Motion-Grounded Video Reasoning: Understanding and Perceiving Motion at Pixel Level (Supplementary Materials)

Andong Deng[1], Tongjia Chen[2], Shoubin Yu[3], Taojiannan Yang[4], Lincoln Spencer[1],
Yapeng Tian[5], Ajmal Saeed Mian[2], Mohit Bansal[3], Chen Chen[1]

[1]CRCV, University of Central Florida, [2]University of Western Australia,
[3]UNC, Chapel Hill, [4]Amazon Web Services, [5] The University of Texas at Dallas

1. Overview

The following appendix is structured to provide supplementary information about our GROUNDMORE dataset, its annotations, and representative examples. We aim to present a comprehensive view of the statistical analysis, annotation process, and key insights that further elaborate on the main text. The appendix is divided into the following sections:

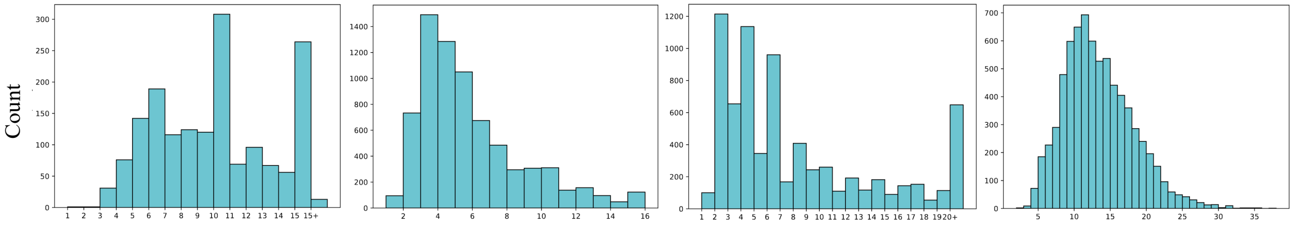
- Section 2 offers detailed statistical insights into the types of questions and scenes captured in our dataset, as well as an analysis of the distribution of objects, verbs, and word clouds in the question annotations.
- Section 3 provides detailed information about the annotation process, including the types of motion-related expressions, the generation of questions through large language models, and the quality validation procedures.
- Section 4 showcases a set of representative examples from GROUNDMORE, illustrating the richness of the dataset through diverse scenes, objects, and questions.
- Section 5 discusses the necessity of including implicit reasoning, highlighting the importance of capturing nuanced motion-grounded video reasoning.
- Section 6 showcases the impact of object numbers on the dataset's performance.
- Section 7 provides details of our MoRA baseline.
- Section 8 demonstrates the qualitative performance of current video LLMs in the two-stage baseline settings.
- Section 10 outlines the limitations of the current version of GROUNDMORE and discusses future work.
- Section 11 outlines the ethical considerations, privacy concerns, and licensing terms associated with GROUNDMORE.

2. GROUNDMORE Statistics

Our GROUNDMORE contains 1,715 videos 7,577 questions and 249K object masks as well as 3,942 objects. And the average video clip duration is 9.61 seconds. GROUNDMORE is split into 1,333 training and 382 test videos. As shown in Figure 1a, most of the clips have a duration between 5s and 15s, which is long enough to include sufficient motion semantics. This range ensures that the clips capture complete actions and interactions, providing a rich context for question formulation. In Figure 1b, it is evident that most motions in GROUNDMORE have a duration from 2s to 6s, highlighting the challenge of temporal localization in our dataset. These short-duration motions require precise temporal understanding and segmentation, adding to the complexity of the GROUNDMORE. Besides, the average motion (segment) ratio in each video clip is 51%. As seen in Figure 1c, for most clips, the number of questions is more than 2, with a significant number having up to 4 or more questions. This indicates that GROUNDMORE provides a diverse set of questions per clip, ensuring a comprehensive evaluation of the clip's content. It also implies that each clip contains multiple distinct motion semantics that warrants varied questioning. In Figure 1d, the distribution shows that most questions are sufficiently long, typically ranging from 7 to 15 words. This length reflects the complexity and detail required in the questions, underscoring the difficulty level of our GROUNDMORE. The substantial word count in questions ensures that they are descriptive and context-rich, further challenging the systems to provide accurate and detailed responses.

2.1. Question and Scene Type.

We provide detailed statistics of GROUNDMORE in this section, including the distribution of question types, scene



(a) Clip Length(s) Distribution (b) Motion Length(s) Distribution (c) Question Number Distribution (d) Question Word Length Distribution

Figure 1. Statistics of GROUNDMORE benchmark.

types, objects, and verbs that appear in our question annotation, etc. As shown in Figure 2a, the **Descriptive** questions constitute the highest proportion at 29.7%, followed closely by **Causal** questions at 28.5%. **Sequential** questions make up 21.7% of the total, while **Counterfactual** questions are the least common, accounting for 20.2%. Our GROUNDMORE shows a balanced distribution w.r.t. question type. Regarding scene type distribution (Figure 2b), **family** scenes dominate with a significant 35.1% share, slightly higher than the **ball game** scenes, which account for 32.7%. **Animal** scenes are also well-represented at 25.4%, whereas **outdoor activity** scenes are relatively rare, comprising only 6.8% of the total scenes in our GROUNDMORE.

2.2. Object Word Distribution.

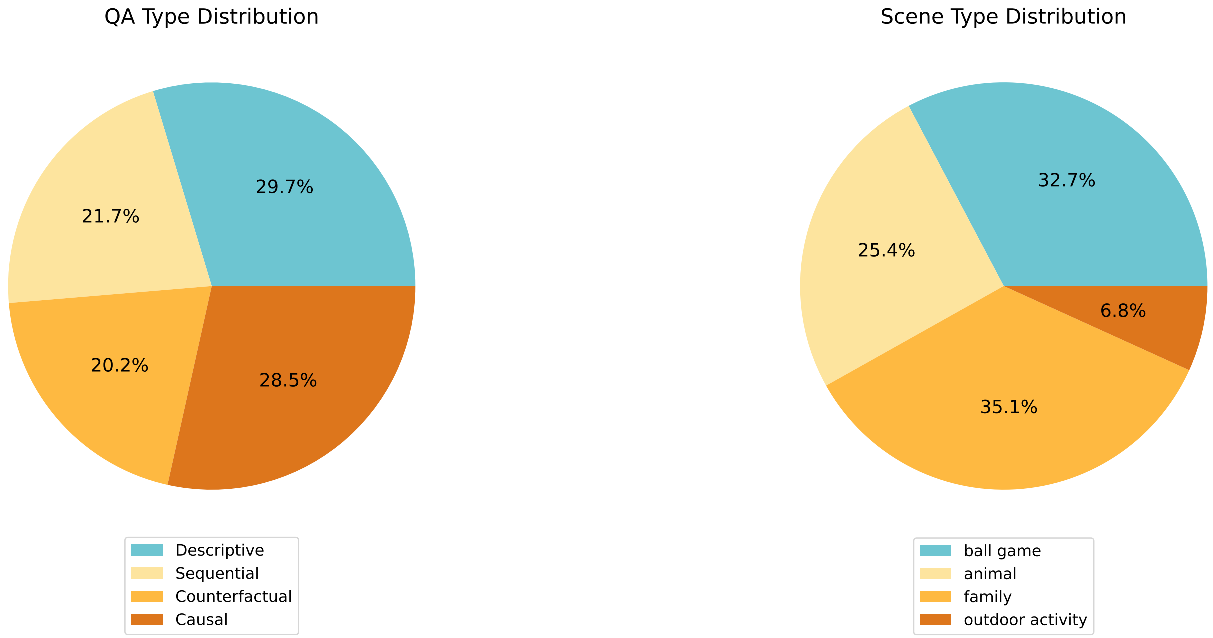
Figure 3 illustrates the top 30 most frequent objects in our GROUNDMORE questions. We categorize these objects into six parent categories: sports equipment, people, animals, furniture, household items, and food, reflecting common items in daily life. As can be seen from the figure, *ball* is the most frequently occurring object, followed by *man*, *dog*, *basketball*, and *girl*. This prevalence is aligned with the high proportion of sports and family videos in our GROUNDMORE, as indicated in Figure 2b. The dominance of sports equipment such as *ball* and *basketball* correlates with the 32.7% share of ball game scenes. Similarly, the frequent appearance of *man*, *girl*, and *woman* objects is consistent with the substantial 35.1% of family scenes, where people are commonly depicted. Additionally, animals like *dog* and *cat* are prominent due to their significant 25.4% representation in animal scenes. The distribution of these objects highlights the diverse and realistic contexts covered in our GROUNDMORE, ensuring a comprehensive evaluation of various question types and scene contexts.

