VRR-QA: Visual Relational Reasoning in Videos Beyond Explicit Cues

Supplementary Material

We organize the supplementary as follows:

- Section A: Data, Tool and Licenses
- Section B: Detailed Results
- Section C: VRR-QA Detailed Statistics
  - *Genres*
  - *Media Type*
  - *Movie Release Timeline*
  - *Difficulty*
  - *Question Word Distribution*
- Section D: Impact of Reasoning Prompt
- Section E: Experiment Statistical Significance
- Section F: Human Baseline
  - *Human baselines for Inferred Counting*
- Section G: Annotation Tool Interface
- Section H: Qualitative Results
  - *Viewpoint and Visibility*
  - *Physical and Environmental Context*
  - *Vertical Spatial Reasoning*
  - *Relative Depth and Proximity*
  - *Lateral Spatial Reasoning*
  - *Motion and Trajectory Dynamics*
  - *Social Interaction and Relationships*
  - *Inferred Counting*
  - *Motivational Reasoning*

A. Data, Tool and Licenses

We publicly release the benchmark, annotation tool, evaluation scripts and evaluation tool with Apache 2.0 license at https://swetha5.github.io/ImplicitQA/. We ran all our evaluations on NVIDIA A6000 48GB GPUs, and released the eval scripts for reproducibility.

B. Detailed Results

We present detailed results for varying model scales, temporal context across different models in Table 4. The best-performing open-source models were variants of the Qwen2.5 VL[31] family at 32B scale. Model scale had a noticeable but diminishing effect: Within Qwen2.5 VL, increasing model size from 3B to 32B parameters provided modest performance gains (approximately +2% accuracy improvement). Smaller models, such as LLaVA-OneVision, struggled significantly in challenging reasoning categories irrespective of scale. Distinct performance variations across reasoning categories emerged. Social Interaction was relatively easier, with accuracy up to 79.31% (qwen2.5 VL).

Analysis on increasing frame count generally improves model performance for certain architectures, but this improvement plateaus or slightly degrades beyond a threshold. For instance, LLaVA-NeXT-Video[39] exhibited a peak performance at 16 frames (33.9%) with a marginal decrease at 32 frames (32.56%). Similarly, LLaVA-OneVision[11] performed slightly better at 32 frames (43.16%) compared to 16 frames (43.4%), indicating negligible performance gains. For Qwen2.5-VL[3] 32B model, increasing frames from 4 to 32 resulted in accuracy improvements, particularly evident in the inferred counting & social interaction categories, achieving substantial accuracy improvements from 27.91% to 41.86% and from 62.07% to 79.31%, respectively. This suggests deeper frame context substantially aids in specific visual implicit reasoning tasks.

C. VRR-QA Detailed Statistics

This section provides additional statistics for VRR-QA benchmark, highlighting the dataset's diversity across multiple dimensions, including

- Genre
- Media type
- Movie Release Timeline
- Difficulty based on hard-ness score

C.1. Genres

To further characterize the content diversity in VRR-QA, we manually annotate both the primary and secondary genres of each video. We assign these genre annotations by considering genres listed on dedicated pages for each movie on publicly available sources such as IMDb[1] and Wikipedia[2].

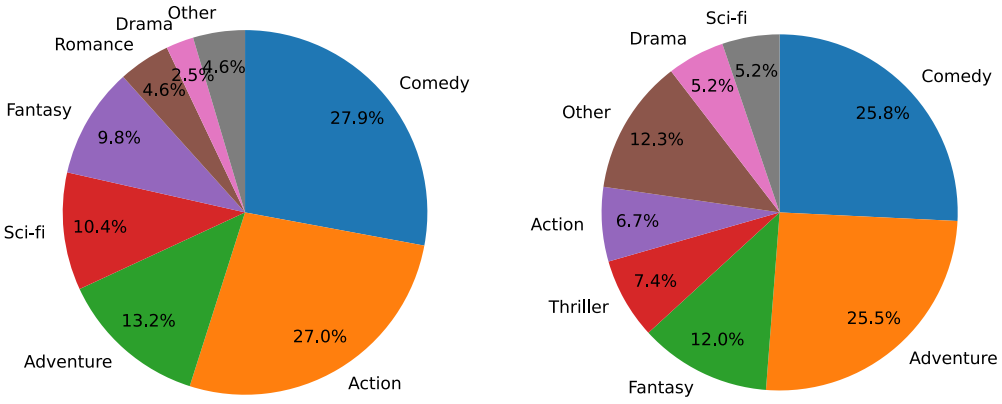
We assign each video a primary genre, which represents a movie's core theme and structure; and a secondary genre which reflects additional aspects of a movie. We observe that these primary and secondary genres come from a total of 15 different genres which are listed(in alphabetical order):

- Action
- Adventure
- Black comedy
- Comedy
- Crime
- Drama
- Fantasy
- Horror
- Mystery
- Psychological horror/thriller
- Romance
- Sci-fi
- Socio-political
- Thriller
- Western

[1] https://www.imdb.com/
[2] https://www.wikipedia.org/

Table 4. Detailed Results on VRR-QA for all visual relational implicit reasoning categories on various VideoLMMs in multiple settings.

