LEVERAGING VLMS AND LLMS FOR COMMON-SENSE REASONING IN VISUAL SCENES

Allan Tan, Anuranjan Pandey, Bhargav Krishnamurthy, Kriti Jha, Pratyush Bhatnagar & Swapnil Chhatre
Viterbi School of Engineering
University of Southern California
Los Angeles, CA 90089, USA
{allantan, akpandey, bhargavk, kritijha, pb65357, chhatre}@usc.edu

ABSTRACT

This research project aims to advance AI capabilities by leveraging Vision Language Models (VLMs) and Large-Language Models (LLMs) to enhance common-sense reasoning within visual contexts. Our approach integrates a vision encoder with architectures for temporal and spatial modeling to deepen AI's comprehension and interaction with visual data, particularly videos. We focus on the frame-by-frame analysis of visual scenes to capture temporal dependencies and contextual nuances, thereby bridging the gap in AI's ability to reason about the physical world. This enables more accurate predictions of future actions and deeper understanding of physical dynamics. Utilizing the Compositional Physical Reasoning (ComPhy) dataset, our model is trained to infer latent physical attributes from videos and to answer complex queries about the dynamics within physical environments. Our research not only aims to enhance AI's safety in navigating complex environments but also seeks to tailor learning experiences and contribute to threat identification and prevention, addressing key challenges in real-world applications.

1 INTRODUCTION

This research project represents an effort to advance the capabilities of Artificial Intelligence (AI) systems by harnessing the power of Vision-Language Models (VLMs) and Large-Language Models (LLMs) to enhance physical common-sense reasoning within visual contexts. By integrating a vision encoder alongside a basic architecture for temporal modeling, the project aims to enhance how AI comprehends and interacts with videos. To model the spatial information better, we use slot attention Locatello et al. (2020). The unified reasoning capabilities adhering to physics have far-reaching potential in any domain, requiring an intuitive understanding of the physical environment and dynamics governed by the laws of physics.

At its core, the project aims to enable AI systems to understand the nuances and complexities present in visual scenes by analyzing them frame by frame. This detailed analysis is crucial for bridging the existing gap in AI's common-sense reasoning about the physical world, thereby laying the groundwork for AI systems to exhibit a more sophisticated understanding of their environment. Such enhanced capabilities not only allow machines to predict future actions with greater accuracy but also empower them to communicate their understanding in a manner that is comprehensible to humans.

To initiate our investigation, we delve into assessing the reasoning capabilities exhibited by a spectrum of cutting-edge Vision-Language Models, such as Visual chat-GPT Wu et al. (2023), MM-React Yang et al. (2023), FROZEN Tsimpoukelli et al. (2021), BLIP-2 Li et al. (2023), LLaMA-Adapter Zhang et al. (2023b), MiniGPT-4 Zhu et al. (2023), and LLaVA Liu et al. (2023). Employing the Compositional Physical Reasoning dataset, introduced by Chen et al. Chen et al. (2022), designed to test reasoning abilities in understanding how objects interact under various physical conditions, we focus on predicting what might occur in forthcoming scenarios based on current configurations. We also aim to provide descriptions of the attributes of the objects, such as relative positioning, size, material composition, shape, etc. This analysis serves as a foundational step in understanding the nuanced performance of these models and their proximity to human-like responses.

The potential applications of this technology are vast and varied. In the realm of autonomous navigation, AI systems equipped with enhanced common-sense reasoning capabilities can navigate complex environments more effectively, leading to safer and more efficient transportation systems. Similarly, in personalized education, these systems can provide tailored learning experiences by understanding and adapting to the individual needs of students. Furthermore, in the realm of security, AI systems capable of generating human-like descriptions of video content can aid in the identification and prevention of threats, enhancing overall safety and security.

2 Related Work

Integrating vision and language models to enhance common-sense reasoning for understanding visual contexts like videos is a challenging task. In recent years, there has been a surge of interest in developing vision-language models that can effectively understand and reason about visual and textual information jointly. This literature review covers the latest advancements in this field, focusing on various approaches, architectures, and applications.

2.1 Video Vision Transformers

Several works have explored the use of transformer-based architectures for video understanding and classification tasks. ViViT Arnab et al. (2021) introduced a pure transformer-based model for video classification, processing spatio-temporal tokens from the input video using a series of transformer layers. TimeSformer Bertasius et al. (2021) proposed a convolution-free approach, adapting the standard transformer architecture to video by enabling spatio-temporal feature learning directly from frame-level patches. Similarly, the all-in-one transformerWang et al. (2022) incorporates the video elements directly into a transformer. Raw video and textual signals are rolled into joint representations using a unified backbone architecture.

2.2 Video-Language Modeling and Alignment

A significant line of research has focused on developing models that can align and jointly process video and textual information. Slot-VLM Xu et al. (2024) proposed a novel framework for generating semantically decomposed video tokens to facilitate language model inference for video question-answering tasks. Another work by Zhao et al. (2024) explored fine-tuning an image-language model with synthesized instructional data, using the adapted model to auto-label millions of videos and generate high-quality captions for video-language tasks. Video-LLaMA Zhang et al. (2023a) introduced a multi-modal understanding approach, training vision-language and audio-language branches separately and fine-tuning on instructional datasets to enhance the model's ability to follow instructions and comprehend visual and auditory inputs. The model builds on top of Q-former model introduced in Blip2 and adopts it for a video medium.

2.3 Masked Visual Modeling

The use of masked visual modeling (MVM) has been explored in the context of video-language (VidL) pre-training. Fu et al. (2023) conducted a systematic study on the potential of MVM in VidL learning, presenting VIOLETv2, an enhanced model pre-trained with an MVM objective, achieving notable improvements on various VidL benchmarks.

2.4 Unified Video-Language Frameworks

Several works have aimed to unify video-language understanding tasks under a common framework. All-in-One introduced an end-to-end video-language model that embeds raw video and textual signals into joint representations using a unified backbone architecture. LAVENDER Li et al. (2022) proposed a unified VidL framework where Masked Language Modeling (MLM) is used as the common interface for all pre-training and downstream tasks, achieving competitive performance on 14 VidL benchmarks.

2.5 Efficient Video-Language Alignment

Addressing the challenge of efficient and effective video-language alignment, VLAP Wang et al. (2024) introduced a novel network with a learnable Frame-Prompter for selecting relevant video frames and a QFormer-Distiller module for cross-modal fusion and distillation. Another work by Lin et al. (2023) explored fast adaptation of pretrained contrastive models for multi-channel video-language retrieval, investigating different approaches for representing videos and fusing video and text information.