2.3. Verb Distribution.

Another key component of our GROUNDMORE is the verb in the motion-related questions. In Figure 4, we present the top 20 most frequent verbs across different scene types, represented by distinct colors. The verb *use* has the highest overall proportion, reflecting its ubiquity in daily activities, with a notable presence in family scenes, as well as significant occurrences in animal and ball game scenes. The verb *dribble* ranks second and is exclusively found in ball game videos, highlighting its specificity to that context. The verb *move* is also prominent, appearing across all four scene types, indicating its general applicability in various contexts. Verbs such as *hold*, *open*, and *put* are more frequently observed in family videos, underscoring their relevance to everyday domestic activities. In contrast, *accelerate* and *shoot* are predominantly associated with ball game scenes, which is consistent with the dynamic nature of these activities. Besides, the distribution of verbs shows a more balanced pattern compared to the object distribution, reflecting a diverse range of actions across different contexts. For instance, while *throw* and *pass* are mainly seen in ball game scenes, verbs like *push* and *grab* are well-represented in both family and ball game contexts. This balanced distribution underscores the comprehensive nature of our GROUNDMORE, capturing a wide array of activities and interactions within various scene types.

2.4. Word Cloud Visualization.

Moreover, we leverage the word cloud of the top 100 words that appear in our GROUNDMORE questions. The word cloud in Figure 5 provides a visual representation of the most frequently occurring words. We can observe that common objects like *"dog"*, *"cat"*, and *"ball"* are prominently featured, which aligns with the object distribution shown in Figure 3. These objects are integral to many of the scenes and questions, reflecting their high frequency in the dataset. In addition to objects, prepositions closely related to motion, such as *"down"*, *"out"*, and *"with"*, are also prevalent. This is consistent with the verb distribution illustrated in Figure 4, where actions often involve directional or positional changes, necessitating the use of these prepositions. Furthermore, adverbs such as *"before"* and *"after"* appear frequently, indicating their importance in describing temporal relationships within the scenes. These temporal adverbs are essential in forming questions related to sequences and causality, which are common in descriptive and sequential question types. Overall, the word cloud highlights the interconnected nature of objects, verbs, and descriptive lan-



(a) QA Type Distribution

(b) Scene Type Distribution

Figure 2. Question and Scene Type Distribution of GroundMoRe.

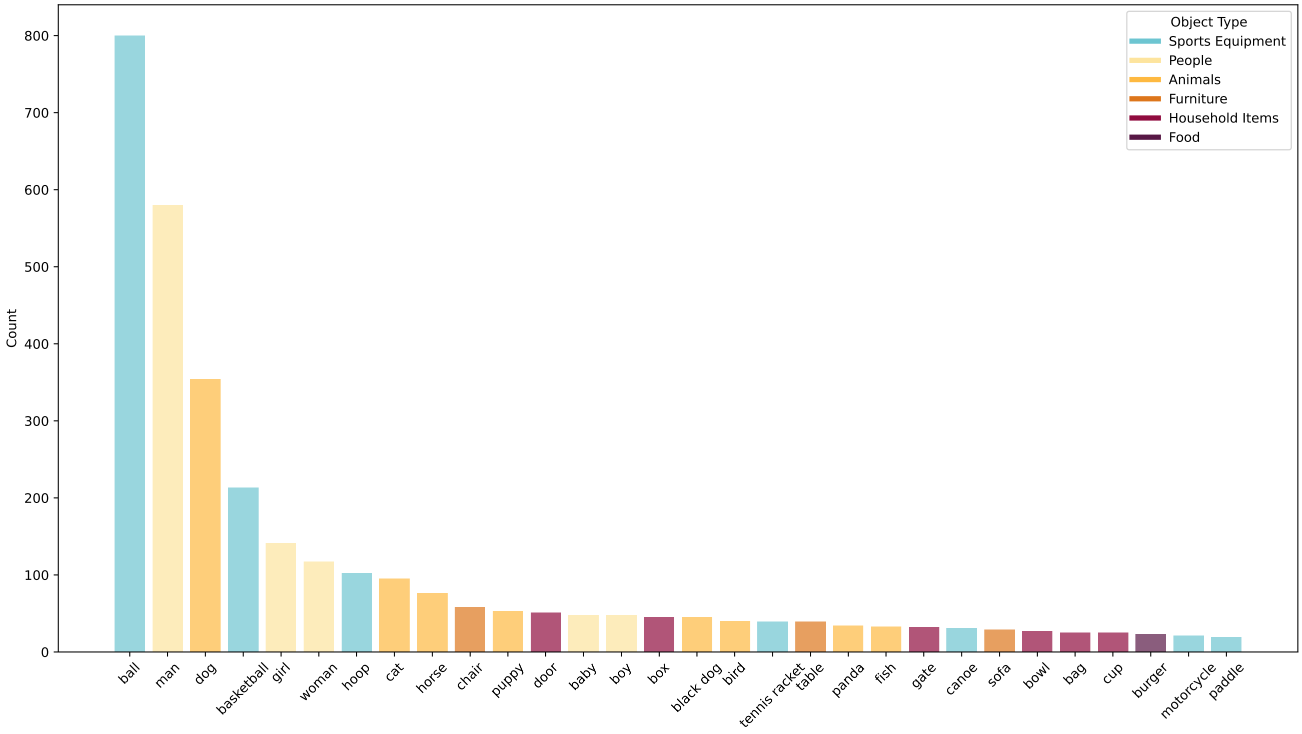


Figure 3. Object distribution of GroundMoRe.

guage within our GroundMoRe, demonstrating the comprehensive coverage of various elements that contribute to the complexity and richness of the dataset.