Model	Scale	#Frames	Lateral Spatial Reasoning	Vertical Spatial Reasoning	Relative Depth and Proximity	Viewpoint and Visibility	Motion & Traj. Dynamics	Motivational Reasoning	Inferred Counting	Physical & Env. Context	Social Interaction & Relations	Avg.	Macro Avg.
LLaVA-NeXT-Video [39]	7b	8	34.78	31.48	28.95	43.90	37.36	40.24	23.26	42.86	55.17	33.72	37.56
	7b	16	36.00	29.60	30.10	48.80	36.30	39.00	30.20	35.70	51.70	33.90	37.50
	7b	32	34.78	29.63	30.08	39.02	36.26	36.59	27.91	35.71	37.93	32.56	34.21
LLaVA-OneVision [11]	7b	16	37.30	46.80	35.00	56.10	57.10	57.30	23.30	50.00	55.20	43.40	46.40
	7b	32	35.40	50.46	33.08	53.66	56.04	56.10	18.60	50.00	65.52	43.16	46.54
LLaVA-Video [40]	7b	8	31.68	41.20	29.32	48.78	57.14	57.32	13.95	50.00	62.07	39.02	43.50
	7b	16	36.00	44.00	31.60	56.10	60.40	62.20	14.00	50.00	62.10	42.10	46.30
	7b	32	36.02	47.22	32.71	51.22	58.24	63.41	16.28	57.14	68.97	43.27	47.91
Qwen2.5-VL [3]	3b	16	39.75	43.98	33.83	63.41	52.75	56.10	20.93	57.14	65.52	42.95	48.16
	7b	1	34.78	38.43	34.21	51.22	46.15	43.90	18.60	57.14	51.72	38.18	41.80
	7b	2	40.99	37.96	36.84	56.10	42.86	47.56	18.60	64.29	51.72	40.19	44.10
	7b	4	40.37	39.35	37.22	56.10	48.35	52.44	23.26	42.86	48.28	41.25	43.14
	7b	8	42.24	46.30	36.09	53.66	50.55	51.22	25.58	57.14	58.62	43.48	46.82
	7b	16	41.60	47.20	32.70	61.00	50.50	51.20	25.60	42.90	62.10	42.80	46.10
	7b	32	40.99	49.07	35.71	53.66	53.85	58.54	25.58	42.86	75.86	**45.07**	48.46
	32b	4	39.75	44.91	39.10	48.78	41.76	57.32	27.91	57.14	62.07	43.27	46.53
	32b	8	38.51	45.83	40.98	56.10	47.25	59.76	23.26	50.00	62.07	44.54	47.08
	32b	16	38.51	48.15	37.59	51.22	49.45	62.20	32.56	50.00	62.07	44.75	47.97
	32b	32	39.75	45.83	35.71	43.90	49.45	64.63	41.86	57.14	79.31	44.86	**50.84**
Qwen2-VL [31]	2b	16	34.78	43.98	47.37	60.98	39.56	43.90	16.28	57.14	51.72	42.84	43.97
	7b	1	38.51	39.35	36.84	53.66	42.86	48.78	13.95	42.86	55.17	39.66	41.33
	7b	2	36.65	40.28	40.60	46.34	50.55	50.00	20.93	42.86	58.62	41.57	42.98
	7b	4	39.13	45.83	39.47	46.34	49.45	51.22	20.93	50.00	58.62	43.05	44.56
	7b	8	37.27	47.69	40.98	53.66	48.35	54.88	16.28	50.00	65.52	44.11	46.07
	7b	16	39.80	46.80	40.60	51.20	52.70	58.50	16.30	35.70	72.40	44.90	46.00
	7b	32	40.37	49.07	40.98	48.78	48.35	60.98	16.28	35.71	62.07	44.96	44.73
LongVILA-R1 [5]	7b	8	39.13	24.09	26.69	48.78	42.86	48.78	22.41	49.06	41.38	33.67	38.13
	7b	16	35.40	28.18	30.08	39.02	42.86	52.44	31.03	50.94	58.62	35.86	40.95
	7b	32	34.78	25.00	29.32	41.46	42.86	56.10	31.03	52.83	51.72	35.16	40.57
	7b	64	34.16	28.64	26.69	53.66	45.05	46.34	29.31	49.06	48.28	34.67	40.13
	7b	128	26.71	26.82	25.19	46.34	43.96	48.78	27.59	49.06	51.72	32.47	38.46
Proprietary Models													
GPT 4.1 [21]	Full	16	42.90	53.20	51.10	48.80	59.30	82.90	41.90	71.40	75.90	54.30	58.60
	Mini	16	44.10	37.96	29.32	39.02	41.76	60.98	32.56	78.57	68.97	40.30	48.14
	Nano	16	35.40	32.87	28.95	58.54	37.36	36.59	18.60	42.86	55.17	34.25	38.48
GPT O3 [22]	Full	16	50.30	72.20	55.30	78.00	71.40	85.40	39.50	78.60	86.20	**64.10**	**68.60**



(a) Primary Genre Distribution

(b) Secondary Genre Distribution

Figure 10. Visualization of VRR-QA video statistics across primary and secondary genres for the top seven most frequent categories. A large proportion of animation videos fall under the Comedy genre, contributing to the higher number of samples annotated as Comedy in both primary and secondary genre distributions.

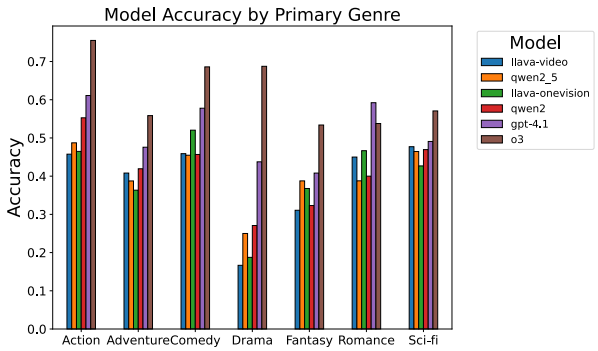


Figure 11. Model accuracy across primary video genres in the VRR-QA dataset. Performance varies significantly by genre, with O3 model consistently leading across genres except Romance.

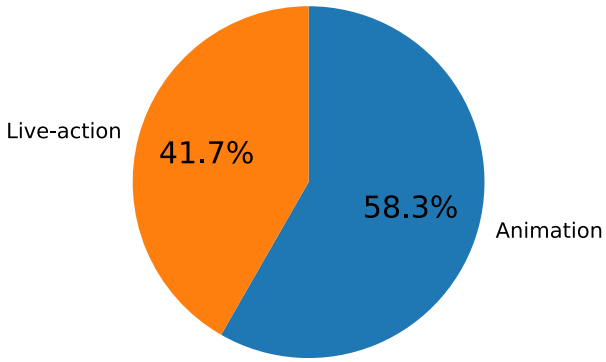


Figure 12. Distribution of Media Type in VRR-QA.

Figure 10 shows the primary and secondary genre distribution of our dataset. We observe that the dataset includes a wide range of genres, with Comedy, Action, and Adventure being the most prominent primary, while Comedy, Adventure and Fantasy being the top 3 secondary. This broad genre coverage ensures the benchmark captures diverse narrative structures, thematic elements, and stylistic conventions - essential for evaluating visual implicit reasoning across contexts.

To investigate how genre influences model performance, we show accuracy across primary genre categories in VRR-QA. As shown in Figure 11, genre plays a substantial role in performance variation. Overall models perform best on Action, Comedy, and Romance. In contrast, performance drops for genres like Drama and Fantasy. Notably, the O3 model outperforms all others across every genre except Romance, suggesting its stronger ability to generalize across narrative structures. The variation across genres also underscores the importance of content diversity in benchmark design, as genre-specific reasoning challenges reveal gaps in current video LMMs capabilities.

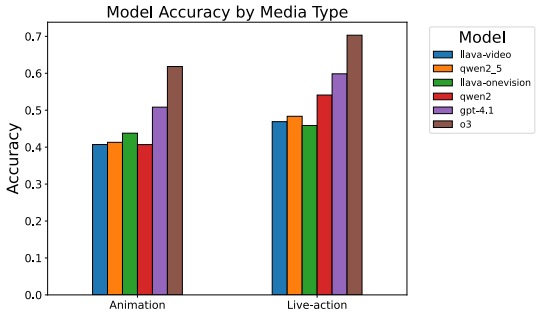


Figure 13. Model accuracy across media types (Animation vs. Live-action). Performance is consistently higher on live-action videos, with the largest gains observed in higher-capacity models such as GPT-4.1 and O3.

