Towards understanding Compositionality in Vision-Language Models

Thesis submitted in partial fulfilment
of the requirements for the degree of

Master of Science in ***Computer Science and Engineering*** *by Research*

by

Darshana Saravanan
2022701012
darshana.s@research.iiit.ac.in
Advisor: Prof. Vineet Gandhi

International Institute of Information Technology
Hyderabad - 500 032, INDIA
June 2025

Copyright © Darshana Saravanan, 2025
All Rights Reserved

International Institute of Information Technology

Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled "Towards understanding Compositionality in Vision-Language Models" by Darshana Saravanan, has been carried out under my supervision and is not submitted elsewhere for a degree.

June 2025

Advisor: Prof. Vineet Gandhi

To my Parents (Sakthipriya J, Saravanan R)

Acknowledgments

Over the past three years at IIIT Hyderabad, I have not only gained valuable knowledge but also learned the importance of perseverance and developed the confidence to choose challenging problems and work steadily towards solving them. This journey would not have been possible without the support and contributions of many people, to whom I am deeply grateful.

First and foremost, I would like to express my sincere gratitude to Prof. Vineet Gandhi for giving me the opportunity to pursue my aspirations and for laying the foundation of my scientific thinking. His unwavering guidance, both in research and in life, has been a source of motivation during challenging times. I am equally grateful to Prof. Makarand Tapaswi for shaping my understanding of the Vision-Language domain. His continuous mentorship, patience, and insightful feedback have played a crucial role throughout this journey. I also thank Prof. Naresh Manwani for his consistent support and encouragement.

This journey would not have been as enjoyable without my friends at CVIT. I am grateful to Vivek, Joy, Kawshik, Yasaswi, Siddharth, Ayan, Deepti, Adhiraj, Vansh, Akshat and Balu for making each day memorable. A special thanks to Darshan for always sharing his knowledge and involving me in our projects. Thank you, Varun, for your tireless collaboration and commitment. I amm grateful to Soumya J. for the fun conversations and for constantly pushing me to try new things. My heartfelt thanks to Anagha, Aparajitha, and Soumya V. for being there to listen and for all the wonderful times together. Thank you, Vaibhav, for the light-hearted banter, and Manu, for our interesting discussions on research. Thank you, Shyam, for our many discussions; your technical insights were invaluable, and your encouragement much appreciated.

I owe a special debt of gratitude to Ms. S. Kiran, my high school computer science teacher, whose passion for teaching strengthened my interest in this field. Her early influence set me on the path that I am on today.

I am thankful to the friends who have stood by me over the years. Thank you, Vaishnavi for always checking up on me, Nivetha for your unwavering belief in me and endless support, and Anirudhan, for being a great teammate when we studied together. A heartfelt thanks to Arthi and Sana for being constants in my life over the past seven years and for supporting me unconditionally.

Lastly, I am forever grateful to my parents, Sakthipriya and Saravanan, for encouraging and supporting me to pursue this path. They have always emphasized the value of education and provided me with the privilege they themselves were not able to access. I also thank my brother Rajendraprasad for being occasionally helpful, and my grandmother Prema for always cheering for me.

Abstract

Human intelligence relies on compositional generalization: the ability to interpret novel situations by flexibly combining familiar concepts and relational structures. This thesis investigates compositionality in vision-language models (VLMs), focusing on their ability to understand and generalise across visual (images, videos) and linguistic inputs.

In the first part, we introduce VELOCITI, a benchmark for evaluating compositional understanding in video-language models through a suite of entailment tasks. Unlike prior compositionality benchmarks constrained to single-agent videos, VELOCITI captures the complexity of real-world videos involving multiple agents and dynamic interactions. VELOCITI assesses how well models recognize and bind agents, actions, and temporal events using both text-inspired and in-video counterfactual negations.

In the second part, we probe the internal activations of VLMs to understand how concepts in an image are bound to their attributes and references in text. Extending the Binding ID mechanism in language models, we demonstrate that VLMs construct binding ID vectors in the activations of both image tokens and their textual references, enabling in-context concept association.

Together, these contributions advance our understanding of compositional reasoning in VLMs and offer tools for probing their capabilities.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

When a child watches a cartoon for the first time in which an animal is depicted preparing soup, they exhibit compositional understanding by decomposing the visual scene into fundamental conceptual elements—such as the animal, vegetables, the action of chopping, a pot, and the action of stirring. By identifying these components and their relationships, the child is able to interpret the scene in terms of "who is doing what, with what, and when." Despite lacking direct experience of animals cooking in the real world, the child demonstrates compositional generalization. By drawing upon prior knowledge of animals, food preparation activities, and the conventional roles associated with objects like vegetables and pots allows them to make sense of the unfamiliar scenario and infer that a meal is being prepared. This illustrates the human ability to generalize to novel situations by flexibly recombining familiar concepts and relational structures.

Compositionality has been regarded as a fundamental requirement for intelligence and extensively studied in the Vision-Language domain. Early works constructed visual question answering (VQA) datasets with training and compositional splits [2, 3, 4]. The train split equips the model to comprehend the visual concepts and relationships in the train data while the compositional split evaluates the extent to which the model can generalize to novel compositions of known concepts. This lead to the development of models that have explicit mechanisms to encode compositional representations [5, 6, 7, 8].

