowers his head towards the bed as he is interacting with the girl who is jumping on it. This action suggests that he might be trying to engage with her or possibly to avoid getting hit by her. The context of the video shows a playful interaction between the man and the girl.

**Qwen-ITP Answer:** (B) to pick the girl up

Figure 3: Example For Grounded VideoQA.

**Video:** 


**Question:** To accurately pinpoint the event “the dartboard is shown with three darts in it.” in the video, determine the precise time period of the event.

**Qwen-ITP Clue:** 9s <-----> 19s

**Qwen-ITP Thinking:** The event “The dartboard is shown with three darts in it” occurs after the man throws the darts. The camera then focuses on the dartboard, showing the three darts embedded in it. This suggests that the event happens sometime after the man's actions.

Figure 4: Example For Temporal Grounding.

**Image:**



**Question:** What is the person in the image using to guide the animal?",

**Option:** (A):The person is using their hands to guide the animal.  
 (B):The person is using a wooden stick to guide the animal.  
 (C):The person is using a long piece of rope to guide the animal.  
 (D):The person is using a short piece of string to guide the animal.


**Qwen-ITP Clue:** [251, 60, 449, 256]

**Qwen-ITP Thinking:** The person in the image is guiding the animal using a long piece of rope. The pig is tied with a rope around its neck, and the individual is holding the other end of the rope, which appears to be several feet long, allowing them to control the pig's movement without direct physical contact.

**Qwen-ITP Answer:** (C):The person is using a long piece of rope to guide the animal.

Figure 5: **Example For Grounded ImageQA.**

**Image:**




**Question:** Please detect woman most left pink shirt.

**Qwen-ITP Clue:** [0, 259, 70, 480]

**Qwen-ITP Thinking:** The individuals visible from left to right include: A woman with dark hair wearing a blue jacket (over a black shirt). A woman with light brown hair wearing a white shirt. A bald man in a black shirt. A standing woman with blonde hair wearing a white shirt.


Figure 6: **Example For Detection.**

**Video:**



**Question:** To accurately pinpoint the event “the dartboard is shown with three darts in it.” in the video, determine the precise time period of the event.

**Qwen-ITP Clue:**



**Qwen-ITP Thinking:** The video depicts the process of four people on horseback chasing each other. The camera is shaking. The person riding at the end overtakes one of the people and swings a polo stick to play polo.

Figure 7: Example For Tracking.