2.6 Video Captioning

SwinBERT Lin et al. (2022) presented an end-to-end transformer-based model for video captioning, taking video frame patches directly as inputs and outputting natural language descriptions. It introduced a Sparse Attention Mask that adaptively learns to focus on significant video frame patches, reducing redundancy and improving captioning performance.

2.7 Commonsense Reasoning

While most existing works have focused on aligning visual and textual information, some studies have explored the capability of vision-language models for commonsense reasoning. VICOR Zhou et al. (2023a) proposed a method for integrating vision-and-language models with large language models for visual commonsense reasoning tasks. Cola Chen et al. (2023) employed a large language model to coordinate the strengths of multiple vision-language models for enhanced reasoning capabilities. Lastly, Rai et al.Zhou et al. (2023b) introduced ROME, a novel probing dataset to evaluate whether pre-trained vision-language models can reason beyond visual common sense.

3 Dataset Description

In this research, we leverage an existing benchmark dataset called Compositional Physical Reasoning (ComPhy). ComPhy serves as a foundational resource for advancing physical reasoning within video modeling by addressing the challenge of inferring hidden physical properties from observable visual cues in dynamic scenarios.

In addition, CLEVR Johnson et al. (2016) is used as a supplementary dataset. Its images are graphically similar to ComPhy in a synthetic nature. It contains nearly 7 times as many datapoints in its training set alone and allows us to more comprehensively pre-train our slot attention mechanism.

In order to instruction tune the model we use the Video Instruction Data curated by the authors of Video-ChatGPT Maaz et al. (2023). This dataset consists of approximately 100,000 video-text pairs, where each pair comprises a question and its corresponding answer. The video-text pairs were generated from the ActivityNet dataset Fabian Caba Heilbron & Niebles (2015), with an average video duration of 180 seconds.

3.1 Key Features of Compositional Physical Reasoning (ComPhy)

1. Compositional Nature: ComPhy emphasizes the compositional relationships between visible and hidden properties of objects, facilitating the exploration of how these factors influence object interactions in video sequences.

2. Inclusion of Hidden Properties: The dataset includes annotations for physical properties such as mass and electric charge, which are essential for understanding object dynamics but not directly observable. This enables researchers to develop models capable of inferring and reasoning about these latent attributes.

3. Diverse Interaction Scenarios: ComPhy encompasses a variety of dynamic interaction scenarios, where objects move and interact under different initial conditions. This diversity challenges models to generalize across varied contexts and infer physical properties amidst dynamic interactions.

4. Evaluation Framework: Evaluation in ComPhy focuses on assessing models' ability to discern and utilize hidden physical properties for answering questions posed on video sequences. This

evaluation framework provides insights into models' proficiency in unraveling compositional relationships and predicting system dynamics.

In our research, we utilize ComPhy as a benchmark dataset to evaluate the performance of our proposed method for video physical reasoning. By leveraging the diverse scenarios and annotated physical properties within ComPhy as shown in Fig 1, we aim to demonstrate the effectiveness and robustness of our approach in inferring hidden physical attributes and predicting object behaviors in dynamic environments.

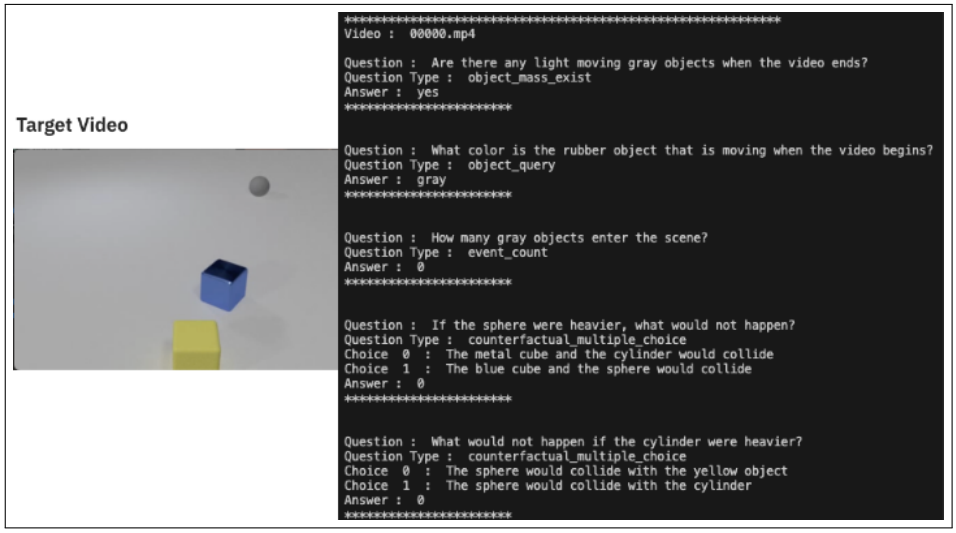

Figure 1: An example of ComPhy dataset

3.2 Dataset Statistics

ComPhy consists of 8,000 training sets, 2,000 validation sets, and 2,000 testing sets. The dataset comprises 41,933 factual questions, 50,405 counterfactual questions, and 7,506 predictive questions, distributed across the training, validation, and testing splits. To ensure the integrity and suitability of the dataset for evaluation, efforts have been made to mitigate biases and ensure informative interactions within video sets.

4 Methodology

The proposed architecture in Fig 2 extracts features from videos by separating them into frames, sampling at different rates, and encoding them using the CLIP ViT-L/14 Dosovitskiy et al. (2020) image encoder resulting in the dimension (H x W x D x T) where 'H' is the height, 'W' is the width, 'D' is the hidden dimension and 'T' is the duration of the input video. The spatial and temporal aspects are modeled using the Spatial Slot Attention Module and Temporal Slot Module, respectively, which produce abstract representations of the input video. In the Fig 2, the output of the Spatial slot attention and Time Slots is of the form (T x N_s) where N_s denotes the number of spatial slots. Whereas, the output of the Time Slots is of the dimension (N_f x M_d), where N_f denotes the number of sampled frames and M_d is (H x W x D). The outputs are then projected and concatenated into a unified representation, which is fed into Vicuna Chiang et al. (2023) a pre-trained large language model along with text instructions. This allows the model to leverage the reasoning capabilities of LLMs on rich video data and text instructions.