The advent of large-scale pre-trained models [9] blurred the separation between train and test data (in-distribution vs. out-of-distribution). These models seem capable of a wide range of tasks with high zero-shot accuracy. However, multiple works have pointed out their weakness in *binding*, the ability to associate concepts with the right relations [10, 11]. A series of benchmarks demonstrate their limitations in compositional reasoning via tasks that involve differentiating images and captions that have the same atomic entities but are composed with different relationships [12, 13, 14, 15, 16].

In the first part of this thesis, we present VELOCITI, a video-caption entailment benchmark for evaluating compositionality in Video-Language Models. While several works have studied compositionality for image-based models, this is a serious challenge for videos as they contain multiple persons, objects, and scenes that dynamically interact and change over time. Some recent works have attempted to study compositionality in video models by evaluating their ability to differentiate minimally differing captions. However, they restrict themselves to single-agent videos and counterfactual captions where one atomic aspect is absent in the video. To overcome this limitation, VELOCITI employs videos from the VidSitu [17] dataset that contain short yet complex videos with multiple persons and interactions. Through a suite of seven tests, VELOCITI disentangles and assess a model's ability to comprehend *agents*, *actions*, and their associations across *multiple events* through time. Further, VELOCITI uses two strategies to generate the counterfactual captions: text-inspired negation manipulates the aligned caption without looking at the video, and in-video negation, a key contribution, uses a different entity appearing in the same video. The latter requires models to bind the correct entity to a given situation and ignore the different but plausible entities that also appear in the video.

Typical vision-language compositionality benchmarks evaluate if the concepts in an image are associated with the right relationships in the representations of the model. The VQA style benchmarks pose questions that can be answered purely based on the information contained in the image. To fully exploit the potential of current multi-modal LLMs, it is useful to perform tasks that require synthesizing information from both vision and language modalities. Multi-modal QA benchmarks pose questions that need to be answered based on an image accompanied by some textual information [18, 19, 20].

In the second part of this thesis, we investigate how Vision-LLMS associate concepts across modalities in context by probing the activations of the model. Given an 'image of a red toy car', the model should associate this image to phrases like 'car', 'red toy', 'red object', etc. [21] propose the Binding ID mechanism in LLMs, suggesting that the entity and its corresponding attribute tokens share a Binding ID, a pair of vectors that have a high similarity in a subspace, in the model activations. This work investigates the mechanism for image-text binding in VLMs using a synthetic dataset and a task that requires models to associate 3D objects in an image with their descriptions in the text. The experiments demonstrate that VLMs assign a distinct Binding ID to an object's image tokens and its textual references, enabling in-context association.

1.1 Contributions

This thesis provides unique insights into the compositional reasoning capabilities of VLMs in vision-centric and multi-modal tasks. The core contributions are listed below.

- We introduce VELOCITI, a diagnostic benchmark to evaluate compositional reasoning in video-language models. VELOCITI disentangles and assesses a model's ability to comprehend *agents*, *actions*, and their associations across *multiple events* through time.
- The performance gap between text-inspired and in-video negation tests in VELOCITI reveals the limitations of current models' ability to bind semantic concepts, even within a short 10-second video.
- We investigate how VLMs synthesize multi-modal information in context. Experiments demonstrate that VLMs utilize binding ID vectors to bind concepts across an image and text.
- We establish that image tokens corresponding to the location of the visual concept represent information related to that concept.

1.2 Organization of the thesis

The rest of the thesis is organized as follows:

1. In Chapter 2, we review the prior work done on understanding and evaluating compositional reasoning in VLMs while highlighting the role of this thesis in expanding our knowledge.
2. In Chapter 3, we introduce VELOCITI, a benchmark for evaluating compositional reasoning in Video-Language models.
3. In Chapter 4, we present our work explaining how Vision-Language models associate objects in an image with their references in text in-context.
4. In Chapter 5 we discuss the concluding remarks.

Chapter 2

Vision-Language Compositional understanding through the ages

In this chapter, we present a broad view of the work done in the domain of compositional reasoning in VLMs and highlight the research gaps this thesis addresses. We discuss the prominent benchmarks used to evaluate compositional understanding in Section 2.1, approaches to improving compositionality in Section 2.2 and controlled studies conducted to understand compositional reasoning in Section 2.3.

2.1 Benchmarks for Compositional Reasoning

Compositional reasoning has garnered significant attention within the vision-language domain, as it is fundamental to understanding relationships within and across visual and linguistic inputs. Although language is consistently employed across benchmarks as the medium for both formulating questions and presenting answers, the following paragraphs categorize prior works based on the modality (images, videos, or multi-modal) that primarily encapsulates the information necessary to arrive at the correct answer.