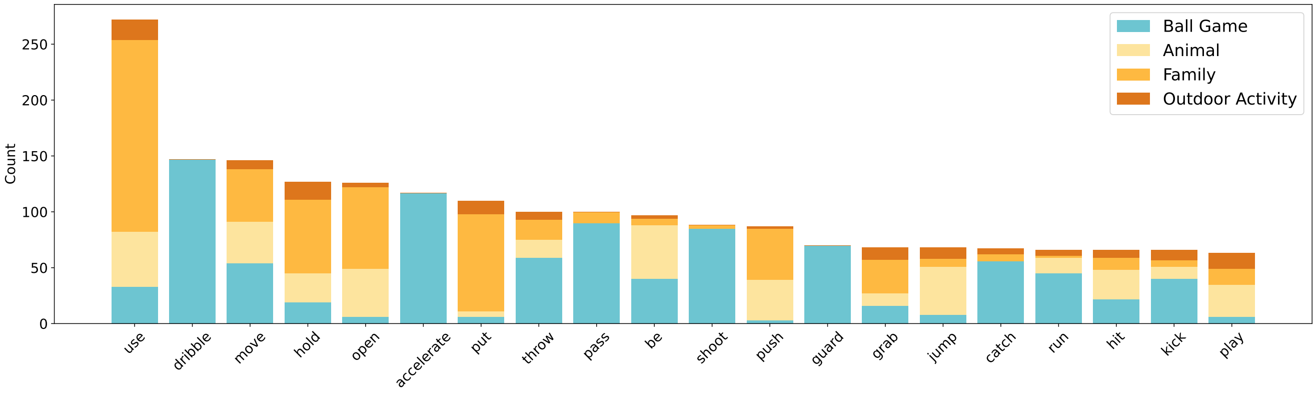


Figure 4. Verb distribution of the motion concepts in GROUNDMORE.

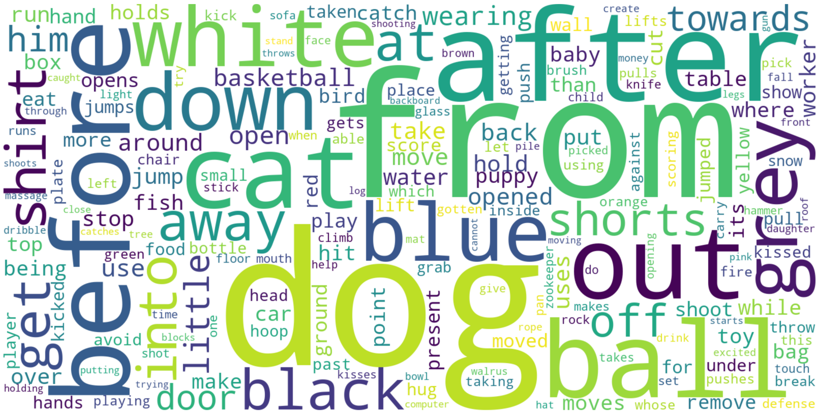


Figure 5. Word cloud of the top 100 words in the question annotation in our GROUNDMORE dataset.

2.5. Sankey Diagram for Interaction.

We provide the Sankey diagram of our proposed GROUNDMORE in Figure 6, which illustrates the interactions within our GROUNDMORE. In this diagram, the elements on the left side represent different initial categories of objects or entities involved in interactions (e.g., People_A, Animals_A, Sport Equipments_A), while the elements on the right side represent the resulting categories of objects or entities after interactions (e.g., People_B, Animals_B, Sport Equipments_B). From the diagram, we can see that human-involved interactions (People_A) have the highest proportions, flowing prominently into both sports and family categories on the right. This is consistent with the scene type distribution (Figure 2b), where sports and family scenes were among the most prevalent. Similarly, the frequent appearance of sports equipment, animals, and household items in both left and right categories aligns with the object distribution shown in Figure 3. The Sankey diagram validates that our GROUNDMORE is well-suited for motion and interaction understanding. It demonstrates the comprehensive coverage of various interactions, emphasizing the importance of human involvement and the diverse range of objects and entities engaged in these interactions. This rich interplay of elements ensures that GROUNDMORE could serve as a robust benchmark for evaluating motion understanding in complex video scenarios.

3. Annotation Details

We recruited a team of 15 computer science students with experience in video understanding as our paid annotators to ensure high-quality annotations, 10 of them were assigned to question annotation and the rest focused on mask. As mentioned in Section **??**, the question annotation is constituted of two stages: 1) motion-related expression annotation; and 2) LLM-assisted QA generation. And we resort to

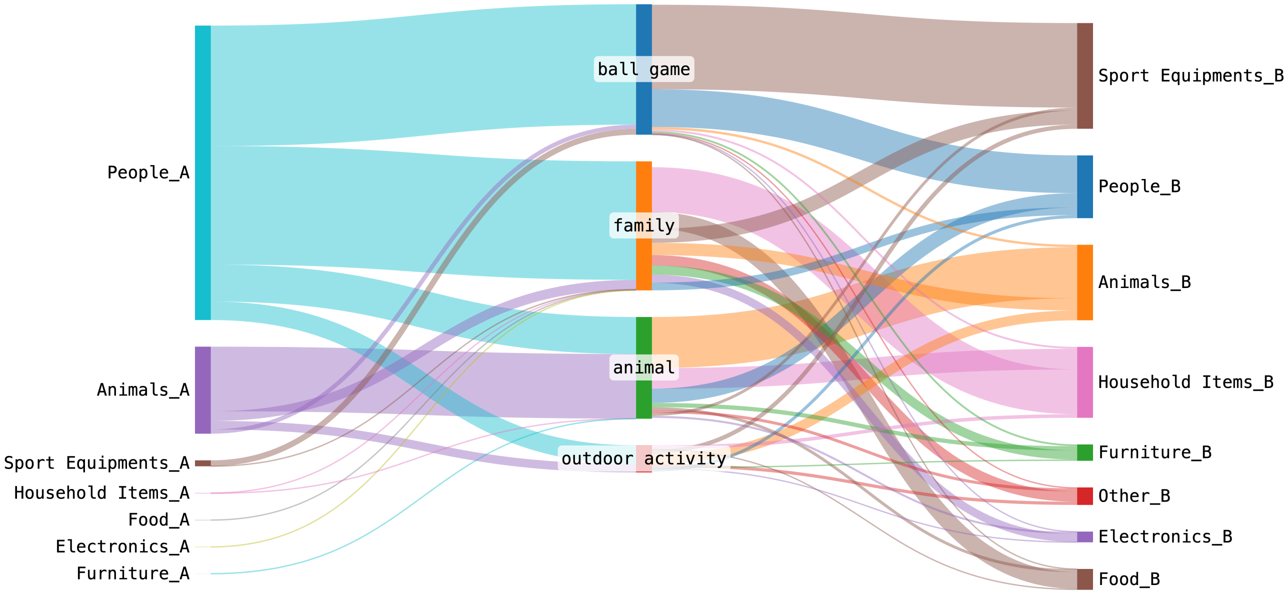


Figure 6. Sankey diagram on the interaction of our GROUNDMORE.