4.1 Feature Extraction

Because the ComPhy dataset is primarily composed of synthetic videos, feature extraction was fairly straightforward and simple. It was done in a manner similar to Slot-VLM and many other VLM

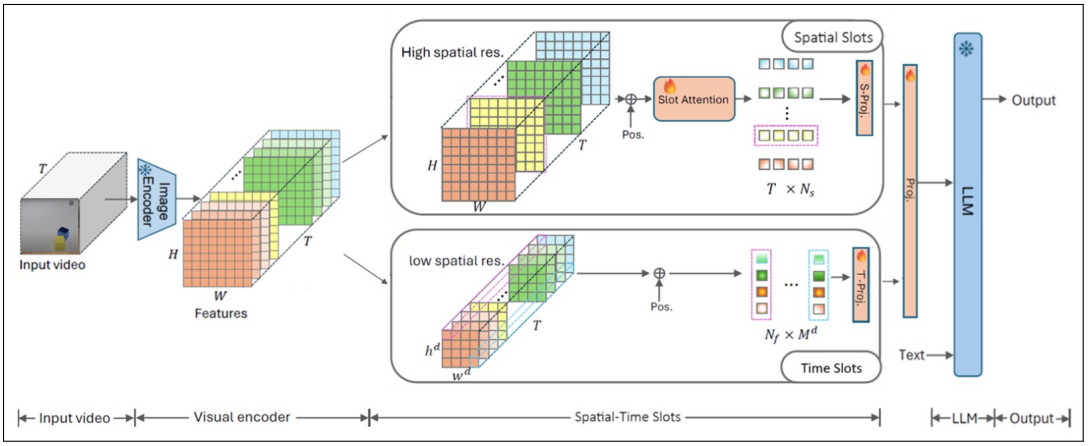

Figure 2: Architecture of the proposed model

models. The process was twofold: isolate each video into individual frames via a sample rate and encode each frame into a vector to ready it for calculation.

To start, OpenCV was used to separate the videos into individual frames. Each video in the ComPhy consisted of 127 frames and had a frame rate of 25 frames per second, meaning the length was about 5 seconds. To reduce redundancy and computational load, the videos were then sampled at a certain frame rate depending on the destination branch. In other words, a set number of frames would be saved per second of video. For our use case, given a 5 second long video, a sample rate of 1 fps was used to generate 6 frames, and a sample rate of 2 fps was used to generate 11 frames per video for the spatial branch and temporal branch, respectively. Next, each frame was resized to 128x128 resolution before being appended to an array, representing a whole video.

Once isolated, features could then be extracted from each frame per video. We utilized the vision-language pre-trained model CLIP (ViT-L/14) as our image encoder. We refer to this encoder as our vision transformer (ViT). Once an image is input, we examine only the penultimate layer of the model output, resulting in a final vector of size 16x16x1024.

4.2 Slot Attention Modules

Previous approaches for modeling spatial and temporal aspects of video features have relied on Pooling in Video-ChatGPT or Q-former Zhang et al. (2023a) techniques to aggregate the features. However, these methods are suboptimal as the coupled representations of the tokens may lose information, hampering video understanding and ultimately affecting the performance of the eventual large language model (LLM) inference.

To overcome the limitations of previous approaches and capture the underlying semantics more effectively, we introduced two dedicated modules: a spatial slot attention module and a time slots module. The spatial slot attention module models the structural arrangements of objects within each video frame, while the time slot module captures the temporal dynamics across frames. By concatenating the token representations from these two modules, we can preserve the decoupled dynamics of the videos, unlike previous approaches that may lose crucial information.

4.2.1 Spatial Slot Attention

The spatial module operates on high-resolution features and produces a set of task-dependent abstract representations called slots. These slots are exchangeable and can bind to any object in the input through a competitive attention procedure over multiple rounds.

Before passing the output of the ViT (16 x 16 x 1024 x 6) tokens to the slot attention module, we added a positional embedding to every token of each frame to capture differences in objects'

positions. These tokens were then processed by the slot attention module, which takes the form of a convolutional neural network (CNN)-based encoder/decoder.

We considered 8 spatial slots for each video. The output of this module is a tensor of dimension (6 x 8 x 64) for each video. To pre-train this module, we incorporated the same technique used by Dinosaur Seitzer et al. (2023), where we reconstructed back the ViT features from the 8 slots generated by the forward pass of the attention module. During this reconstruction, the module learns to better model the generated slots. We pre-trained the model for 25 epochs on the CLEVR dataset using an Nvidia A100 (40GB) GPU.

4.2.2 Time Slots

To capture temporal aspects more effectively, we used a low-resolution approach by average pooling the features from (16 x 16 x 1024) to (4 x 4 x 1024). We sampled each video at 2 frames per second, resulting in 11 frames per video. As we aimed to capture event-specific slots, the number of time slots are equal to the number of sampled frames.

Lastly, we added positional embeddings to each spatial patch over the T frames, allowing the module to model sequences and changes in events across frames. The final embeddings from this branch are (11 x 16 x 1024).

4.3 Projection layers and LLM

To integrate the spatial and temporal representations for LLM-based reasoning, we employ dedicated projection layers. The 6 x 8 x 64 spatial representations from the slot attention module are first projected to 6 x 1024 vectors. Similarly, the 11 x 16 x 1024 temporal representations capturing dynamics across frames are projected to 11 x 1024 vectors. These spatial and temporal projections are then concatenated and collectively projected to a 17 x 4096 representation, aligning with the input dimensionality of the language model. This unified 17 x 4096 representation is subsequently fed into a pre-trained large language model (LLM), specifically Vicuna, along with text instructions for common-sense reasoning and inference tasks. Instruction fine-tuning was done using 100,000 video-text pairs generated from a subset of the ActivityNet dataset, providing additional data to enhance the pre-training phase. All training was successfully performed on 2 Nvidia A100 80GB GPUs.

5 Results and Discussion

Our results primarily consist of two parts: sanity checking via visualization and a comparison between a baseline and our trained model.

5.1 Sanity Check

To ensure the integrity and utility of the feature vectors fed into the Vision Language Models (VLMs), it is crucial to understand their contributions and how they enhance the model's outputs. We conducted sanity checks by visualizing these intermediary results, which provided invaluable insights into the model's performance and areas for improvement. For instance, our visualizations revealed that in certain reconstructions, like the one shown in the fourth row of Figure 4, the yellow object towards the edge of the frame was missing, which led to poor query responses related to those objects. These visual checks are instrumental in understanding the limitations and capabilities of our models in real-world scenarios.