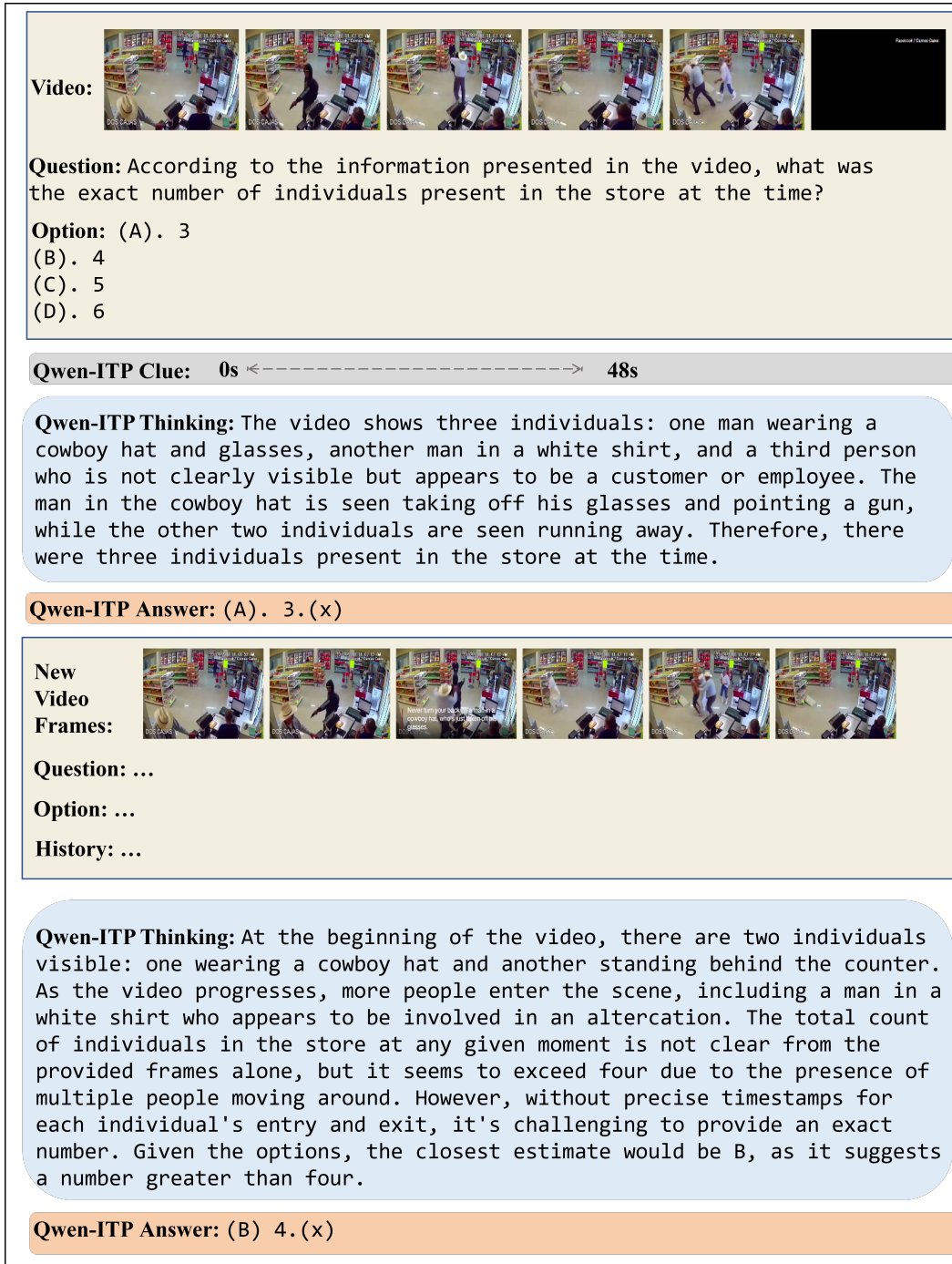


Image:



"The Devilfish in Egyptian Waters," 1882

**Question:** A similar cartoon could be made about each of the following countries prior to 1890 EXCEPT?

**Option:** (A): Netherlands.  
(B): France.  
(C): Belgium.  
(D): Germany.

**Qwen-ITP Clue:** [207, 426, 310, 622]

**Qwen-ITP Thinking:** The image depicts a cartoon from 1882 that portrays England as a powerful entity with tentacles extending across various regions, symbolizing its influence and control over these areas. The question asks which country among the given options (Netherlands, France, Belgium, Germany) would not have been depicted similarly due to the historical context before 1890. Given the time frame, it is unlikely that any of these countries would have been portrayed in such a manner because they were not as dominant or influential as England at that time. However, the question specifically asks for an exception, implying that one of these countries might have been depicted differently.

**Qwen-ITP Answer:** (D): Germany(✓)

Image:



**Question:** ...

**Option:** ...

**History:** ...

**Qwen-ITP Thinking:** The image is a political cartoon depicting England (symbolized by a figure with a top hat and "ENGLAND" on his belt) with tentacles extending to various colonies, illustrating British imperialism. The question asks which country prior to 1890 could not have a similar cartoon made about it, given their colonial activities. While Belgium established the Congo Free State in 1885 (just before 1890), its colonial presence was limited primarily to the Congo. Unlike the other options, Belgium lacked the multiple colonies required to justify a cartoon with numerous "tentacles" akin to the one shown.

**Qwen-ITP Answer:** (C): Belgium(x)

Figure 9: Failure Case For Image Iteration Perception

## References

- [1] EvolvingLMMs Lab. open-rl-multimodal: A fork to add multimodal model training to open-rl. <https://github.com/EvolvingLMMs-Lab/open-rl-multimodal>, 2025. Accessed: 2025-05-22.
- [2] Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wanjun Zhong, Yufei Wang, Lanqing Hong, Jianhua Han, Hang Xu, Zhenguo Li, et al. G-llava: Solving geometric problem with multi-modal large language model. *arXiv preprint arXiv:2312.11370*, 2023.
- [3] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, pages 5267–5275, 2017.
- [4] Lianghua Huang, Xin Zhao, and Kaiqi Huang. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *TPAMI*, 43(5):1562–1577, 2019.
- [5] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.
- [6] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022.
- [7] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, pages 11–20, 2016.
- [8] Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Mateusz Malinowski, Yi Yang, Carl Doersch, et al. Perception test: A diagnostic benchmark for multimodal video models. *Advances in Neural Information Processing Systems*, 36: 42748–42761, 2023.
- [9] Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *Advances in Neural Information Processing Systems*, 37:8612–8642, 2024.
- [10] Penghao Wu and Saining Xie. V?: Guided visual search as a core mechanism in multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13084–13094, 2024.
- [11] Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. Can i trust your answer? visually grounded video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13204–13214, 2024.
- [12] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *ECCV*, pages 69–85. Springer, 2016.