C.2. Media Type

We further categorize the videos into Live-Action and Animation to highlight the diversity in visual domains present in VRR-QA. As shown in Figure 12, the dataset maintains a balanced composition across both categories, with 58.3% of the videos being animated and 41.7% live-action. This mix ensures exposure to varied stylistic, motion, and rendering characteristics that challenges LMMs.

To further understand model generalization across different visual domains, we show performance on animation and live-action videos. As shown in Figure 13, all models demonstrate stronger performance on live-action content. The gap is especially more for larger models like GPT-4.1 and O3, which outperform others by a substantial margin. These results indicate that models may rely more effectively on grounded visual signals and realistic spatial cues present in live-action videos, whereas stylized representations in animation pose additional challenges for visual relational implicit reasoning. This highlights the need for further adaptation for animation-rich inputs.

C.3. Movie Release Timeline

We annotate the release year for each video and present the distribution by decade in Figure 14. The VRR-QA dataset spans a broad temporal range, covering films from the 1960s to current decade. A film's release period is often indicative of its visual and narrative style - including factors such as picture quality, cinematographic techniques, editing conventions, character costumes, and action design. This temporal diversity in VRR-QA enhances its realism and ensures broader generalization by exposing models to varied cinematic styles and storytelling conventions across eras.

C.4. Difficulty

To better understand model behavior under varying levels of difficulty, we propose a hardness-based partitioning of the

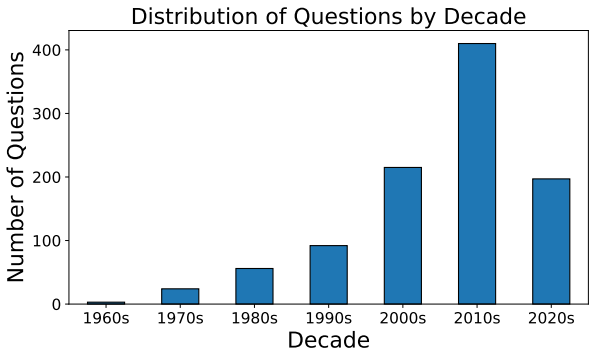


Figure 14. Distribution of videos in VRR-QA by release year. The dataset spans over ***7 decades***, capturing a wide range of visual styles, production techniques, and narrative conventions across different time periods.

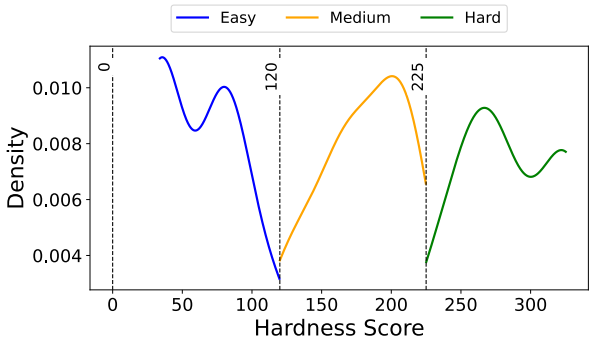


Figure 15. Density distribution of questions in VRR-QA based on hardness scores. Questions are categorized into three difficulty levels - Easy (0–120), Medium (120–225), and Hard (225+) - based on model performance scores. The distribution is approximately uniform, ensuring a balanced evaluation across varying difficulty levels.

VRR-QA dataset. Each question is assigned a hardness score derived from model performance: questions answered incorrectly by all models contribute more to the score, while those answered correctly by all models contribute none. Specifically, the hardness score is computed by summing each incorrect model's average accuracy. This metric reflects how broadly difficult a question is across model architectures. As shown in Figure 15, the distribution of hardness scores is approximately uniform across the three difficulty categories - Easy, Medium, and Hard - with roughly equal numbers of questions in each group. This balanced partitioning allows for a fair evaluation of model performance across difficulty levels.

Figure 16 shows model accuracy broken down by these categories. While all models perform well on Easy questions, accuracy drops substantially for Medium and Hard

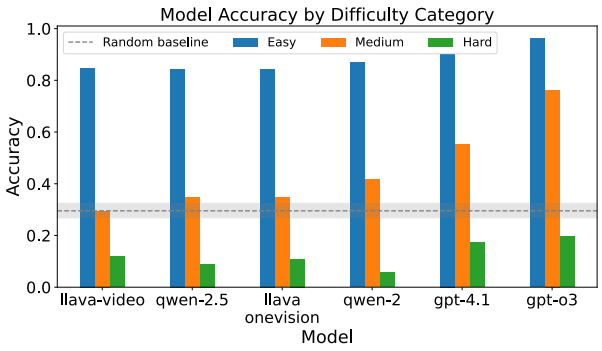


Figure 16. Model accuracy across difficulty categories. While all models perform strongly on Easy questions, performance drops significantly on Medium and Hard examples.

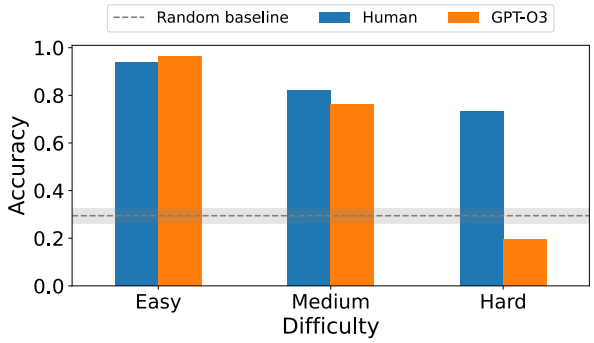


Figure 17. Accuracy comparison between GPT-O3 and non-expert human annotator across difficulty categories. While both perform well on Easy and Medium questions, human accuracy remains robust on Hard questions, whereas GPT-O3 performance drops significantly. Ground truth annotations were provided by expert annotators, underscoring the reasoning gap between models and even non-expert humans on complex questions.

examples. GPT-4.1 and GPT-O3 demonstrate better generalization across difficulty levels, whereas other models perform near or below random chance on the hardest questions. These findings reveal a steep difficulty gradient and highlight the value of hardness-aware analysis for assessing reasoning robustness. To contextualize model performance, we compare GPT-O3's accuracy against human performance. As shown in Figure 17, GPT-O3 performs comparably to humans on Easy questions and shows only a modest drop on Medium questions. However, the gap becomes pronounced on Hard examples: while human accuracy remains relatively high, GPT-O3's performance declines sharply, approaching random chance. It is important to note that the human baseline reflects responses from non-expert participants, while ground truth annotations in VRR-QA were created by expert annotators with domain familiarity. The relatively strong performance of non-experts highlights the accessibility of