5.1.1 Visualization of attention

We took inspiration from the DINO visualization module to visualize the attention from the ViT fast branch. By visualizing attention from the ViT fast branch and integrating it with the temporal branch's feature vectors, one can potentially uncover how the model incorporates temporal information into its decision-making process. This can lead to insights into how the final multimodal model handles temporal information across frames as shown in Fig 3.

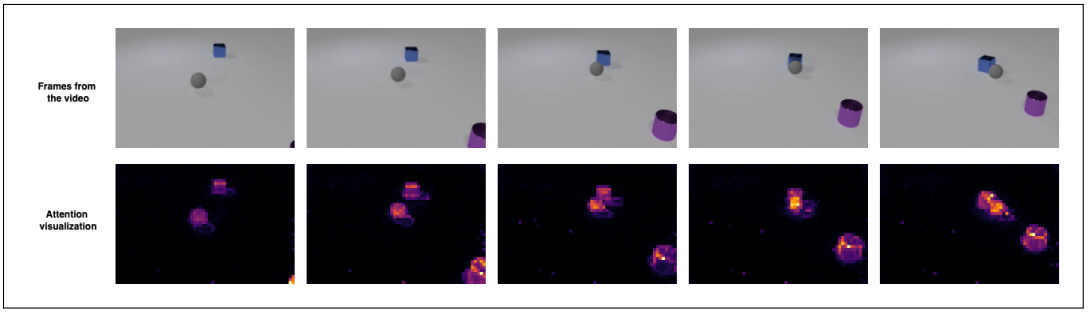

Figure 3: Visualization of attention from the time branch

5.1.2 Incorporating Spatial Features through slot attention

The slot attention model is typically structured as an encoder-decoder architecture. The encoder processes the input data, extracting meaningful representations using ViT, while the decoder generates outputs based on these representations. The core of the slot attention model lies in its self-attention mechanism, allowing the model to weigh the significance of different parts of the input data during predictions. This mechanism is crucial for capturing intricate patterns and dependencies within the data, enhancing the model's performance. We've visualized what gets captured in the slots at Fig 4 to show the efficacy of the spatial module.

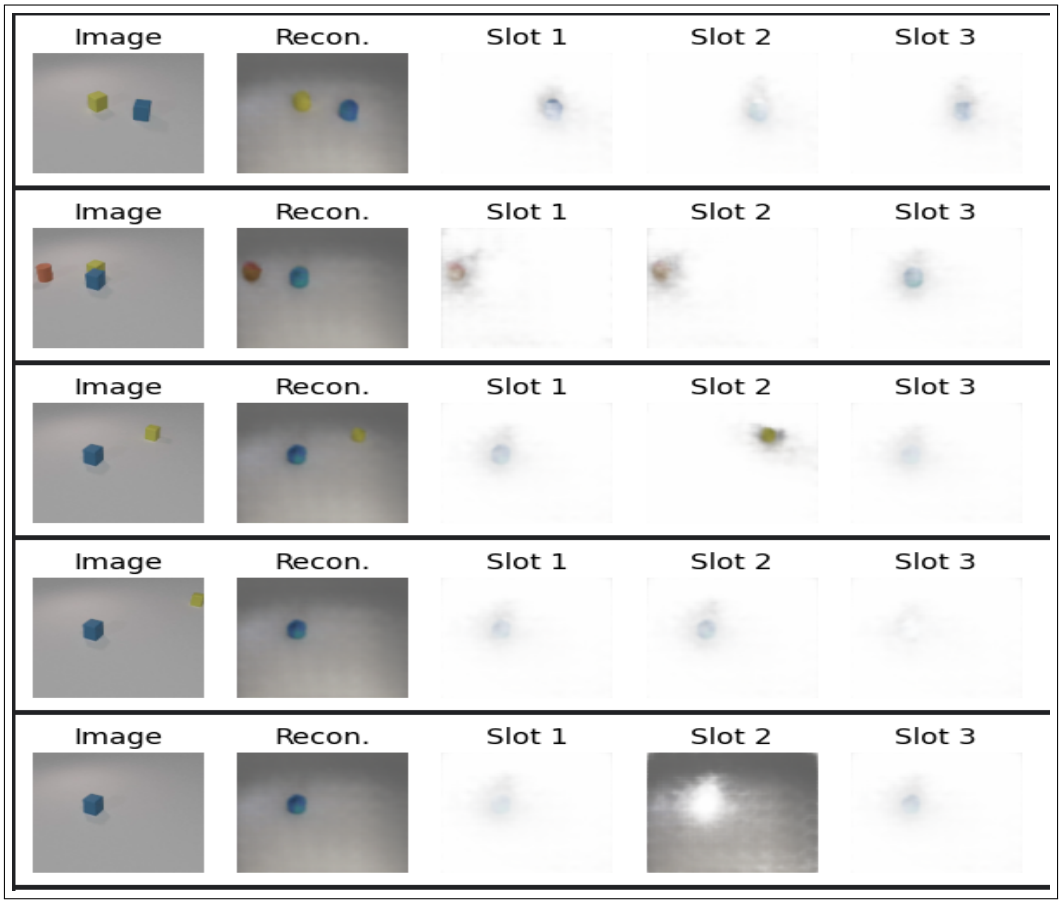

Figure 4: Visualization of attention from the spatial branch

5.2 THE EFFECT OF STITCHING FRAMES INTO A SINGLE IMAGE

Natively, LLaVA does not support video format and primarily only takes input in the form of image-question pairs. As a result, to get any meaningful and comparable output from the model we would need to provide input of videos in the form of an image. The workaround we developed was to stitch together the frames of a video into a single image. Because of the short duration of each of the ComPhy videos, we were able to generate an image of 5 frames in a manner similar to what was done during feature extraction. This was then fed into LLaVA along with a question to receive a response. Thus, we were able to transform LLaVA into a rudimentary baseline, which we could use to compare and evaluate performance with our trained model. Fig 5 highlights the effectiveness of this rather simplistic approach.