XMem++ [1] as our semi-automated mask annotation tool. The interface is shown in Figure 7.

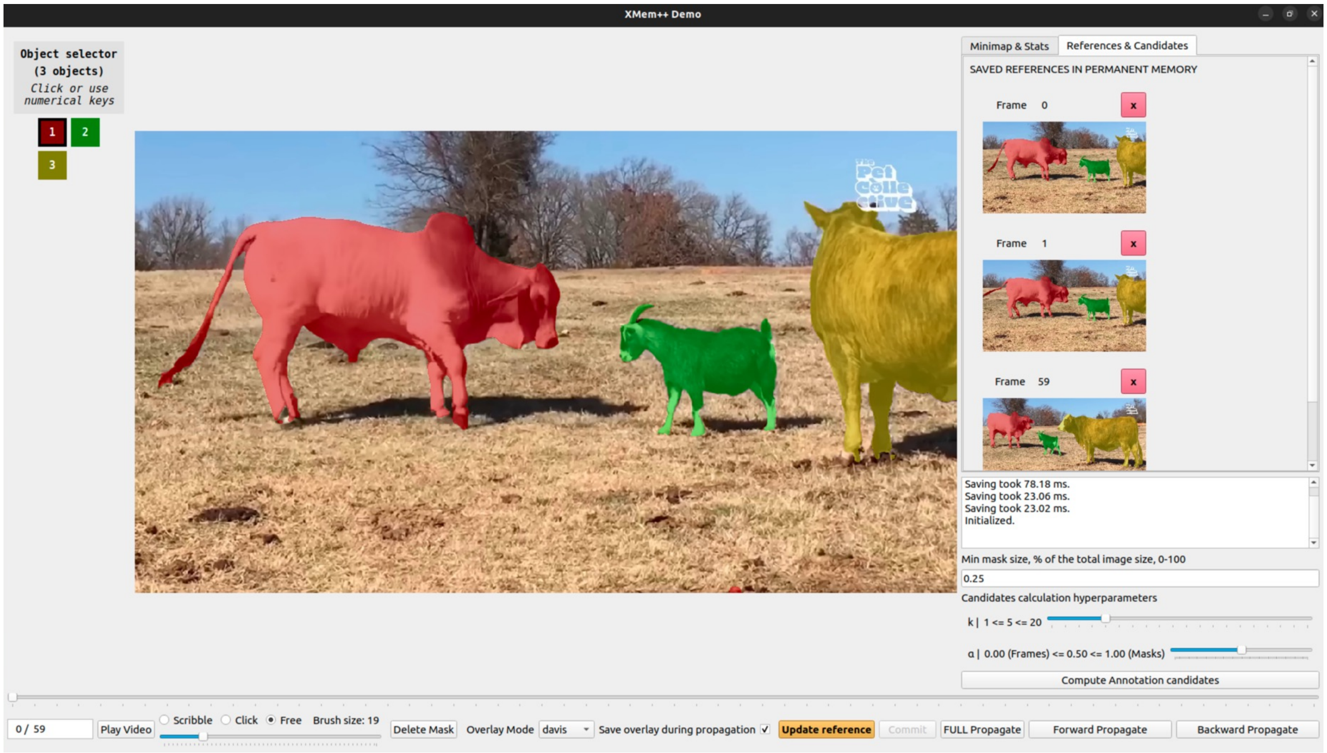
3.1. Expression Annotations.

Expression annotation is to annotate the ongoing motions or events in a given video. We define three different expression types: interaction-causal, interaction-temporal, and descriptive expression. The motions that can be described within these three types of expressions could generally cover most of the daily scenarios. The interaction-causal expression has the format ¡obj_A, motion, obj_B, to do something¿ which depicts a scene where the motion takes place based on some hidden motivations. For instance, as shown in the first row in Figure 8, the causal-driven expression of this case elucidates the motivation behind the motion of *passed the knife to the man in the grey shirt* is to *let him cut the watermelon*. Interaction-temporal expressions, following the format ¡obj_A, motion, obj_B, before/after another motion¿, describe the chronological relations between temporally adjacent actions, which enables motion understanding based on temporal conditions. As shown in the second row in Figure 8, *the man in black* performs two consecutive actions, *get rid of the defense from the man in white* and *shot the basketball*. In most similar cases, the temporally adjacent motion not only has temporal relations but also has cause-and-effect; therefore, such expressions could help analyze the existence of one motion based on another. The third one is the descriptive expression, which contains either general scene description or motion-based abstract attributes (e.g., *energetic, naughty, faster, etc.*). As shown in the last row in Figure 8, *consumed more energy* could be viewed as an abstract attribute represented by the fact that the man is doing massage for the dog. Given this expression type, the models are required to perform both spatiotemporal reasoning and commonsense reasoning to understand the scene content.

3.2. Question Annotations.

As shown in Figure 9, we specifically design the prompt to leverage the text generation ability of GPT-4o. For each expression, we first specify the target objects that would be annotated during the mask annotation. For instance, in the first row of Figure 8, considering the bidirectional nature of an interaction, we will ask GPT-4o to generate questions for both *the man in the yellow shirt* and *the knife* by providing their object ID: {"1": "the man in the yellow shirt", "2": "the knife"}.

Causal questions are generated from expressions of interaction-causal expressions. Due to the bidirectional nature of the interactions, we will generate questions targeting both subject and object. For instance, for the expressions in the first row of Figure 8, we will generate questions as follows: *Who passed the knife to the man in the grey shirt to let him cut the watermelon?* and *What did the man in the yellow shirt pass to the man in the grey shirt to let him cut the watermelon?* We generate questions for both the subject and the object of motion to ensure complete spatial context reasoning. **Sequential** questions are generated from interaction-temporal expressions. Similarly, since it is also interaction-related, we will generate two questions for each expression as shown in the middle



```
Saving took 78.18 ms.
Saving took 23.06 ms.
Saving took 23.02 ms.
Initialized.
```

Min mask size, % of the total image size, 0-100: 0.25

Candidates calculation hyperparameters

k | 1 <= 5 <= 20

α | 0.00 (Frames) <= 0.50 <= 1.00 (Masks)

0 / 59

○ Scribble ○ Click ● Free Brush size: 19

Overlay Mode: davis

Save overlay during propagation ☑

Figure 7. Annotation Interface of XMem++.

row of Figure 8. **Counterfactual** questions are also generated from interaction-temporal expressions. But it focuses on those scenarios where temporal-adjacent motions have cause-and-effect. For example, in the middle row of Figure 8, the fact that **the man in black got rid of the defense from the man in white** is a prerequisite that he could perform a jump shot. Therefore, the questions can be as follows: *Who needs to be got rid of defense from by the man in black or he cannot shoot the basketball?* and *What cannot be shot if the man in black did not get rid of the defense from the man in white?* **Descriptive** questions are simply converted from descriptive questions as shown in the third row of Figure 8. It will follow the same rule aforementioned if an interaction is involved.