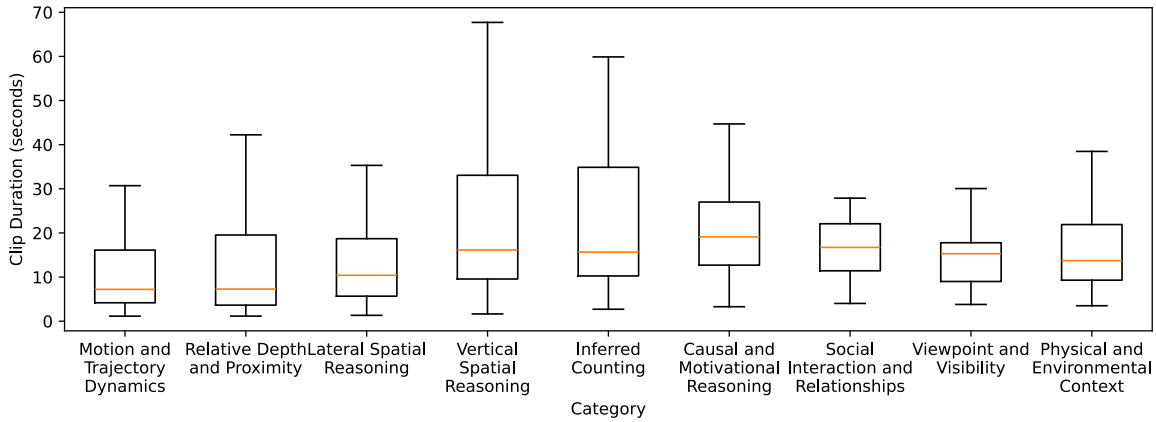


Figure 18. Question durations for each category.

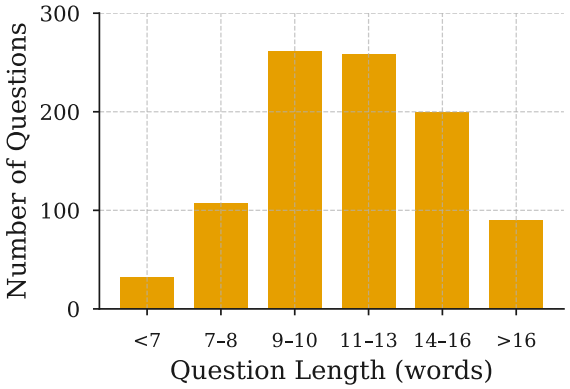


Figure 19. VRR-QA Question word counts

visual implicit reasoning for humans, even without expertize, while also emphasizing the performance ceiling that current models have yet to reach.

C.5. Question Word Distribution

In Figure 19 we present the word count distribution for questions in VRR-QA.

We also show differences in temporal length between categories in Figure 18, with counting and vertical spatial reasoning videos being the longest.

D. Impact of Reasoning Prompt

The reasoning prompt used to evaluate the impact of structured reasoning on GPT models is illustrated in Figure 20. This prompt is specifically designed to guide the model in analyzing video content for inferred reasoning across the various categories defined in VRR-QA. It breaks down the task into sequential steps-analyzing the video, summarizing key spatial relationships, highlighting important elements, answering the multiple choice question, and providing additional insights-thereby encouraging systematic focus on visual implicit reasoning.

The output format enforces a structured response, including a concise summary, bullet-pointed key themes and spatial cues, the selected answer, and contextual reasoning. This structured approach is intended to enhance the model's ability to infer unstated relationships and improve its overall accuracy.

As demonstrated in Figure 8 of the main paper, incorporating this reasoning-based prompt yields significant performance gains, improving the accuracy of GPT-4.1 Full by 3.9% and GPT-4.1 Mini by 4.8%. A detailed breakdown of the results is provided in Table 5. In Figure 21, we present a qualitative example illustrating the effectiveness of the reasoning prompt. When prompted without reasoning, GPT-4.1 incorrectly identifies the spatial relationship between characters as perpendicular. With the structured reasoning prompt, the model successfully breaks down spatial positions, directionality, and frame context to arrive at the correct answer: They are facing directly toward each other. This example highlights how the reasoning prompt not only improves accuracy but also fosters interpretability by making the model's decision-making process more transparent and spatially grounded.

E. Experiment Statistical Significance

In Table 6, we report the statistical error margins for the Qwen-2.5 VL model. To assess the variability of the model's performance, we conducted five independent runs using the same evaluation setup. For each of the nine visual implicit reasoning categories in VRR-QA as well as the overall accuracy, we compute the mean and standard deviation.

Reasoning Prompt

system_prompt = “ “ “
Assist users in understanding and gaining insights from video content, with particular emphasis on inferring the relative spatial positioning of various characters, objects, etc., and answer related multiple choice questions.
Guide users in digesting and analyzing the content of video material by breaking down the key themes, summarizing the narrative or subject matter, identifying important elements such as spatial relationships, characters, plot points, or pivotal moments, and answering any multiple choice questions related to the video.

 # Steps
1. **Analyze Video Content**: Identify main themes, characters, and narrative structure with a focus on spatial relationships.
2. **Summarize**: Write a brief summary of the video's key points and overall narrative, emphasizing spatial positioning.
3. **Highlight Key Elements**: Point out significant moments or elements in the video, particularly those that reveal spatial positioning and relationships.
4. **Answer Multiple Choice Question**: Review the provided multiple choice question and select the most accurate answer, considering spatial inferences.
5. **Provide Insights**: Offer any additional insights or context, specifically regarding spatial relationships, that could help in comprehending the video's content.

 # Output Format
Provide a structured text output that includes:
- A brief summary paragraph of the video focusing on spatial positioning.
- Bullet points highlighting key themes, spatial relationships, characters, and significant elements.
- The selected answer for the multiple choice question.
- Additional spatial relationship-focused insights as a paragraph.

 # Notes
- Ensure summaries are concise but comprehensive, with an emphasis on spatial understanding.
- Focus on elements that are crucial to understanding the video's spatial intent or message.
- Use clear and precise language to define spatial positioning.
- Make sure the multiple choice answer is clearly indicated and explained through spatial reasoning.
” ” ”

Figure 20. Reasoning prompt used to guide GPT models in analyzing video content. The prompt breaks down the task into structured steps - focusing on spatial relationships, narrative, summarizing key elements, selecting the correct answer, and providing insights - encouraging systematic reasoning aligned with VRR-QA’s visual implicit question categories.