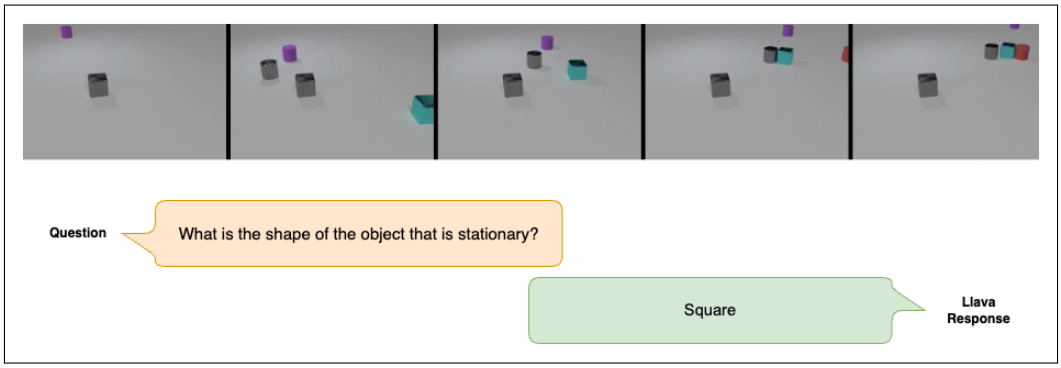

Figure 5: LLaVA response on stitched frames

There is also a downside to this. Such an approach makes sense for short videos. If the video frames of longer videos are stitched together, the volume of information in each frame is overloaded. This method fails to scale for longer videos.

5.3 MODEL EVALUATION ON THE COMPHY DATASET

As part of our analysis, we conducted tests to evaluate the reasoning capabilities of our models concerning counterfactual and factual reasoning based on visual input. Figures 6 and 7 present typical examples where the model's predictions are compared against the ground truth and LLaVA's response. In the first example (Fig 6), we assess the model's ability to predict the interaction outcomes between objects if one of the object's physical properties, such as weight, is altered. The ground truth suggests that making the sphere heavier would prevent it from colliding with the cylinder, highlighting the model's sensitivity to changes in physical dynamics.

In the second example (Fig 7), we challenge the model's ability to accurately count the number of objects in a video frame. The model's performance can be compared with the LLaVA model and the ground truth to gauge its precision in basic observational tasks.

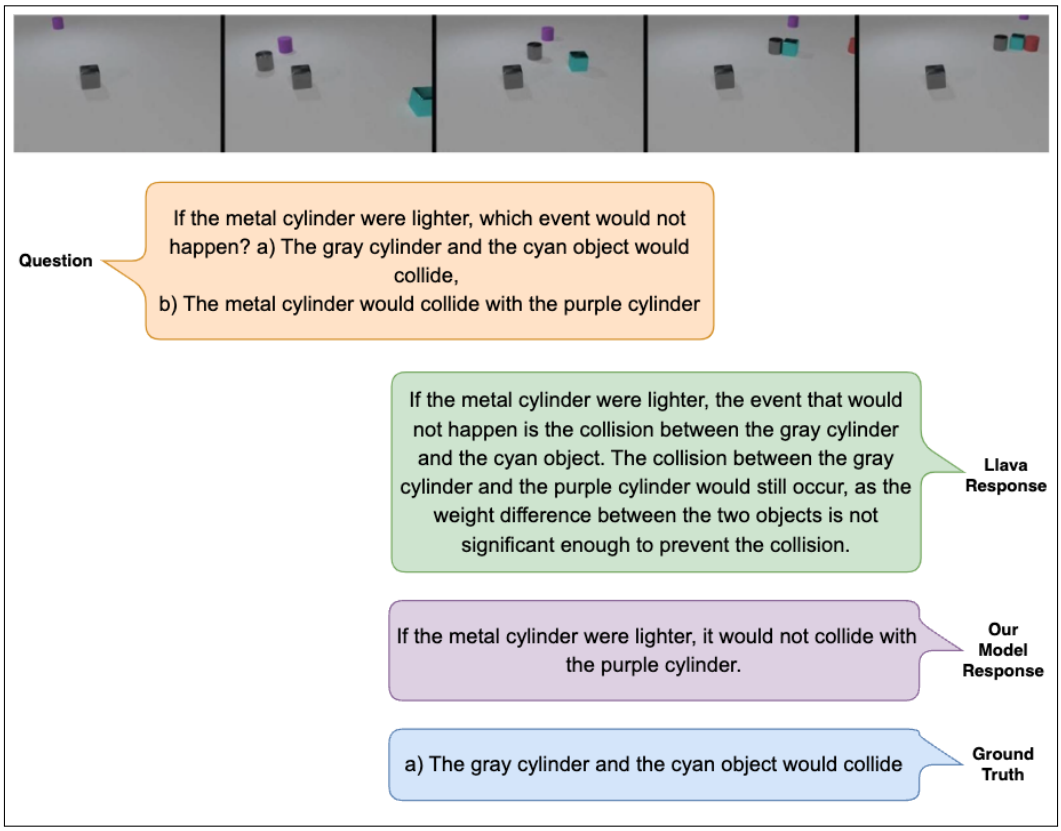

Figure 6: Sample responses for a counter-factual question

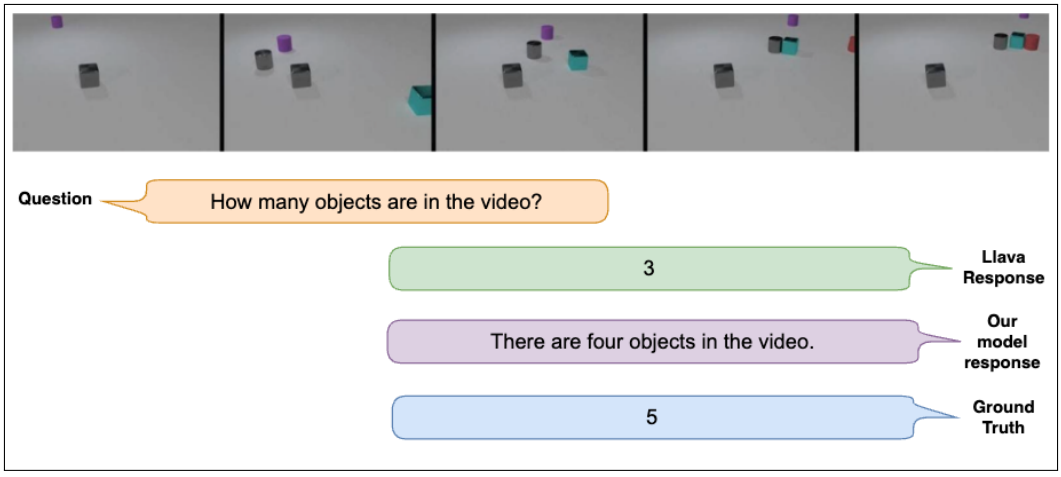

Figure 7: Sample responses for a factual question

This part of our study underscores the challenges and capabilities of current VLMs and LLMs in interpreting and reasoning within dynamic and visually complex environments, as encountered in the ComPhy dataset.