3.3. Quality Validation.

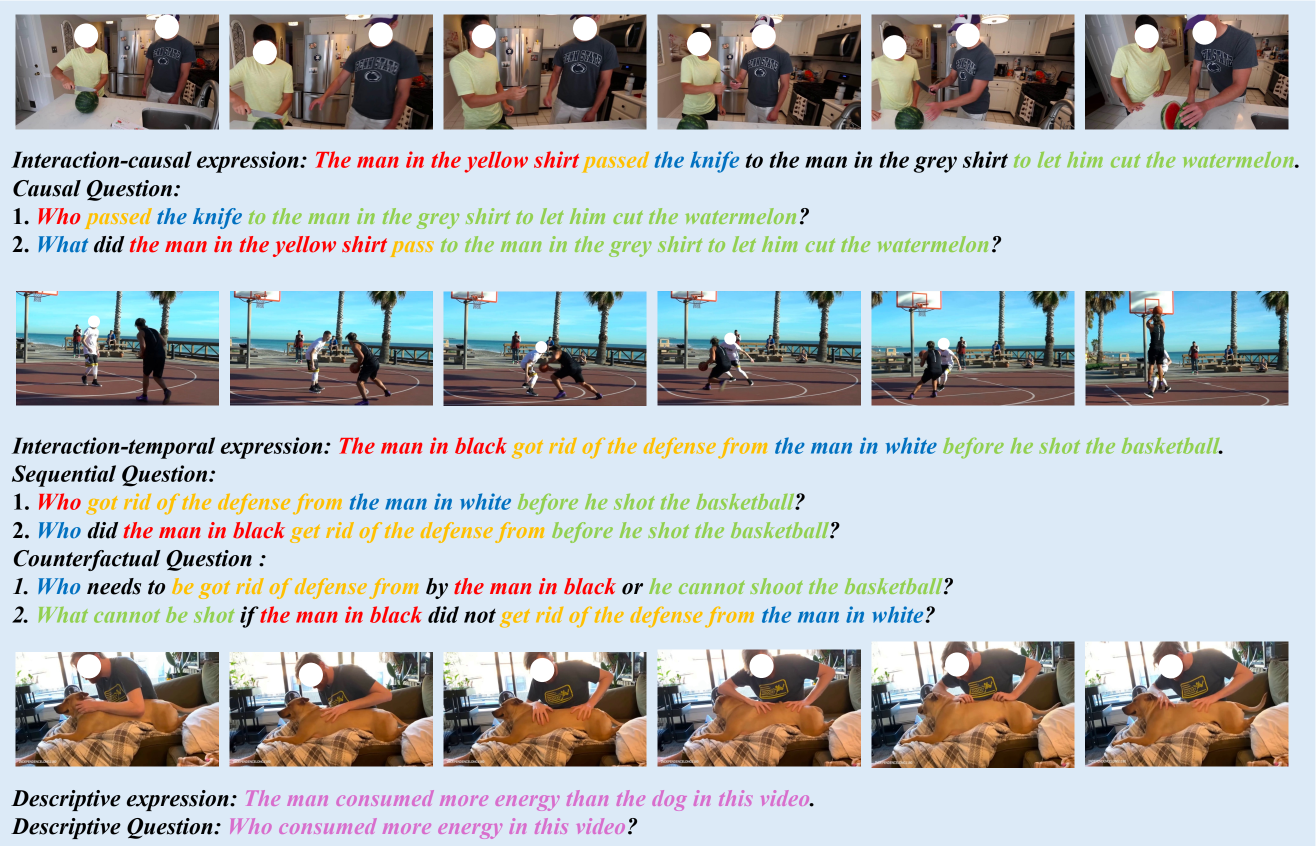
After the generation of questions by GPT-4o, the resulting questions and their corresponding answers will be distributed to different annotators in the same question annotation group for quality control. Importantly, these annotators will not have been involved in the original expression annotation to ensure objectivity. The annotators will be instructed to perform the following steps:

1. **Check relevance**: Verify whether the generated question logically aligns with the current video context and scene.
2. **Answer validation**: Answer the question independently and compare the response with the original annotation. The goal is to ensure consistency between the generated answer and the initial annotation.
3. **Single-object validation**: Confirm that the answer references a single object when appropriate. If the answer references multiple objects and is not explicitly required, the question-answer pair should be revised.

If any issues are identified with the question or the answer, the annotator is required to update the question-answer pair. For example, if the generated question is *"Who is playing baseball?"* and the answer is *"The boy and the dog"*, the annotator should revise the pair to better reflect clarity and context, such as: *"Who is playing baseball with the dog? The boy."* and *"Who is playing baseball with the boy? The dog."*

Similarly, the masks will undergo a quality check by different annotators within the mask annotation group. The first task for the reviewer is to assess whether the mask corresponds to the object(s) indicated in the answer. If a mismatch is found between the mask and the answer, a third annotator will be consulted to provide an additional opinion. The final decision on whether to accept or reject the mask will be based on the majority decision. Mismatched masks will be discarded entirely since re-annotating from scratch is typically more efficient than attempting to fix them.

If the masks match the answer, the annotator will pro-



Interaction-causal expression: *The man in the yellow shirt passed the knife* **to the man in the grey shirt** *to let him cut the watermelon*.
Causal Question:
1. *Who passed the knife to the man in the grey shirt to let him cut the watermelon?*
2. *What* **did** *the man in the yellow shirt pass to the man in the grey shirt to let him cut the watermelon?*

Interaction-temporal expression: *The man in black got rid of the defense from the man in white before he shot the basketball*.
Sequential Question:
1. *Who got rid of the defense from the man in white before he shot the basketball?*
2. *Who* **did** *the man in black get rid of the defense from before he shot the basketball?*
Counterfactual Question :
1. Who **needs to** *be got rid of defense from* **by** *the man in black* **or** *he cannot shoot the basketball?*
2. What cannot be shot **if** *the man in black* **did not** *get rid of the defense from the man in white?*

Descriptive expression: *The man consumed more energy than the dog in this video*.
Descriptive Question: *Who consumed more energy in this video?*

Figure 8. Question generation examples for different types of motion-related expressions.

ceed to evaluate the overall quality, focusing on any potential missing regions, incorrect regions, or other inaccuracies. In the end, all mask-answer pairs must meet predefined quality standards to ensure their validity for downstream tasks.