Images. Shifting the focus beyond traditional tasks such as image classification and object detection, the Visual Genome dataset [22] introduced rich annotations for attributes and relationships through scene graphs, enabling a more holistic understanding of visual scenes (e.g., distinguishing between "man chasing dog" and "dog chasing man"). Building on this foundation, the GQA dataset [23] generates compositional questions by traversing scene graphs, thereby requiring models to perform multi-step reasoning. CLEVR [24], a synthetic visual question answering dataset, emphasizes compositional visual reasoning by focusing on tasks such as attribute identification, counting, and logical comparisons. Complementing these efforts, Winoground [12] frames image-caption matching as a fine-grained reasoning task, where two captions—containing the same words in different orders—must be correctly matched to one of two images based on subtle relational differences. While Winoground consists of only 400 handcrafted examples, more scalable datasets such as ARO [13], CREPE [25], and VL-Checklist [26] leverage existing annotations from datasets like Visual Genome, SWiG [27], COCO [28], and Flickr [29] to create broader image-caption matching tasks that evaluate models' compositional understanding. ARO tests reasoning over attributes, relationships, and word order by generating incorrect captions through word swaps or shuffling. VL-Checklist introduces errors by altering a specific object, attribute, or relationship in the caption, while CREPE adopts a similar strategy using negations, replacements, and swaps. However, these perturbation-based methods were found to introduce biases in the form of implausible or grammatically incorrect captions, which models could exploit to artificially boost performance. To address this, SugarCREPE [14] was proposed with improved distractor generation techniques. Unlike these image-to-text tasks, Cola [30] evaluates compositionality in the reverse setting—text-to-image retrieval—where models must identify the correct image from a set based on a compositional text query.

Videos. Building on the idea of structured scene representations introduced by Visual Genome, Action Genome [31] extends this framework to the video domain by providing spatio-temporal scene graphs. These graphs aim to decompose actions over time by annotating specific video segments and identifying only the objects directly involved in each action. Leveraging this representation, the AGQA dataset [2] poses compositional questions that require models to reason over sequences of interactions between entities to produce accurate answers. AGQA-Decomp [32] further advances this idea by evaluating the consistency of a model's responses to sub-questions derived from a single compositional query about a visual event. In parallel, synthetic datasets like CATER [33] focus on long-term compositional reasoning by tracking complex object movements over time. Benchmarks such as Contrast Sets [34], VideoCon [35], VITATECS [36], and TempCompass [15] introduce tasks that challenge models to distinguish between captions that differ subtly in temporal attributes such as action type, speed, direction, or event sequence. Inspired by the design of Winoground, Vinoground [16] introduces a video-caption matching task involving two captions and two video frames, where success depends on recognizing temporal differences in actions and object transformations.

Distinct from these prior efforts, our work VELOCITI introduces a novel challenge by focusing on multi-agent, multi-action interactions within video scenes. Unlike most benchmarks that use single-agent videos [2, 34, 36, 15, 16], VELOCITI includes more complex scenarios involving multiple interacting agents. Furthermore, it incorporates both text-inspired and in-video negation tests, designed to disentangle and highlight specific weaknesses in recognizing and reasoning over temporally and compositionally rich content.

Multi-modal. In parallel with these vision-centric efforts, multi-modal question answering datasets have emerged to evaluate reasoning across both visual and textual modalities. MMQA [18] presents a challenging benchmark requiring joint reasoning over text, tables, and images. MUMUQA [19] focuses on real-world news contexts, where grounding across image-caption pairs is required to correctly identify entities being referred to in the question and reason about their relationships to arrive at the answer. VTQA [20] expands this further by introducing diverse, multi-hop questions spanning a variety of genres and topics, necessitating deep cross-modal compositional reasoning.

Together, these image, video, and multimodal benchmarks reflect the increasing complexity in evaluating compositional understanding, emphasizing the need for models that can truly understand the real world.

2.2 Approaches to modelling compositionality

Early research into compositional understanding in vision-language models (VLMs) emphasized designing neural architectures that explicitly reflect the compositional structure of language and vision. Neural Module Networks (NMNs) [5] introduced a question answering framework capable of dynamically assembling task-specific networks from a set of composable modules based on the structure of natural language queries. Similarly, the MAC network (Memory, Attention, and Composition) [7] approached reasoning tasks by decomposing them into a sequence of attention-driven steps, each executed by a recurrent MAC cell that separates the "control" component responsible for selecting the reasoning operation from the "memory" component, which aggregates intermediate results. SCAN (Symbol-Concept Association Network) [6] further advanced compositional learning by enabling neural models to acquire grounded visual concepts through largely unsupervised learning of symbol associations. Expanding this line of work to video understanding, HCRN [8] introduced Conditional Relational Networks (CRNs), a general-purpose neural unit used to construct hierarchical architectures for VideoQA, enabling multi-level reasoning through modular composition.