Table 5. Results with Reasoning Prompt on VRR-QA for all visual implicit reasoning categories.

Model	Scale	Lateral Spatial Reasoning	Vertical Spatial Reasoning	Relative Depth and Proximity	Viewpoint and Visibility	Motion & Traj. Dynamics	Motivational Reasoning	Inferred Counting	Physical & Env. Context	Social Interaction & Relations	Avg.	Macro Avg.
GPT 4.1 [21]	Full	42.90	53.20	51.10	48.80	59.30	82.90	41.90	71.40	75.90	54.30	58.60
GPT 4.1-Reasoning [21]	Full	50.90	63.00	51.10	46.30	63.70	79.30	41.90	71.40	86.20	**58.20**	**61.50**
GPT 4.1 [21]	Mini	44.10	37.96	29.32	39.02	41.76	60.98	32.56	78.57	68.97	40.30	48.14
GPT 4.1-Reasoning [21]	Mini	36.02	48.61	33.08	56.10	54.95	65.85	39.53	64.29	72.41	**45.07**	**52.32**

F. Human Baseline

To establish a reference point for model performance, we evaluated with non-expert human annotator on the VRR-QA benchmark using our custom-built visual Implicit-Eval tool. The annotator was not provided with prior exposure to the dataset or answer keys and completed the evaluation at a natural pace, averaging approximately one minute per question. The annotator achieved an overall accuracy of 83.0%, with a macro-average score of 85.6% across reasoning categories. While performance was consistently strong across most categories, Inferred Counting emerged as the most challenging for the human. This aligns with the category’s reliance on temporal cues and visual implicit aggregation of visual information across multiple scenes, factors that often test not only reasoning but also memory and attention.

We further analyze the human accuracy across questions grouped by difficulty level using our hardness-based scoring method. As shown in Figure 17, the non-expert human

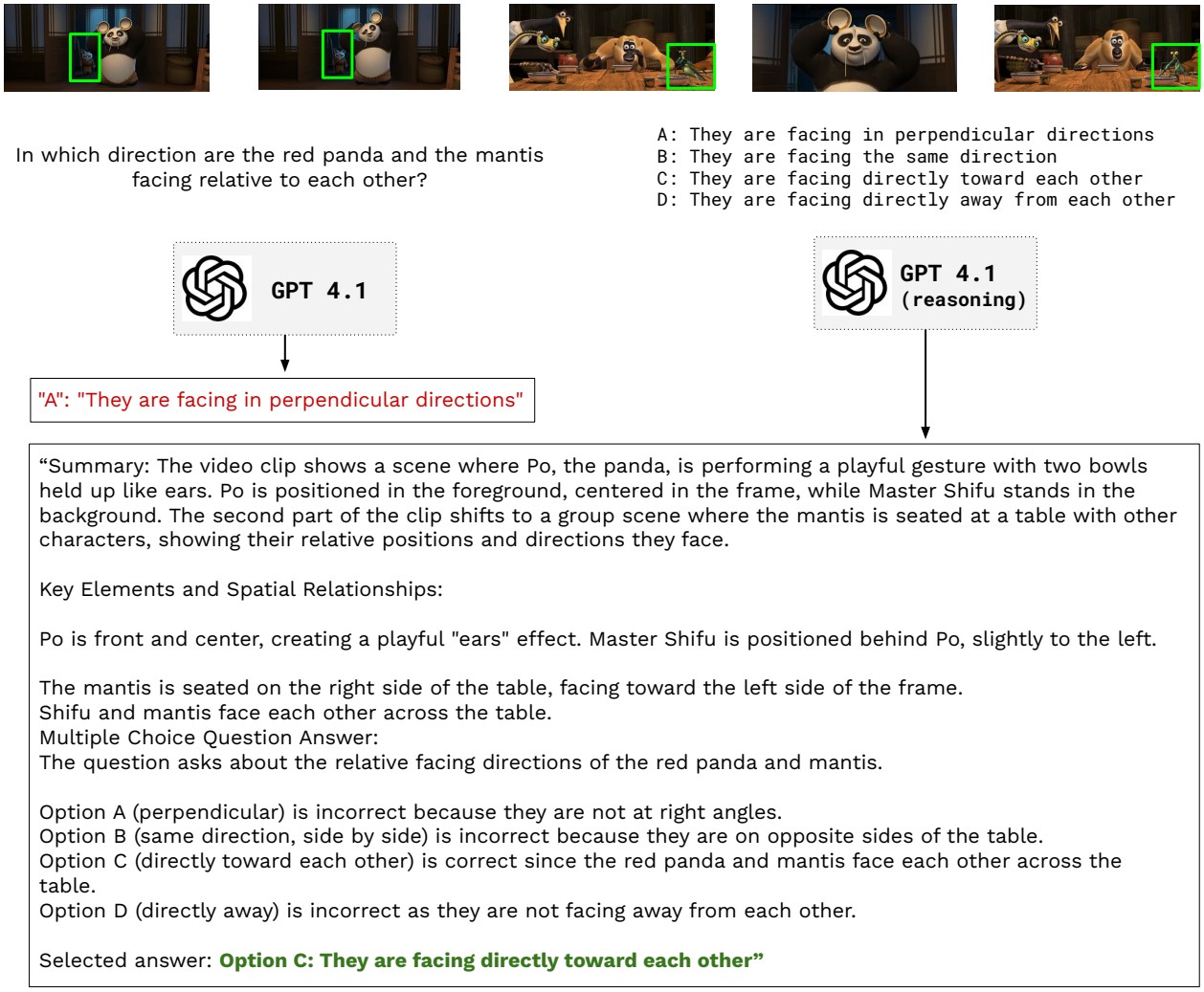


Figure 21. Qualitative example from VRR-QA demonstrating how the reasoning prompt improves GPT-4.1’s performance. Without structured reasoning, the model incorrectly selects Option A. With the reasoning prompt, the model provides a detailed spatial analysis of character positions and orientations, ultimately selecting the correct answer, Option C. This showcases the benefit of guiding the model through spatially grounded reasoning steps.

Table 6. Mean and Standard Deviation Across Categories (in %)

	Lateral Spatial Reasoning	Vertical Spatial Reasoning	Relative Depth and Proximity	Viewpoint and Visibility	Motion & Traj. Dynamics	Motivational Reasoning	Inferred Counting	Physical & Env. Context	Social Interaction & Relations	Avg.
Mean	41.61	47.41	32.71	60.00	50.55	51.71	26.51	42.86	62.07	42.93
Std Dev	0.00	0.25	0.00	1.34	0.00	0.67	1.27	0.00	0.00	0.12

achieved near-ceiling performance on Easy questions and maintained strong performance on Medium ones. Remarkably, even on Hard questions defined by consistent failure across models, the human annotator performed well above chance, achieving over 70% accuracy. This analysis underscores the significant gap between human and model capabilities on complex reasoning tasks. Even non-expert humans demonstrate strong generalization and visual implicit understanding, particularly in scenes that demand spatial, temporal, or motivational reasoning, highlighting the limitations of current state-of-the-art Video LMMs.