3.4. Annotator Compensation.

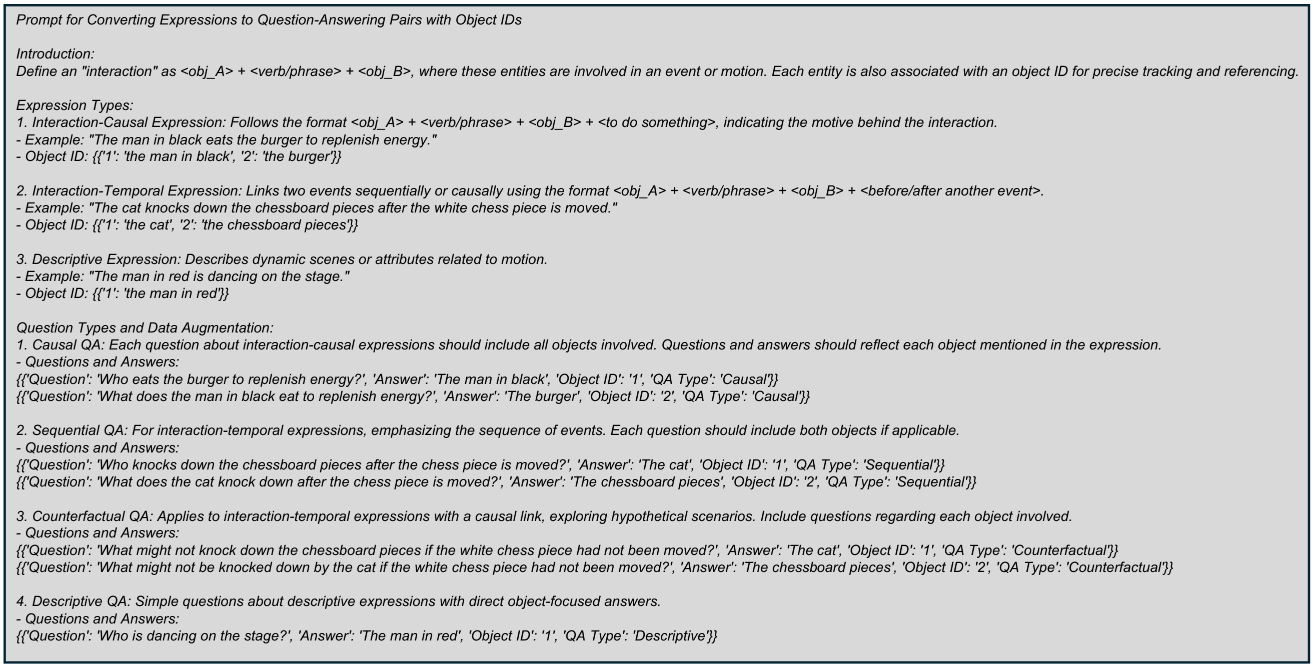
We compensated the question annotators $0.50 per expression and paid $1.00 per clip for mask annotations. Additionally, during the quality validation process, we provided an extra compensation of $0.20 per instance (a question-clip pair).

4. GroundMoRe Examples

We provide additional visualizations of our proposed GroundMoRe in Figure 10. As shown, our GroundMoRe requires advanced motion reasoning abilities in diverse scenarios. As illustrated in the fourth row of the figure, the question ``What might not be held by the man if it had not been unwrapped from the paper?" requires the model to reason the wrapping relationship between ``the man", ``the paper" and ''the piston" as well as the causal connections in the challenging *counterfactual* setting. Additionally, we can observe from the case in the seventh row that our GroundMoRe includes spatiotemporal grounding context as well as motion-related attributes understanding. The answer to the question ``Who might not have fallen into the blue cushion on the wall if he had not tripped while trying to defend?" can only be determined at the end of the video clip. For the question ``Who is the more offensive player?", the model must infer motion-based implicit attributes from the video sequence, demonstrating a strong need for world-level commonsense reasoning ability. These details further demonstrate the complex motion reasoning context of our GroundMoRe.

Besides, the raw videos are processed into individual frames and stored in a folder named with the format ”youtube_id_start-time_end-time”. The annotation is in a JSON format, structured as follows:

```
{
  "questions": {
    "1": {
      "action_end": "0:15",
```



Prompt for Converting Expressions to Question-Answering Pairs with Object IDs

Introduction:
Define an "interaction" as <obj_A> + <verb/phrase> + <obj_B>, where these entities are involved in an event or motion. Each entity is also associated with an object ID for precise tracking and referencing.

Expression Types:
1. Interaction-Causal Expression: Follows the format <obj_A> + <verb/phrase> + <obj_B> + <to do something>, indicating the motive behind the interaction.
- Example: "The man in black eats the burger to replenish energy."
- Object ID: {{'1': 'the man in black', '2': 'the burger'}}

2. Interaction-Temporal Expression: Links two events sequentially or causally using the format <obj_A> + <verb/phrase> + <obj_B> + <before/after another event>.
- Example: "The cat knocks down the chessboard pieces after the white chess piece is moved."
- Object ID: {{'1': 'the cat', '2': 'the chessboard pieces'}}

3. Descriptive Expression: Describes dynamic scenes or attributes related to motion.
- Example: "The man in red is dancing on the stage."
- Object ID: {{'1': 'the man in red'}}

Question Types and Data Augmentation:
1. Causal QA: Each question about interaction-causal expressions should include all objects involved. Questions and answers should reflect each object mentioned in the expression.
- Questions and Answers:
{{'Question': 'Who eats the burger to replenish energy?', 'Answer': 'The man in black', 'Object ID': '1', 'QA Type': 'Causal'}}
{{'Question': 'What does the man in black eat to replenish energy?', 'Answer': 'The burger', 'Object ID': '2', 'QA Type': 'Causal'}}

2. Sequential QA: For interaction-temporal expressions, emphasizing the sequence of events. Each question should include both objects if applicable.
- Questions and Answers:
{{'Question': 'Who knocks down the chessboard pieces after the chess piece is moved?', 'Answer': 'The cat', 'Object ID': '1', 'QA Type': 'Sequential'}}
{{'Question': 'What does the cat knock down after the chess piece is moved?', 'Answer': 'The chessboard pieces', 'Object ID': '2', 'QA Type': 'Sequential'}}

3. Counterfactual QA: Applies to interaction-temporal expressions with a causal link, exploring hypothetical scenarios. Include questions regarding each object involved.
- Questions and Answers:
{{'Question': 'What might not knock down the chessboard pieces if the white chess piece had not been moved?', 'Answer': 'The cat', 'Object ID': '1', 'QA Type': 'Counterfactual'}}
{{'Question': 'What might not be knocked down by the cat if the white chess piece had not been moved?', 'Answer': 'The chessboard pieces', 'Object ID': '2', 'QA Type': 'Counterfactual'}}

4. Descriptive QA: Simple questions about descriptive expressions with direct object-focused answers.
- Questions and Answers:
{{'Question': 'Who is dancing on the stage?', 'Answer': 'The man in red', 'Object ID': '1', 'QA Type': 'Descriptive'}}

Figure 9. QA generation prompt.

```
      "action_start": "0:00",
      "answer": "The man",
      "obj_id": "1",
      "q_type": "Causal",
      "question": "Who uses the cut jug to
          scoop water out of the canoe?"
    },
    "2": {
      "action_end": "0:15",
      "action_start": "0:00",
      "answer": "The cut jug",
      "obj_id": "2",
      "q_type": "Causal",
      "question": "What does the man use to
           scoop water out of the canoe?"
    }
  }
}
```

Each entry in the JSON file consists of a series of questions associated with the video. Each question contains the following fields:

- `action_start` and `action_end` specify the time segment in the video corresponding to the action.
- `answer` provides the correct response to the question.
- `obj_id` uniquely identifies the object involved in the question.
- `q_type` indicates the question type, such as ”Causal”.
- `question` is the text of the question related to the action in the video.