With the emergence of CLIP, research in compositional understanding has increasingly shifted toward improving training paradigms rather than architectural design. A prominent strategy involves fine-tuning models using minimally altered, misaligned captions to enhance their sensitivity to fine-grained compositional cues [13, 37, 38, 39, 40]. Similar techniques have also been applied to video understanding, targeting both dual-encoder models [41] and Video-LLMs [35]. The Dense and Aligned Captions (DAC) approach [42] enhances the quality and richness of captions in existing datasets and fine-tunes models using Multiple Instance Learning and negative text augmentation. CE-CLIP [43] introduces intra-modal contrastive loss and cross-modal ranking loss to strengthen alignment between modalities. FSC-CLIP [44] addresses the limitations of relying on single global representations by utilizing dense cross-modal representations and a hard negative loss regularization strategy to better preserve multi-modal capabilities. IL-CLIP [45] takes a novel approach by iteratively learning a shared codebook and distilling from it, encouraging the encoders to adopt smoother, lower-frequency approximations of their previous states—an approach hypothesized to promote more compositional representations over time. Another active line of work focuses on synthesizing misaligned or minimally varied image-text pairs for model fine-tuning [46, 47, 48, 49, 50]. SPARCL [51] proposes a synthetic data generation pipeline that produces subtly varied image-text pairs, specifically targeting inconsistencies in cross-modal alignment.

2.3 Controlled studies on compositionality in VLMs.

Concept Association Bias (CAB) refers to a phenomenon observed in vision-language models (VLMs), such as CLIP and BLIP-2, wherein the models attempt to infer missing concepts when there is a discrepancy between the visual and textual inputs [10]. For example, when shown an image containing both a lemon and an eggplant, and prompted with the question "What color is the lemon?", CLIP may incorrectly predict "purple" with high confidence. This error arises because CLIP associates "purple" more strongly with the eggplant, and the text description "purple lemon" aligns more closely with the combined image content than the correct but partial description "yellow lemon." Another work employs controlled synthetic images to show that while CLIP performs well on simple adjective-noun compositions, it struggles with compositional reasoning tasks involving multiple objects and spatial relationships, which require accurate variable binding [11]. Further investigation reveals that although correct attribute-object bindings exist within the individual image and text modalities, CLIP's Bag-of-Words like behavior stems from limitations in cross-modal alignment, which depends on cosine similarity. Enhancing this alignment by applying a linear transformation to text embeddings prior to similarity computation significantly improves CLIP's ability to bind attributes to their corresponding objects.

While these studies shed light on the behavior of contrastive VLMs, the behavior of auto-regressive Vision-LLMs remains less explored. In this thesis, we extend the Binding ID Mechanism [21], originally proposed for language models, to demonstrate that VLMs also construct binding ID vectors in the activations of both image tokens and their textual references, enabling in-context concept association.

Chapter 3

VELOCITI: Benchmarking Video-Language Compositional Reasoning

Near a parking lot, a man in a black hat smiles in a friendly way at a woman in a purple shirt. To a reader, this dense description paints a clear picture about a short snippet (event) of a video clip. We build a mental model of two people (referred here by their clothing), at a specified location, and a short interaction between them. Reading further, *the woman claps as a man in grey pants spins on one leg*. We are able to associate that it is the same woman who is now cheering at a third person (likely) that is performing stunts.

The above example illustrates an intelligent agent's ability to perform compositional reasoning. For video-language models, we scope this in two steps: (i) comprehend atomic entities, *e.g. people* and *actions*; and (ii) reason about them compositionally and across time by building associations[1].

In recent years, strong visual (image) encoders are combined with powerful Large Language Models (LLMs) to advance general-purpose vision [52, 53, 54, 55, 56]. A similar approach is adapted for videos to create Video-LLMs [57, 58, 53, 54]. Keeping pace with the development of new models, there is a flurry of work on evaluating them. Video researchers are also creating benchmarks to study long video comprehension [59, 60, 61, 62]. However, we take a step back and ask, *are today's Video-LLMs ready to take on such challenges?* Specifically, are they good at compositional reasoning in short videos, arguably a prerequisite to tackle complex and long videos?

To this end, we introduce the *VELOCITI*, a benchmark that studies **V**ideo **e**t **L**anguage **C**ompos**i**tionality through **Ti**me. We adopt the video-language entailment (VLE) evaluation setup [35] where a model is prompted to predict whether a video entails a caption. Through a suite of seven tests, we are able to disentangle and assess a model's ability to comprehend *agents*, *actions*, and their associations across

[1] Associations can be thought as *implicit* tuples that a model attempts to build while watching a video. Some examples include *person-attribute* tuples: (man1, black hat), (woman, purple shirt), (man2, grey pants); *agent-action* tuples: (man1, smiles at, woman), (man2, spins on one leg), (woman, claps at, man2); or *action-manner* tuples: (smile, friendly way).