Table 7. Human baselines on the Inferred Counting Category.

Expertise Level	Video View Count	Avg Response Time per Question (s)	Accuracy (%)
Non-expert	once	44.5	65.9
Non-expert	multiple	105.6	82.8
Expert	once	36.9	73.7
Expert	multiple	75.6	87.3

F.1. Human baselines for Inferred Counting

The key difficulty with the inferred counting category is that solving the question correctly is a tedious task, requiring back and forth between the frames of the video - to carefully track the number of occurrences across multiple frames, that must be mentally aggregated. During the human baseline benchmarking process the non-expert annotators were asked to only watch the video once, without the ability to scroll back and forth and as a result on the counting task the performance was low. We perform additional experiments, where the non-expert annotators were allowed to scroll back and forth between frames and review the video multiple times. This flexibility significantly improved performance achieving 82.8% while spending more than double time as shown in Table 7. For these questions we also benchmark expert human baseline and find significant improvement over the non-experts.

We would like to note that we were aware of the difficult nature of the inferred counting task during the annotation process itself. As discussed in the main paper, multiple expert annotators reviewed each sample before being added to the dataset. Out of the 58 questions in the Inferred counting category, there was strong agreement among expert annotators:

- 5 Annotators agreed: 49 questions (84.5% of cases)
- 4 Annotators agreed: 5 questions (8.6% of cases)
- 3 Annotators agreed: 4 questions (6.9% of cases)

These results indicate that, while challenging, the questions are well-defined, solvable, and carefully validated requiring real reasoning capabilities.

G. Annotation Tool Interface

As discussed in Section 3.1 in the Main paper, we have designed and built an annotation tool for intuitive and efficient workflow, which allows annotators to follow a structured approach towards formulating multiple choice QA pairs for desired video clip segments. The tool is optimized for fast navigation and efficient verification. Said annotation can be systematically performed by adhering to the following procedure. We have divided the procedure into 4 intuitive sub-procedures which further comprise of 3 steps each. We describe these subprocedures below.

- **Video Download:** Our tool allows the user to input a video URL and download the associated video. The user needs to run backend.py, navigate to the web interface using the generated link, input desired URL and click on the “Download Video" button. The tool downloads the video and shows it in the embedded display.
- **Clip Segment Selection:** The user can put time markers at any desired place using the provided time bar, and viewing the video in the embedded display. The user can pause, rewind and analyze the video frame-by-frame to determine the best possible clip segment suiting to their requirements. Then the user can preview their selected clip segment by clicking the “Preview Selection" button.
- **QA Annotation:** After selecting the desired segment, the user can formulate the appropriate question and type it in the designated space. The tool has provision to input as many choices in the designated spaces as needed, using the “Add Answer Choice" button. After finalizing, the user can click the “Save Annotation button to record their multiple choice QA pair.
- **Verification:** The generated QA pair can be viewed in the mini-display at the bottom. Once the user has recorded all their annotations for the video clip, they can verify said annotations by clicking on the “View Annotations" Tab at the top of the page. There, they can see all the videos annotated in that particular session. Clicking on the “View Annotations" button associated to a video takes the user to the page for that video where all annotations are listed along with the embedded display which can play the relevant clip. Clicking on a question displays all details associated with that question including the relevant timestamps ensuring complete verification.

H. Qualitative Results

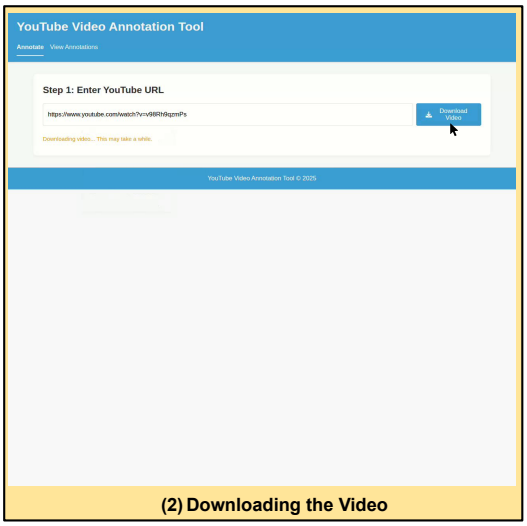
To illustrate the complexity and diversity of visual implicit reasoning required in our benchmark, we present qualitative examples spanning all nine reasoning categories in Figure 23,24,25. For each example, we show the relevant video frames, question, answer choices, correct answer, and model predictions. These examples highlight a wide range of reasoning challenges -from spatial positioning and motion inference to social understanding and inherent explanation. Notably, GPT-O3 demonstrates superior performance in most cases.

H.1. Viewpoint and Visibility

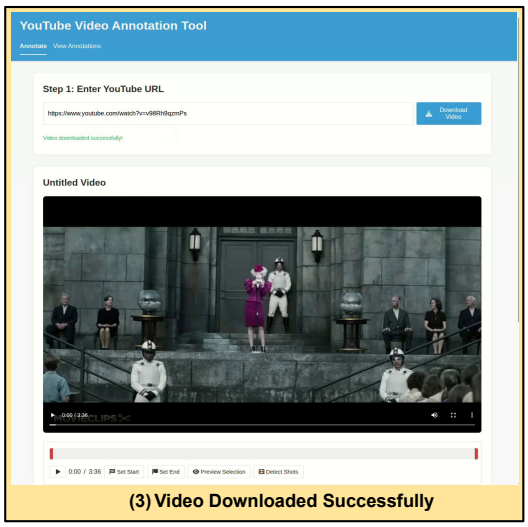
In the Viewpoint and Visibility example shown in Figure 23, only GPT-O3 correctly infers the adopted perspective of the panda character, showcasing its ability to track camera shifts and narrative cues.