5. Dataset Necessity

In previous MeViS [3], the more challenging motion expressions increase the difficulty of the dataset compared with previous benchmarks, since the target objects have to be distinguished from others by sophisticated motion understanding. In our GROUNDMORE, we not only consider the abundant temporal reasoning clues in the motion expressions, we also take the *implicit reasoning* into account and we view it as a core challenge in Motion-Grounded Video Reasoning. Moreover, we hypothesize that containing motion expressions though, the object information in the input language in MeViS might still result in an identity leakage and make the model ignore the motion description but rely on the target information itself. To validate this, we made a modification on the original expressions in MeViS valid-u data so that the object name will be replaced by *”something”*, making the original explicit expressions into implicit ones. After this, we ran the evaluation process as usual and only found that the performance had an obvious drop, about 20% as shown in Table 1. In our GROUNDMORE, since we intentionally omit the target identity by using the questions as our implicit expressions, we force the models to focus on the motion clues and perform reasoning before the segmentation process. In this way, the motion information is guaranteed to be leveraged. This interesting discovery in Table 1 not only demonstrates the weak implicit expression processing ability in existing models but also validates the necessity of our task and dataset, i.e., our

Expressions Type	J&F	J	F
original (explicit)	40.23	36.51	43.90
implicit	32.33	28.81	35.86

Table 1. Comparison of explicit and implicit expression on MeViS valid-u.

implicit questions are not similar to the motion expressions.

6. Impact of Object Numbers

The number of objects will affect the results a lot, which is also consistent with the intuition that more objects in the videos will bring more difficulties in localizing target objects. Due to the time limit, we cannot obtain the overall analysis now, but we do obtain a subset results. Specifically, we randomly sample two subsets (containing 120 instances each) from GroundMoRe, the first subset contains videos that include less than 3 objects, and the second one with more than 6 objects (we ignore visual-insignificant objects here). The results (MoRA zero-shot) are shown in Table 2.

	J&F	J	F
#OBJ ≤ 3	23.61	23.77	23.45
#OBJ ≥ 6	14.38	14.52	14.24

Table 2. The impact of object numbers in GroundMoRe.

7. Details of MoRA

We build our baseline model mainly based on LISA [5]. We extend the image-based model to the temporal domain by introducing the spatiotemporal pooling [6] after frame encoding, and embedding the video features into LLM space after the linear projection layer. The linear projection layer is a 1-layer MLP that project the visual feature from the visual hidden dimension to the language model hidden dimension. The output of MoRA is designed templates that include special tokens: **[SEG]** and **[LOC]**. The **[SEG]** token corresponds to the visual embedding that contains the target object semantic that can be decoded by SAM [4] decoder with the frame embeddings. The **[LOC]** learns the temporal boundary semantic and the corresponding embedding can be decoded to a binary temporal mask, which suppresses the temporal false positive (target object exists but the motion in the question does not take place) from the direct output of the SAM decoder.

8. Video LLMs in Two-Stage Baselines

Compared to the results in the main paper, we can still observe that SeViLA outperforms other video QA models in the two-stage setting. A key reason is that SeViLA generates concise and precise answers, avoiding the inclusion of redundant information that could negatively impact the performance of RefVOS models.

For example, given the question *”What does the man in white dribble?”*, the answers from the video QA models are as follows:

- **SeViLA**: ”a basketball.”
- **VideoChat2**: ”The man in white is dribbling a basketball in the video.”
- **VILA**: ”The man in white dribbles the ball around the court while the man in black tries to block him.”

Similarly, for the question *”Who snatches the ball after the man in grey accelerates towards him?”*, the answers are:

- **SeViLA**: ”the man in red.”
- **VideoChat2**: ”The man in red snatches the ball after the man in grey accelerates towards him.”
- **VILA**: ”The man in grey snatches the ball after the man in red accelerates towards him.”

9. MoRA on RefYouTube-VOS

We also evaluate the performance of MoRA on a referring video object segmentation benchmark, RefYouTube-VOS [9]. As shown in Table 3, MoRA can achieve reasonable results compared with other task-specific models. It is worth noting that MoRA provides spatiotemporal masks given the videos and the query, which is not a suitable design for the RefVOS task.

Methods	RefYoutubeVOS		
	J&F	J	F
MTTR [2]	55.3	54	56.6
ReferFormer [10]	64.9	62.8	67.0
UniRef [11]	67.4	65.5	69.2
SgMg [7]	65.7	63.9	67.4
HTR [8]	67.1	65.3	68.9
MoRA	57.8	57.4	58.2

Table 3. Performance on RefYouTube-VOS dataset.

10. Limitation and Future Work

Although our dataset has included a wide range of video scenarios, there are still many scenarios and motion types to be considered, e.g., motion in first-person-view videos. Besides, in the current version, we only consider single-object as target (even though multiple objects appear in the scene), which is less complicated than simultaneously grounding multiple targets.

Besides, we will also consider more modalities, such as audio (which could provide more nuance information beyond visual clues) and keypoint (which could introduce direct motion features), to construct more comprehensive training data as well as the evaluation benchmark.

11. Ethics Statement

Copyright and Fair Use Disclaimer. The collection and use of GROUNDMORE are conducted in accordance with the principles of Fair Use [1] as outlined in U.S. copyright law, particularly for purposes such as research, scholarship, and commentary. The dataset is provided under a strict non-commercial use policy. Any use of GROUNDMORE must adhere to these restrictions, and users are prohibited from using the dataset in any way that may infringe on the rights of the original content creators. By accessing the dataset, users agree to comply with these terms and with the principles of Fair Use.

Privacy Considerations. Since GROUNDMORE includes segments from videos that may contain identifiable human faces and actions, we acknowledge the importance of addressing privacy concerns. The dataset is restricted to non-commercial use only, with the primary aim of advancing research and education. We have taken additional steps to ensure ethical standards are maintained by submitting the dataset for review by the Institutional Review Board (IRB) at our university, and the IRB submission is currently under review.

License. GROUNDMORE is distributed under the Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0)[2]. This license allows others to remix, adapt, and build upon the dataset for non-commercial purposes, provided that appropriate credit is given. Commercial use of the dataset is strictly prohibited.

Data Usage Responsibility. We encourage all users of GROUNDMORE to adhere to ethical research standards, including fairness, transparency, and respect for individual privacy. Researchers are expected to consider the ethical implications of their work and to ensure that any models or technologies developed using GROUNDMORE do not inadvertently reinforce biases or infringe on individual rights.