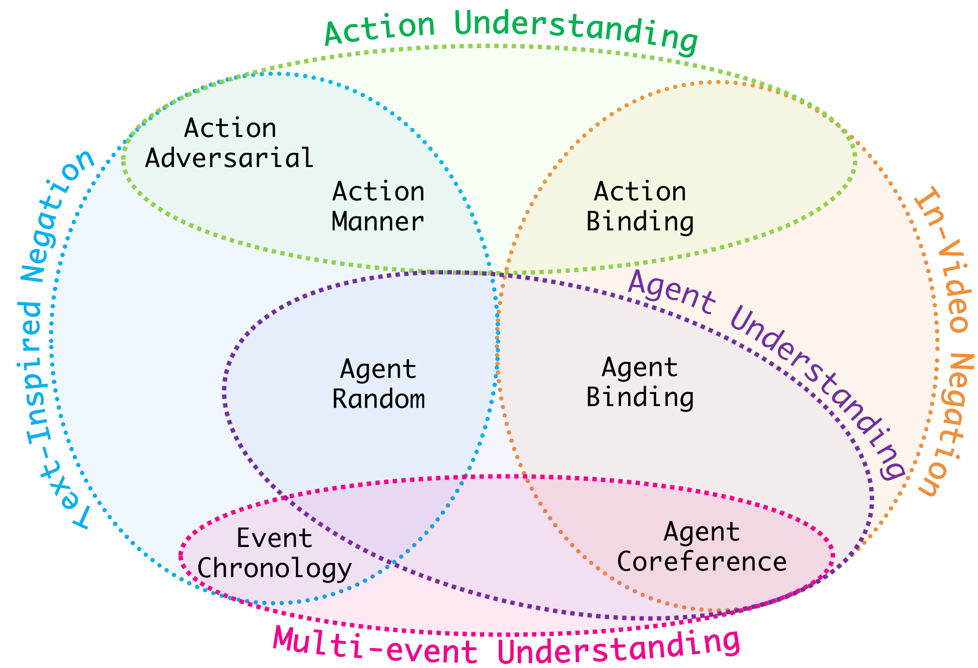


Figure 3.1. A Venn diagram grouping VELOCITI's seven tests (in black) that evaluate a model across different facets: Agent Understanding, Action Understanding, and Multi-event Understanding. The benchmark is formulated as video-language entailment, where negative captions are created by manipulating text (Text-inspired Negation) or from other parts of the same video (In-Video Negation). Best seen in color.

multiple events through time. As illustrated in Fig. 3.1, we group the 7 tests based on: (i) the specific facet of a model's ability (agent, action, multi-event), or (ii) the strategy used to create the negative caption (Text-Inspired Negation *vs*. In-Video Negation).

Our videos are sourced from the VidSitu dataset and are accompanied by action and semantic role label (SRL) annotations for multiple events in a short movie clip [1]. The videos are diverse and feature multiple agents and actions across complex editing and shot changes, while dense SRL succinctly describes *who* did *what* with/to *whom*, *where*, and (sometimes) *how*. Importantly, each SRL only describes a single event, requiring models to implicitly localize the event in the video before solving the test.

Contributions. We summarize our contributions and findings below: (i) We propose VELOCITI, a new benchmark that evaluates compositional reasoning of video-language models. Our test suite sheds light on a model's ability to perceive and reason about *agents* and *actions* across *multiple events*, identifying challenges for improvement (Section 3.1). (ii) Our experiments reveal important findings: a) Understanding actions is harder than agents and b) tests incorporating in-video negation are more challenging than text-inspired negation.

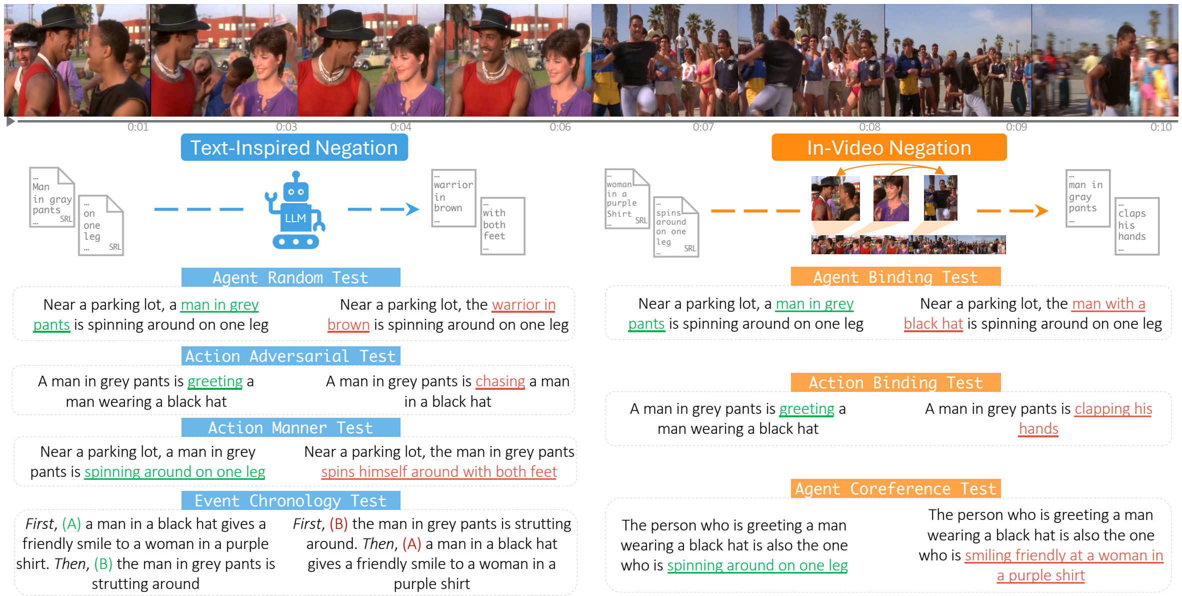


Figure 3.2. VELOCITI evaluates Video-LLMs' video-language entailment capabilities on complex movie clips with dense semantic role label (SRL) annotations from the VidSitu dataset [1]. Positive and negative captions are shown side-by-side for each test with the key difference highlighted with green/red. Negative captions are created by (i) manipulating text using an LLM (Text-Inspired Negation) or (ii) replacing agents or actions by others that appear in the same video (In-Video Negation). We also demonstrate how the same positive caption can be used to create negative captions differently (see Agent Random *vs*. Agent Binding test; or Action Adversarial *vs*. Action Binding test). Each test evaluates models for different facets of compositional reasoning as described in Section 3.1.2. The 10 s video clip used in this example can be viewed here: `https://www.youtube.com/embed/bt6-F11LZsQ?start=25&end=35`.