Video Download | Clip Segment Selection | QA Annotation | Verification

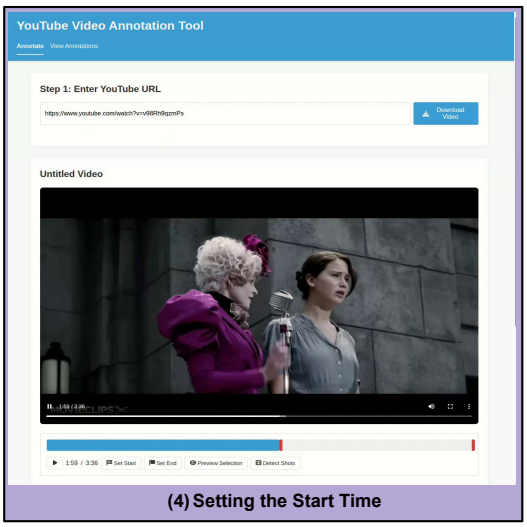
(1) Initial Web Interface



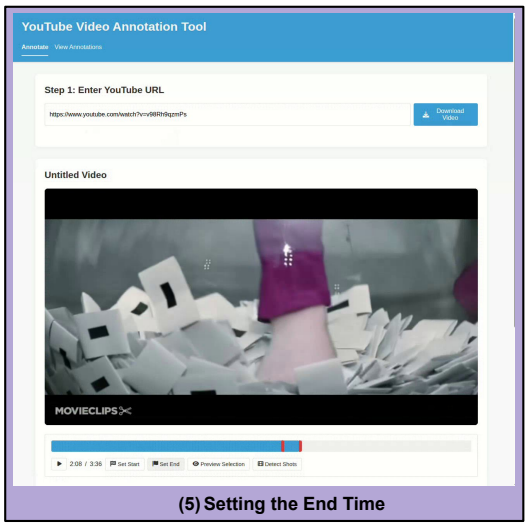
(2) Downloading the Video



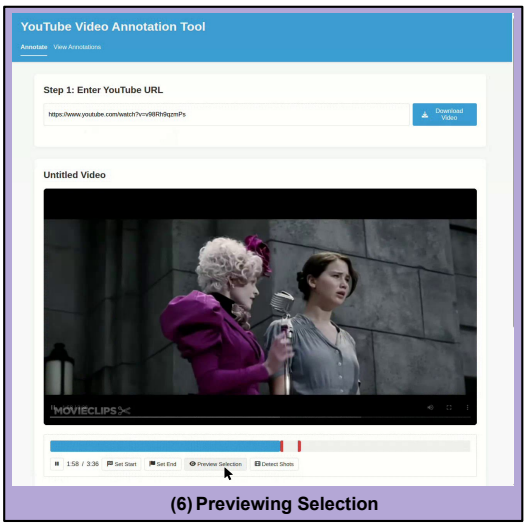
(3) Video Downloaded Successfully



(4) Setting the Start Time



(5) Setting the End Time



(6) Previewing Selection

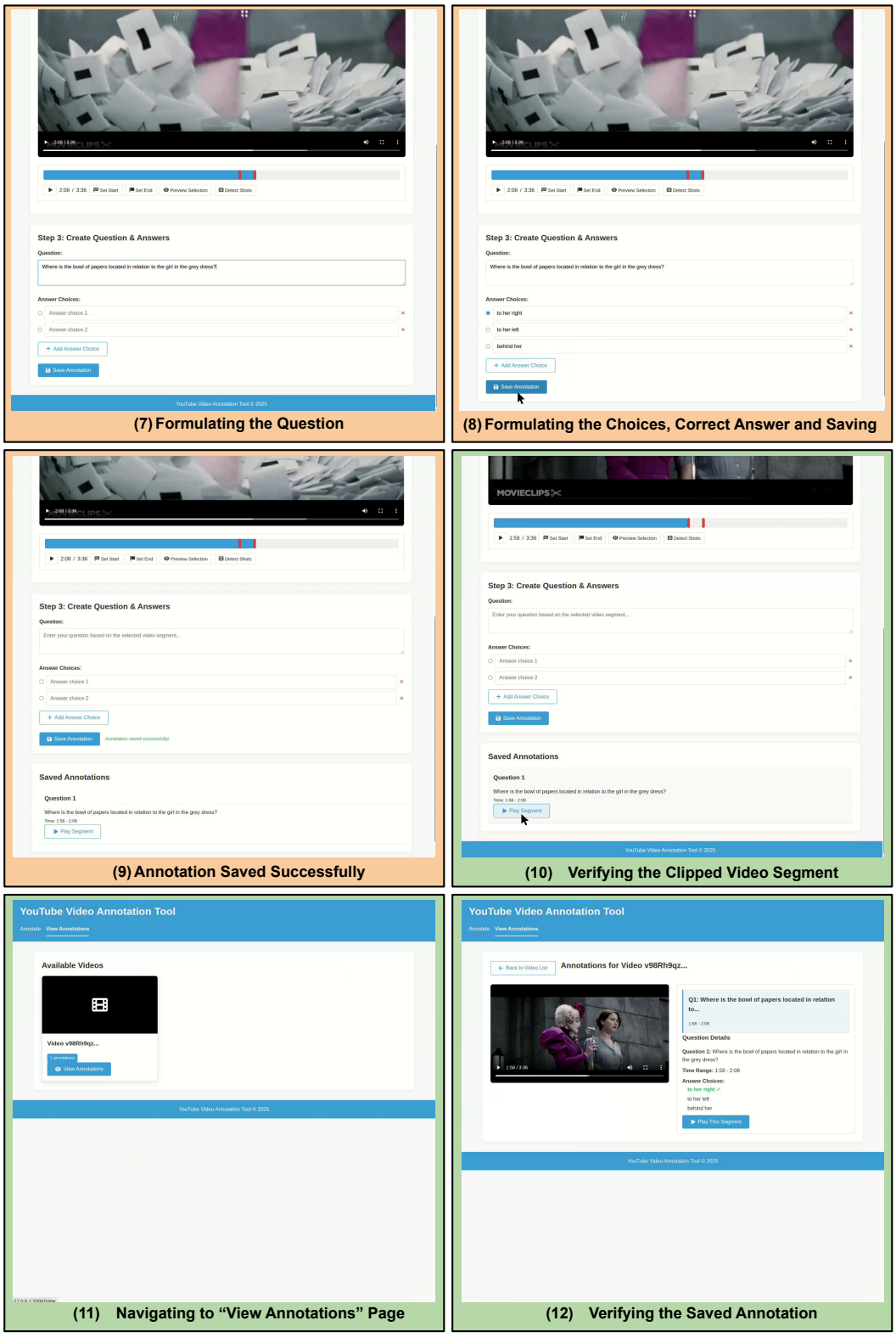


Figure 22. **Schematic illustration of the annotation workflow using our FrameQuiz tool.** The process is organized into **four sub-procedures**: Video Download, Clip Segment Selection, QA Annotation, and Verification. Each sub-procedure contains three steps, resulting in a **12-step pipeline** for generating high-quality multiple-choice QA pairs from video clips. Final annotations are stored locally following verification.

H.2. Physical and Environmental Context

In the Physical and Environmental Context scenario as in Figure 23, GPT-O3 again outperforms others by correctly identifying the white car driven by the woman in black, leveraging spatial cues across frames.