6 FUTURE WORK

Enhancing the visual embeddings with temporal sequencing is a vital aspect to consider. For now we have limited our scope of research only to counter factual questions but ComPhy has a wide variety of other questions also like predictive and factual questions. Also the videos in ComPhy dataset are from a controlled environment, which has a fixed camera having the perfect lighting conditions, but in real world that is not the case and our model have the scope of improvement for variable environmental condition.

In the current implementation, the slot attention module uses CNN based encoder to pick out highly relevant spatial features. This can be extended with a powerful pre-trained transformer based vision encoder.

There can be experiments on late fusion, early fusion, or attention-based fusion mechanisms to effectively combine information from diverse modalities like Audio, Lidar, etc while preserving their unique characteristics.

We can extend our experiments to larger vision models beyond the LLaVA 7B used in this study. Investigating the performance and scaling properties of substantially larger models could provide valuable insights and potentially lead to enhanced capabilities.

Additionally, we can explore joint training of the different modules involved in the multimodal embedding process. This could involve fine-tuning the CLIP embeddings, attention slots, and projection layers simultaneously, with the aim of achieving better coherence and alignment between the different components.

We can also leverage Chain-of-Thought (CoT) Wei et al. (2022) prompting technique to encourage the language model to break down complex video understanding tasks into a series of intuitive reasoning steps.

REFERENCES

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. ViViT: A Video Vision Transformer. November 2021. URL http://arxiv.org/abs/2103.15691. arXiv:2103.15691 [cs].

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is Space-Time Attention All You Need for Video Understanding? June 2021. URL http://arxiv.org/abs/2102.05095. arXiv:2102.05095 [cs].

Liangyu Chen, Bo Li, Sheng Shen, Jingkang Yang, Chunyuan Li, Kurt Keutzer, Trevor Darrell, and Ziwei Liu. Large Language Models are Visual Reasoning Coordinators. October 2023. URL http://arxiv.org/abs/2310.15166. arXiv:2310.15166 [cs].

Zhenfang Chen, Kexin Yi, Yunzhu Li, Mingyu Ding, Antonio Torralba, and Joshua B Tenenbaum. COMPHY: COMPOSITIONAL PHYSICAL REASONING OF OBJECTS AND EVENTS FROM VIDEOS. 2022.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL https://lmsys.org/blog/2023-03-30-vicuna/.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929, 2020. URL https://api.semanticscholar.org/CorpusID:225039882.

Bernard Ghanem Fabian Caba Heilbron, Victor Escorcia and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 961–970, 2015.

Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. An Empirical Study of End-to-End Video-Language Transformers with Masked Visual Modeling. May 2023. URL http://arxiv.org/abs/2209.01540. arXiv:2209.01540 [cs].

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning, 2016. URL https://arxiv.org/abs/1612.06890.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models, June 2023. URL http://arxiv.org/abs/2301.12597. arXiv:2301.12597 [cs].

Linjie Li, Zhe Gan, Kevin Lin, Chung-Ching Lin, Zicheng Liu, Ce Liu, and Lijuan Wang. LAVENDER: Unifying Video-Language Understanding as Masked Language Modeling. June 2022. URL http://arxiv.org/abs/2206.07160. arXiv:2206.07160 [cs].

Kevin Lin, Linjie Li, Chung-Ching Lin, Faisal Ahmed, Zhe Gan, Zicheng Liu, Yumao Lu, and Lijuan Wang. SwinBERT: End-to-End Transformers with Sparse Attention for Video Captioning. June 2022. URL http://arxiv.org/abs/2111.13196. arXiv:2111.13196 [cs].

Xudong Lin, Simran Tiwari, Shiyuan Huang, Manling Li, Mike Zheng Shou, Heng Ji, and Shih-Fu Chang. Towards Fast Adaptation of Pretrained Contrastive Models for Multi-channel Video-Language Retrieval. April 2023. URL http://arxiv.org/abs/2206.02082. arXiv:2206.02082 [cs].

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning, December 2023. URL http://arxiv.org/abs/2304.08485. arXiv:2304.08485 [cs].

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-Centric Learning with Slot Attention, October 2020. URL http://arxiv.org/abs/2006.15055. arXiv:2006.15055 [cs, stat].

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models, June 2023. URL http://arxiv.org/abs/2306.05424. arXiv:2306.05424 [cs].

Maximilian Seitzer, Max Horn, Andrii Zadaianchuk, Dominik Zietlow, Tianjun Xiao, Carl-Johann Simon-Gabriel, Tong He, Zheng Zhang, Bernhard Schölkopf, Thomas Brox, and Francesco Locatello. Bridging the Gap to Real-World Object-Centric Learning, March 2023. URL http://arxiv.org/abs/2209.14860. arXiv:2209.14860 [cs].

Maria Tsimpoukelli, Jacob Menick, Serkan Cabi, S. M. Ali Eslami, Oriol Vinyals, and Felix Hill. Multimodal Few-Shot Learning with Frozen Language Models, July 2021. URL http://arxiv.org/abs/2106.13884. arXiv:2106.13884 [cs].

Alex Jinpeng Wang, Yixiao Ge, Rui Yan, Yuying Ge, Xudong Lin, Guanyu Cai, Jianping Wu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. All in One: Exploring Unified Video-Language Pre-training. March 2022. URL http://arxiv.org/abs/2203.07303. arXiv:2203.07303 [cs].

Xijun Wang, Junbang Liang, Chun-Kai Wang, Kenan Deng, Yu Lou, Ming Lin, and Shan Yang. VLAP: Efficient Video-Language Alignment via Frame Prompting and Distilling for Video Question Answering. February 2024. URL http://arxiv.org/abs/2312.08367. arXiv:2312.08367 [cs].

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Huai hsin Chi, F. Xia, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *ArXiv*, abs/2201.11903, 2022. URL https://api.semanticscholar.org/CorpusID:246411621.

Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual ChatGPT: Talking, Drawing and Editing with Visual Foundation Models, March 2023. URL http://arxiv.org/abs/2303.04671. arXiv:2303.04671 [cs].