3.1 VELOCITI Benchmark

We evaluate compositional reasoning using dynamic 10 s movie clips and SRL annotations from the VidSitu dataset. We propose *seven tests* to evaluate model's comprehension of agents and actions across multiple events through time. Each test consists of $\{V, C^+, C^-\}$: video clip V, a positive caption C^+ that is aligned with a part of the video, and a negative caption C^- that is *not* aligned to the video. We require models to *independently* assess each caption and classify them as V entails C^+ *and* V does not entail C^-.

3.1.1 From SRL to a Video-Caption Pair

In VidSitu, videos are divided into five 2 s events [1] (total 10 s duration). Each event is annotated by the most salient action and the corresponding SRL capturing: *who* is doing the action (agent), *with / to whom* (patient or receiver), *with what* (instrument, if applicable), *where* (scene or location), *how* (manner or adverb), and *why* (purpose).

We use an open LLM (LLaMA-3 [63]) to convert the structured SRL dictionary of each event into a caption. The LLM is prompted to combine the atomic concepts into a fluent caption (prompt in the supplement). We filter 864 videos from the validation set and generate 3101 high-quality (V, C) pairs with captions that are faithful to the SRL annotations. Depending on the test, these captions are directly used as C^+ or used to *form* C^+.

Note, movie events are not bounded by 2 s intervals and the SRL annotations may spill into the neighboring events. Thus, we make a conscious choice to pair the caption C with the entire 10 s video V. To correctly decide whether V entails C, a model needs to implicitly *localize* to the appropriate temporal region in the video. This prevents a single frame bias as reported by Atemporal Probe in [64].

3.1.2 VELOCITI Tests

We motivate and describe the seven tests below. Fig. 3.2 shows an example of each test grouped based on the process used to create C^-: (i) *text-inspired negation* typically creates C^- without looking at the video; and (ii) *in-video negation*, a key contribution of our work, uses a different entity appearing in the same video to create C^-. Both are important as they help us identify pitfalls of current models.

0. Control Test. We start with a control test to establish a baseline understanding. Here, C^+ is as described in Section 3.1.1 and C^- is simply a positive caption of some other random video, making it easily discernible.

1. Agent Random Test(AgRand). C^- is created by replacing the correct *agent* with another agent that does not appear in the video (C^+ is as above). Solving this test requires a model to implicitly localize the event based on the action and identify who is present/absent in the video. We ensure that the replacing agent is not a hypernym (*e.g.*, “man in a shirt” is not replaced by “man”). The SRL dictionary is updated with the random agent and the LLM generates C^-.

2. Agent Binding Test(AgBind) also replaces the agent. Different from above, the replaced agent is chosen from the *same video* making it an in-video negation. This subtle difference requires models to identify the correct agent and bind or associate it with the event description. Models cannot rely purely on presence/absence to solve this task. Fig. 3.2 shows how the same C^+ can be modified to create both agent tests. Similar to above, the LLM generates C^-.

3. Agent Coreference Test(AgCref). Coreference groups two or more phrases that refer to the same entity [65]. In a video, an agent can be referred to by their actions, *e.g.* in Fig. 3.2, the agent: *man in grey pants* is referred by: *the person who is* (i) *greeting a man wearing a black hat* or (ii) *spinning around on one leg*. To create this test, we identify videos with the same person acting in two or more events and construct two references for that person. C^+ is formed by combining the referring expressions of the *same* agent, while C^- combines referring expressions of *different* agents. This test also features complex in-video negation as all concepts mentioned in both C^+ and C^- appear in the video. The captions are created using the template: *The person who is [Event A] is also the one who is [Event B].* Solving this test requires models to associate the correct interactions of an agent across two events. Since the agent description is masked by *the person*, a model requires multi-level compositional reasoning, making this test particularly challenging.

4. Action Adversarial Test(ActAdv). C^+ is as described in Section 3.1.1 and C^- is created by replacing the *action* with an adversarial alternative (a plausible action determined through the text description) that does not appear in the video. Solving this test requires identifying the action that the agent is performing. Given the SRL dictionary, the LLM is prompted to first generate the adversarial action followed by C^-.

5. Action Manner Test (ActMan) typically features a C^+ that includes an adverb, emotion, or facial expression. C^- is generated by replacing this manner with a contradictory yet plausible alternative. Solving this test is challenging as it requires understanding subtle variations in an action. Similar to the test above, the LLM is prompted to first generate the contrasting manner followed by C^-.