H.3. Vertical Spatial Reasoning

As shown in Figure 23, all models successfully answer a Vertical Spatial Reasoning question involving relative positions in a multi-level scene.

H.4. Relative Depth and Proximity

In contrast, more nuanced categories reveal sharper contrasts in performance as shown in Figure 24. For Relative Depth and Proximity, GPT demonstrates strong spatial inference by accurately localizing characters and interpreting their orientations.

H.5. Lateral Spatial Reasoning

For Lateral Spatial Reasoning as shown in Figure 24, we see that most model get it correct except GPT-4.1.

H.6. Motion and Trajectory Dynamics

In the Motion and Trajectory Dynamics example shown in Figure 24, most models correctly track the direction of movement, though GPT-O3 misjudges the path—suggesting sensitivity to camera motion.

H.7. Social Interaction and Relationships

The Social Interaction and Relationships case shown in Figure 25, involving subtle facial cues and body language, is correctly answered only by GPT-O3 and GPT-4.1, reflecting their advanced multimodal understanding.

H.8. Inferred Counting

For Inferred Counting shown in Figure 25, models struggle to aggregate information across frames, with GPT-O3 and Qwen-VL identifying the correct number, while GPT-4.1 undercounts.

H.9. Motivational Reasoning

Finally, in the Motivational Reasoning example in Figure 25, GPT-O3 and GPT-4.1 correctly attribute the escape of the rats to being discovered, while others fail to connect with the relevant event. These examples collectively highlight the diversity and difficulty of visual implicit reasoning tasks in `VRR-QA`.

At the end of the clip, the camera adopts the point of view of a character. Which of the following animals is that character?

A: Tiger
B: Panda
C: Snake
D: Monkey

B: Panda

Viewpoint and visibility

GPT 03 (reasoning)	B: Panda
GPT 4.1	C: Snake
Qwen-VL	C: Snake
LLaVA-OneVision	D: Monkey

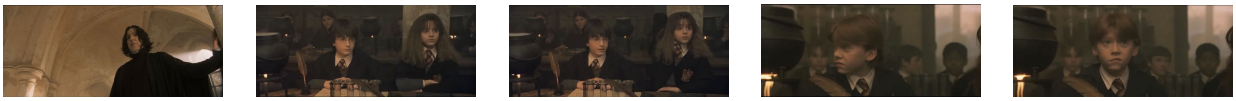
Which car is being driven by the woman in black?

A: Truck
B: Black car
C: White car
D: Police car

C: White car

Physical and environmental context

GPT 03 (reasoning)	C: White car
GPT 4.1	B: Black car
Qwen-VL	B: Black car
LLaVA-OneVision	B: Black car



Where is the redhead boy located relative to the man with the long black hair and black clothes?

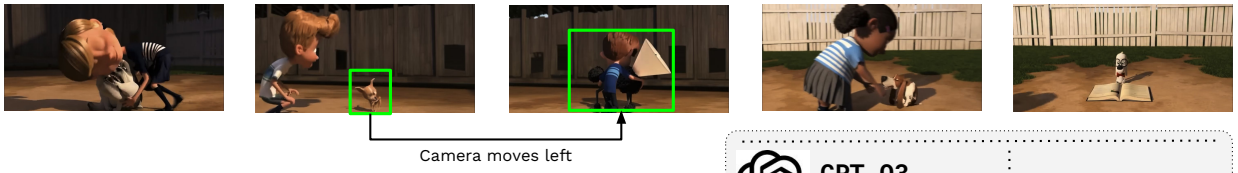
A: In front and below
B: Behind and above
C: Behind and below
D: In front and above

A: In front and below

Vertical Spatial Reasoning

GPT 03 (reasoning)	A: In front and below
GPT 4.1	A: In front and below
Qwen-VL	A: In front and below
LLaVA-OneVision	A: In front and below

Figure 23. More Qualitative VRR-QA examples, targeting distinct visual implicit-reasoning dimensions.



In the camera frame of reference, what is the location of the poodle relative to the chihuahua?

A: Left
B: In front
C: Behind
D: Right

D: Right

Relative depth and proximity

GPT 03 (reasoning)	D: Right
GPT 4.1	D: Right
Qwen-VL	C: Behind
LLaVA-OneVision	A: Left

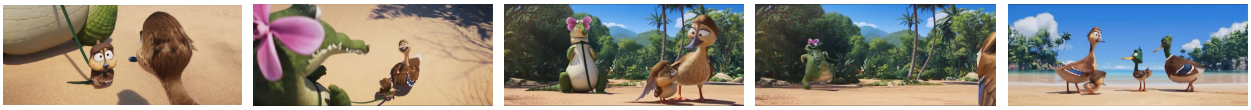
In which direction are the man in orange and the turtle looking relative to each other?

A: Perpendicular
B: Same direction
C: Away from each other
D: Towards each other

D: Towards each other

Lateral spatial reasoning

GPT 03 (reasoning)	D: Towards each other
GPT 4.1	C: Away from each other
Qwen-VL	D: Towards each other
LLaVA-OneVision	D: Towards each other



In which direction is the crocodile running relative to the green ducks?

A: Away from them
B: Perpendicular, left to right
C: Perpendicular, right to left
D: Towards them

A: Away from them

Motion and trajectory dynamics

GPT 03 (reasoning)	C: Perpendicular, left to right
GPT 4.1	A: Away from them
Qwen-VL	D: Towards them
LLaVA-OneVision	A: Away from them

Figure 24. More Qualitative VRR-QA examples, targeting distinct visual implicit-reasoning dimensions.

Why are the three old men looking at the operating men in such a manner?

A: They do not approve of the operating men
B: They do not care about the situation
C: They completely approve of the operating men
D: They are excited about the operation

A: They do not approve of the operating men

Social Interaction and Relationships

GPT 03 (reasoning)	A: Do not approve
GPT 4.1	A: Do not approve
Qwen-VL	C: Completely approve
LLaVA-OneVision	C: Completely approve

How many purple balls appear in the clip?

A: 5 C: 4
B: 3 D: 2

B: 3

Inferred Counting

GPT 03 (reasoning)	C: 3
GPT 4.1	D: 2
Qwen-VL	C: 3
LLaVA-OneVision	C: 3

Why did the rats escape?

A: To avoid mouse traps
B: They were found by the old lady
C: They didn't find what they were looking
D: To avoid being found by the old lady

B: They were found by the old lady

Causal and Motivational Reasoning

GPT 03 (reasoning)	B: They were found by the...
GPT 4.1	B: They were found by the...
Qwen-VL	D: To avoid being found by the ...
LLaVA-OneVision	D: To avoid being found by the ...

Figure 25. More Qualitative VRR-QA examples, targeting distinct visual implicit-reasoning dimensions.