References

[1] Maksym Bekuzarov, Ariana Bermudez, Joon-Young Lee, and Hao Li. Xmem++: Production-level video segmentation from few annotated frames, 2023. 5

[2] Adam Botach, Evgenii Zheltonozhskii, and Chaim Baskin. End-to-end referring video object segmentation with multimodal transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4985–4995, 2022. 9

[3] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Chen Change Loy. Mevis: A large-scale benchmark for video segmentation with motion expressions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2694–2703, 2023. 8

[4] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 9

[5] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. *arXiv preprint arXiv:2308.00692*, 2023. 9

[6] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv:2306.05424*, 2023. 9

[7] Bo Miao, Mohammed Bennamoun, Yongsheng Gao, and Ajmal Mian. Spectrum-guided multi-granularity referring video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 920–930, 2023. 9

[8] Bo Miao, Mohammed Bennamoun, Yongsheng Gao, Mubarak Shah, and Ajmal Mian. Towards temporally consistent referring video object segmentation. *https://arxiv.org/abs/2403.19407*, 2024. 9

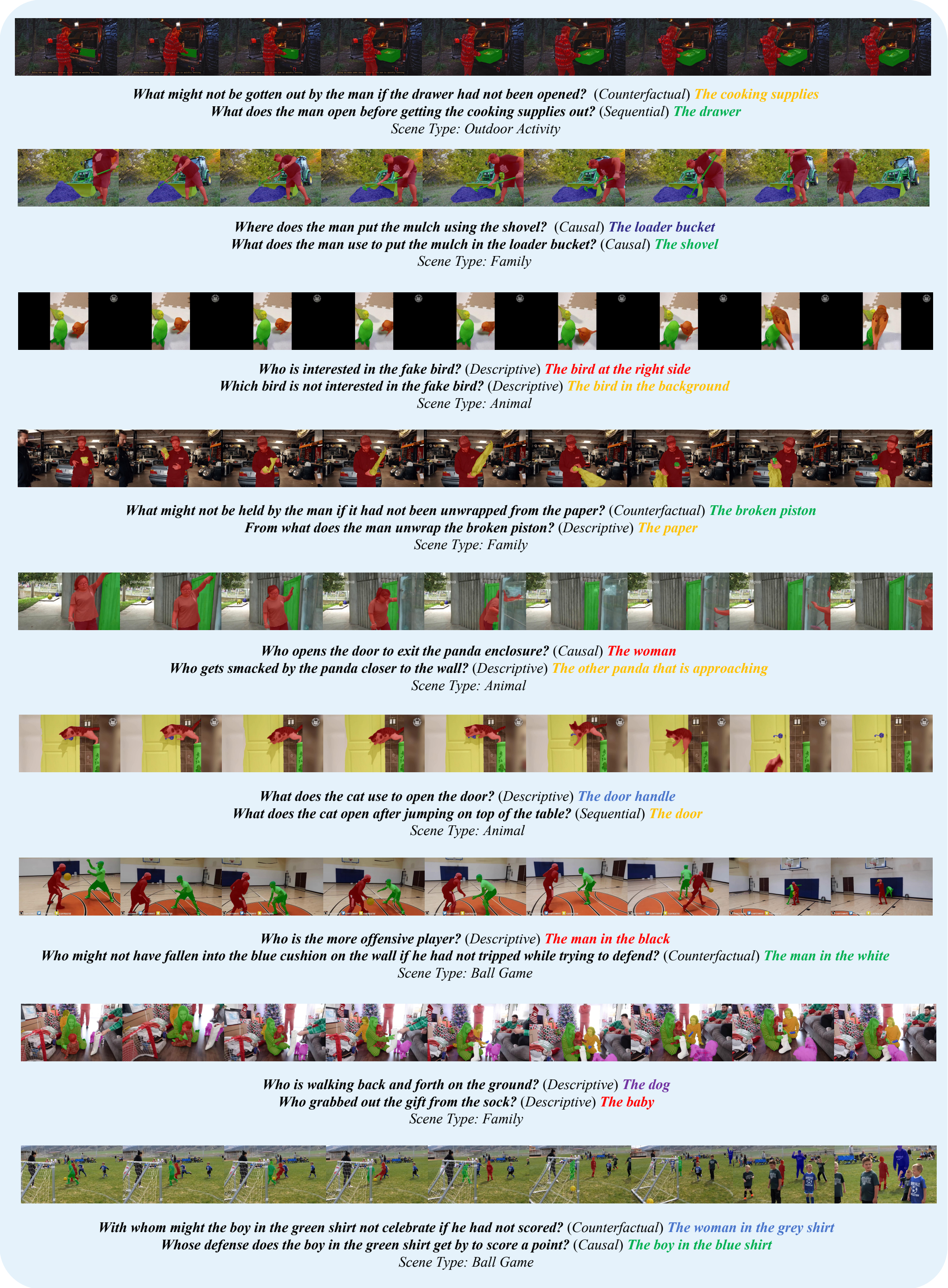
[9] Seonguk Seo, Joon-Young Lee, and Bohyung Han. Urvos: Unified referring video object segmentation network with a large-scale benchmark. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pages 208–223. Springer, 2020. 9

[10] Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. Language as queries for referring video object segmentation. *arXiv preprint arXiv:2201.00487*, 2022. 9

[11] Jiannan Wu, Yi Jiang, Bin Yan, Huchuan Lu, Zehuan Yuan, and Ping Luo. Segment every reference object in spatial and temporal spaces. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2538–2550, 2023. 9

[1] For more information on Fair Use, see https://www.copyright.gov/fair-use

[2] For more details on the license, see https://creativecommons.org/licenses/by-nc/4.0/



What might not be gotten out by the man if the drawer had not been opened? (*Counterfactual*) ***The cooking supplies***
What does the man open before getting the cooking supplies out? (*Sequential*) ***The drawer***
Scene Type: Outdoor Activity

Where does the man put the mulch using the shovel? (*Causal*) ***The loader bucket***
What does the man use to put the mulch in the loader bucket? (*Causal*) ***The shovel***
Scene Type: Family

Who is interested in the fake bird? (*Descriptive*) ***The bird at the right side***
Which bird is not interested in the fake bird? (*Descriptive*) ***The bird in the background***
Scene Type: Animal

What might not be held by the man if it had not been unwrapped from the paper? (*Counterfactual*) ***The broken piston***
From what does the man unwrap the broken piston? (*Descriptive*) ***The paper***
Scene Type: Family

Who opens the door to exit the panda enclosure? (*Causal*) ***The woman***
Who gets smacked by the panda closer to the wall? (*Descriptive*) ***The other panda that is approaching***
Scene Type: Animal

What does the cat use to open the door? (*Descriptive*) ***The door handle***
What does the cat open after jumping on top of the table? (*Sequential*) ***The door***
Scene Type: Animal

Who is the more offensive player? (*Descriptive*) ***The man in the black***
Who might not have fallen into the blue cushion on the wall if he had not tripped while trying to defend? (*Counterfactual*) ***The man in the white***
Scene Type: Ball Game

Who is walking back and forth on the ground? (*Descriptive*) ***The dog***
Who grabbed out the gift from the sock? (*Descriptive*) ***The baby***
Scene Type: Family

With whom might the boy in the green shirt not celebrate if he had not scored? (*Counterfactual*) ***The woman in the grey shirt***
Whose defense does the boy in the green shirt get by to score a point? (*Causal*) ***The boy in the blue shirt***
Scene Type: Ball Game

Figure 10. Additional Visualizations of our GroundMoRe. We provide visualizations of videos alongside their corresponding segmentation masks, questions, answers (color corresponds to the segmentation masks), and scene types.