6. Action Binding Test (ActBind). Here, C^- is created by retaining the agent from C^+ and swapping the action and its modifiers with those from a different event within the *same video*. Solving this test requires models to localize events where the agent appears and bind them with the correct action. This is another test with in-video negation as actions described in both C^+ and C^- appear in the video. To create C^-, we identify an event in the same video with a different action performed by a different agent. Next, we replace the SRL dictionary of C^+ with the action (and relevant modifiers) and prompt the LLM to generate C^-.

7. Event Chronology Test (EvChr). Our final test studies a model's ability to confirm whether the video and caption follow the same event progression. Multiple events descriptions can be related through time using *before, after, first, then* [66]. C^+ is created by concatenating event captions (from Section 3.1.1) with the template *"First, [Event A]. Then, [Event B]."* where event A *precedes* B. C^- simply reverses them to *"First, [Event B]. Then, [Event A]."* The events are sampled at least $2\,\mathrm{s}$ apart to prevent a chance of overlap.

3.2 Results and Discussion

In this section we first describe the evaluation metric and then discuss the results.

3.2.1 Evaluation Metric

We adopt Video-Language Entailment (VLE) as the evaluation scheme for VELOCITI. For contrastive models, the alignment between a video and caption is measured using cosine similarity. For generative models, we define the entailment score similar to [67]. Given an instruction I containing a video V and a caption C, model M is prompted to answer whether the video entails the caption through 'Yes'/'No'.

$$e(V,C) = \frac{p_M(\text{'Yes'}|I(V,C))}{p_M(\text{'Yes'}|I(V,C)) + p_M(\text{'No'}|I(V,C))}, \tag{3.1}$$

where p_M denotes the model's probability distribution over the entire vocabulary.

The instruction is described in Fig. 3.3. Our metric considers that a model is correct when $e(V,C^+) > e(V,C^-)$. The random accuracy is 50%.

Entailment Prompt

Carefully watch the video and pay attention to the sequence of events, the details and actions of persons.

Here is a caption that describes the video: $Caption$

Based on your observation, does the given video entail the caption?

Figure 3.3. Entailment evaluation prompt.

Model	Ctrl	Ag Rand	Ag Bind	Ag Cref	Act Adv	Act Man	Act Bind	Ev Chr	Avg
Random	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0
VERA	50.9	58.4	53.4	63.7	67.6	58.3	53.3	53.4	58.3
SigLIP	95.3	79.0	54.4	50.4	66.4	55.0	54.0	48.2	58.2
ViFi-C	93.7	82.8	58.9	56.3	63.2	60.3	59.0	48.1	61.2
Neg-C	93.4	83.5	55.3	50.4	61.6	61.1	52.4	50.1	59.2
P-LLaVA	90.7	74.6	48.9	63.7	71.0	57.0	51.8	49.8	59.5
V-LLaVA	89.7	75.1	49.6	64.3	61.6	48.5	52.0	53.8	57.8
Owl-Con	90.8	73.2	49.9	48.1	72.4	61.8	52.7	42.5	57.2
QVL-7B	97.7	93.0	74.6	63.7	75.3	76.2	70.0	63.5	73.8
OV-7B	98.6	94.4	78.8	69.0	79.7	76.9	74.2	**84.0**	79.6
OV-72B	**99.4**	**95.8**	**83.3**	**80.5**	**84.2**	**81.2**	**78.4**	81.9	**83.6**

Table 3.1. Accuracy on VELOCITI. Random accuracy is 50%. The performance of contrastive models and older Video-LLMs is close to random. However, recent models (*e.g.* OV) produce better relative entailment scores.

3.2.2 Results

Models. We evaluate both contrastive models (ViT-SO400M-14-SigLIP-384 (SigLIP) [68], VIFICLIP-B16 (ViFi) [69], and NegCLIP-B32 (Neg-C) [13])) and open Video-LLMs (PLLaVA [57], Video-LLaVA (V-LLaVA) [58], Owl-Con [35], Qwen2-VL (QVL) [53], and LLaVA-OneVision (OV) [54]).

Sanity checks. We present a language only baseline that evaluates if C^+ is more plausible than C^-. VERA [70] scores 58.3% (close to random 50%), confirming that language biases are insufficient to solve the tests. Further, almost all models achieve over 90% in the control test, demonstrating that this task is not out-of-distribution.

Agent understanding tests. Contrastive models and older Video-LLMs perform decently in the AgRand test; however, their performance drops to near random in AgBind and AgCref tests. New Video-LLMs such as QVL and OV achieve over 90% in the AgRand test. Although they perform well above chance in AgBind and AgCref, the drop in performance is significant compared to the AgRand test.

Action understanding tests. There is a trend of gradual increase of performance in ActAdv test. Since the replacements here are not as random as in AgRand, even the best model, OV-72B reaches only upto 84% compared to the 95.8% in AgRand test. For all models, distinguishing mannerisms of actions

(ActMan) is more difficult than distinguishing subtle actions (ActAdv). In ActBind test, contrastive models and the older Video-LLMs perform randomly. QVL and OV perform significantly better but their performance is still lower than in ActAdv, indicating that associating actions to agents is more difficult than distinguishing subtle actions for the current models.