Jiaqi Xu, Cuiling Lan, Wenxuan Xie, Xuejin Chen, and Yan Lu. Slot-VLM: SlowFast Slots for Video-Language Modeling. February 2024. URL http://arxiv.org/abs/2402.13088. arXiv:2402.13088 [cs].

Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. MM-REACT: Prompting ChatGPT for Multimodal Reasoning and Action, March 2023. URL http://arxiv.org/abs/2303.11381. arXiv:2303.11381 [cs].

Hang Zhang, Xin Li, and Lidong Bing. Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding. October 2023a. URL http://arxiv.org/abs/2306.02858. arXiv:2306.02858 [cs, eess].

Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. LLaMA-Adapter: Efficient Fine-tuning of Language Models with Zero-init Attention, June 2023b. URL http://arxiv.org/abs/2303.16199. arXiv:2303.16199 [cs].

Yue Zhao, Long Zhao, Xingyi Zhou, Jialin Wu, Chun-Te Chu, Hui Miao, Florian Schroff, Hartwig Adam, Ting Liu, Boqing Gong, Philipp Krähenbühl, and Liangzhe Yuan. Distilling Vision-Language Models on Millions of Videos. January 2024. URL http://arxiv.org/abs/2401.06129. arXiv:2401.06129 [cs].

Kaiwen Zhou, Kwonjoon Lee, Teruhisa Misu, and Xin Eric Wang. ViCor: Bridging Visual Understanding and Commonsense Reasoning with Large Language Models. October 2023a. URL http://arxiv.org/abs/2310.05872. arXiv:2310.05872 [cs].

Kankan Zhou, Eason Lai, Wei Bin Au Yeong, Kyriakos Mouratidis, and Jing Jiang. ROME: Evaluating Pre-trained Vision-Language Models on Reasoning beyond Visual Common Sense. October 2023b. URL http://arxiv.org/abs/2310.19301. arXiv:2310.19301 [cs].

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models, October 2023. URL http://arxiv.org/abs/2304.10592. arXiv:2304.10592 [cs].

A Appendix

A.1 Sample responses for counter-factual questions

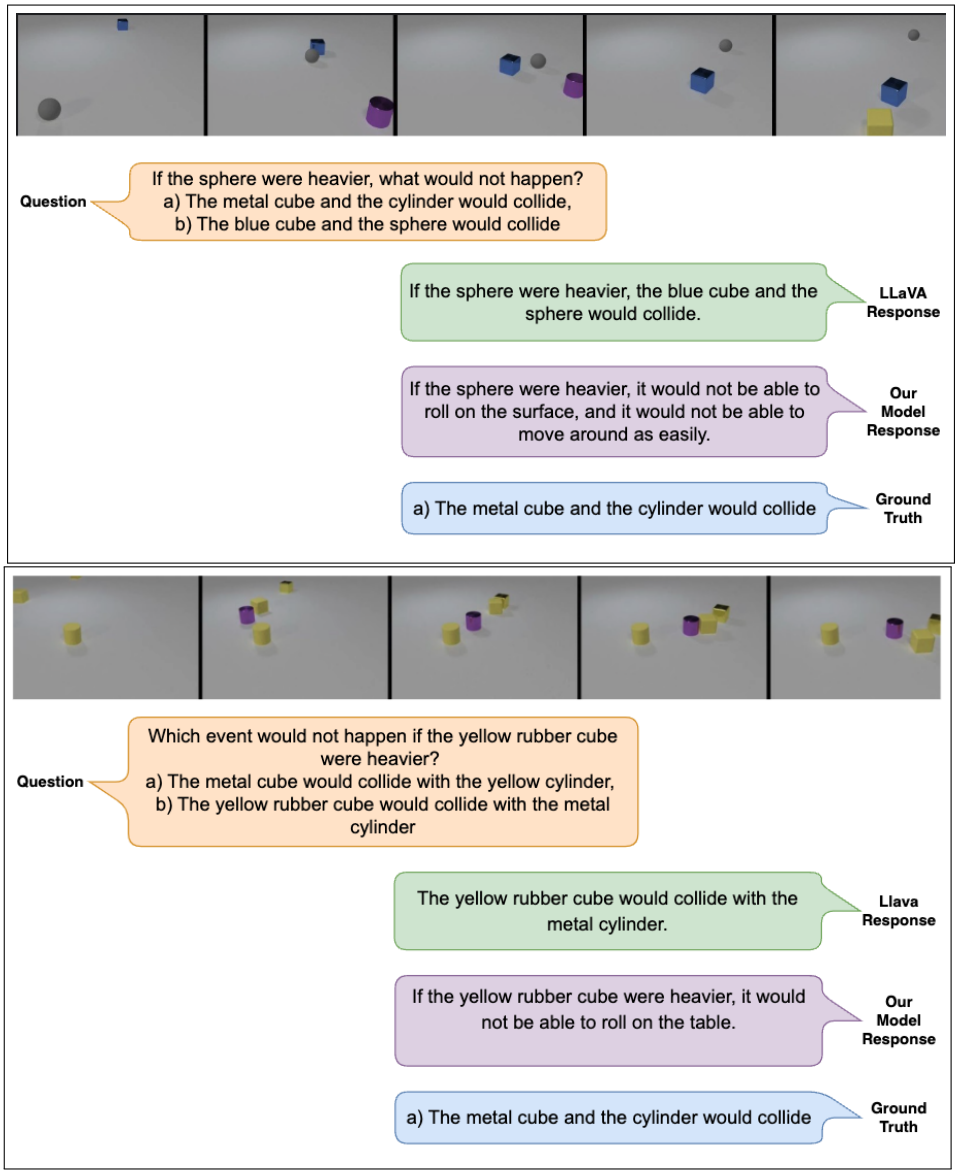

Figure 8: Sample responses for counter-factual questions

A.2 SAMPLE RESPONSES FOR FACTUAL QUESTIONS

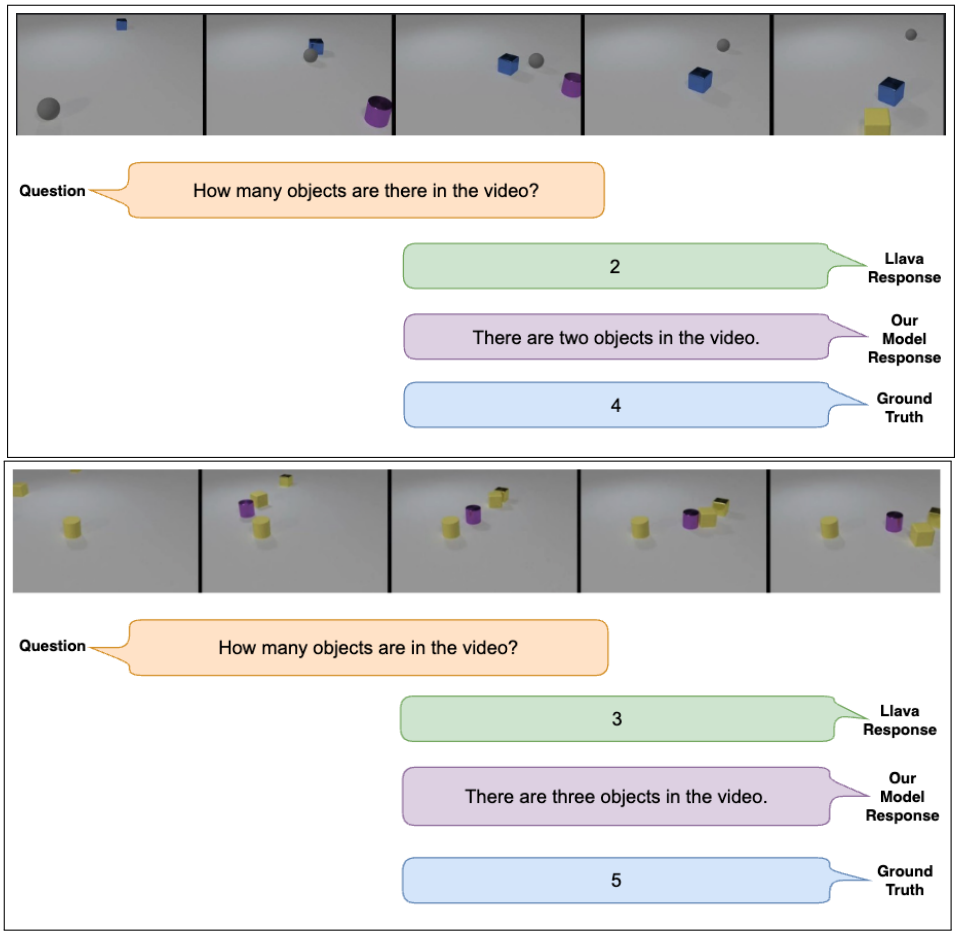

Figure 9: Sample responses for factual questions

A.3 SAMPLE RESPONSES FOR DESCRIPTIVE QUESTIONS

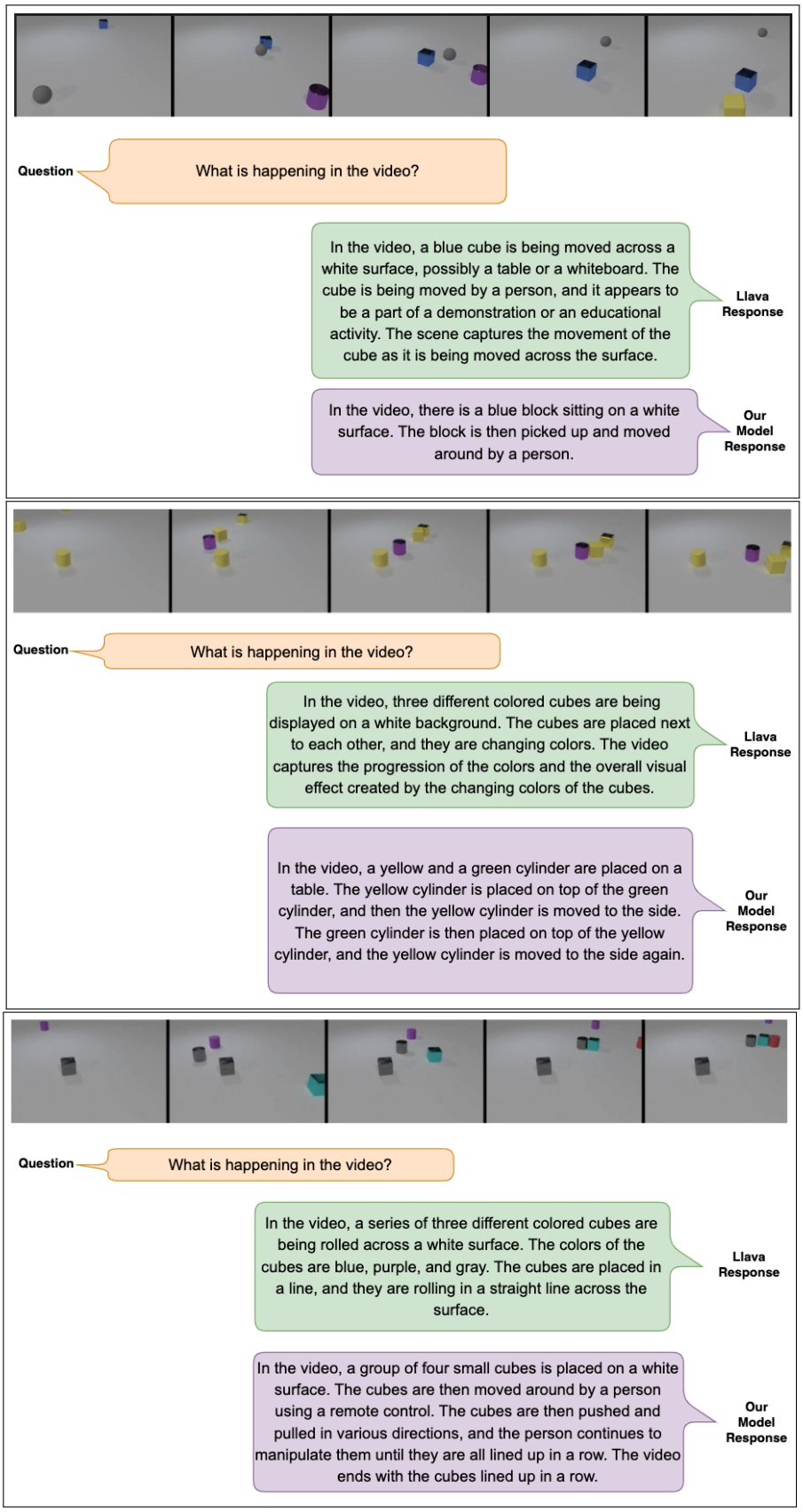

Figure 10: Sample responses for descriptive questions