Event Chronology Test test. In EvChr test, by design, CLIP-based video models that use mean-pooling fail since the order information is lost. Older Video-LLMs also perform close to random while OV-7B and OV-72B show vast improvement.

General trends. From Table 3.1 we can clearly observe that action understanding is more difficult than agent understanding. Further, in both agents and actions, tests involving in-video negations is significantly more difficult than text-inspired negations, indicating that binding agents and actions within a short 10 svideo is still a challenge.

3.3 Benchmark Creation and Details

In this section, we provide details about our benchmark. In particular, we share all prompts used for creating positive captions and various tests (Section 3.3.1, Section 3.3.2), provide benchmark statistics (Section 3.3.3), discuss the strategy used to manually verify and clean all the tests (Section 3.3.4).

3.3.1 Prompt for Converting SRL Dictionary to a Positive Caption

The prompt for generating the positive caption given an SRL dictionary is shown in Fig. 3.4. This refers to the discussion from Sec. 3.1 in the main paper. We use a two-stage strategy that first inserts all elements of the SRL dictionary in a sentence and then refines it for proper grammatical structure.

3.3.2 Prompts for Creating Test Samples

The prompt above (Fig. 3.4) helps create the positive caption for multiple tests. Specifically, Agent Random Test, Agent Binding Test, Action Adversarial Test, Action Manner Test, and Action Binding Test, all use the above strategy, while Agent Coreference Test and Event Chronology Test adopt templates that are filled in with the complete (or partial) positive captions.

The negative prompts for Agent Random Test, Agent Binding Test, and Action Binding Test are also created in the same manner as above by first replacing the specific `Verb` or `Arg0` in the dictionary followed by strategy above.

Test	Videos	# Samples	Subset
Ctrl	850	2635	150
AgRand	588	873	150
AgBind	615	1459	150
AgCref	183	339	150
ActAdv	355	438	150
ActMan	378	458	150
ActBind	540	1356	150
EvChr	521	1234	150

Table 3.2. Number of videos and samples across different tests in VELOCITI.

Finally, the prompt for generating the Action Adversarial Test negative caption is shown in Fig. 3.5 and for Action Manner Test negative captions in Fig. 3.6. Both involve generating a `target` replacement that seems reasonable followed by converting the SRL dictionary into a caption.

3.3.3 Benchmark Statistics

We present some statistics highlighting the diversity and nuance in the VELOCITI benchmark. Since this benchmark is a subset of VidSitu [1], we observe similar trends as presented in their work.

Fig. 3.7a shows that around 87% of the videos contain 4 or more unique verbs, and Fig. 3.7b shows that about 85% of videos contain 2 or more unique agents (people performing actions). We evaluate binding by leveraging the fact that one agent can perform multiple actions in the video, and the richness of the SRL annotations ensure that these events are described adequately. In Fig. 3.7c, we observe that over 70% of the events contain 4 or more SRLs (*e.g.* agent, patient, manner, *etc.*), indicating the detail-oriented nature of the annotations. Finally, Fig. 3.7d shows that over 72% of agents occur twice or more in their corresponding video annotation. These agents would likely be performing two different actions, and we utilize this to create two references to the same agent in tests such as Agent Coreference Test.

3.3.4 Quality Control

To ensure that the data generated from the automated pipelines discussed earlier are correct, we filtered the data samples manually, following specific guidelines discussed in this section. The final count of the data samples is reported in Tab. 11.

SRL dictionary to caption. The instructions and the interface for evaluating caption quality is described in Fig. 3.8. For each sample, three choices were provided: positive if the caption is correct, negative if the caption is wrong, and neutral if the caption cannot be negative but contains some ambiguity due to which it could not be considered positive. Out of the 380 samples that were manually verified, 356 were marked as positive, 21 were neutral, and 3 were negative. The number of positive and neutral samples was high (99.2%).

All tests. For each sample of all tests, we perform a meticulous cleanup. The instructions and the interface are presented in Fig. 3.9. For each video, the green bar contains a positive caption, and the red bar contains a negative caption. The positive and the negative captions are known while filtering. Only the samples for which both positive and negative captions are deemed appropriate are retained. All test samples in VELOCITI are verified by humans to ensure that C^+ aligns with the video and C^- is misaligned.

3.4 Conclusion

We introduced VELOCITI, a benchmark to evaluate the compositional capabilities of video-language models by disentangling and assessing the comprehension of agents, actions, and their associations across multiple events. Our experiments showed that action understanding is harder than agent understanding, and solving tests with in-video negation is harder than text-inspired ones. Overall, our work established that compositional reasoning on short videos is still unsolved and remains challenging for video-language models.

System Prompt

Using the provided dictionary containing verb and argument-role pairs in the style of PropBank, follow these steps to generate two captions

Naive Caption: Generate a caption that faithfully reflects all details from the dictionary without adding or omitting any information. Ensure that every argument detail is accurately included in the Naive Caption.

Fluent Caption: If the Naive Caption is already fluent and naturally phrased, directly copy it to the Fluent Caption. If necessary, refine the Naive Caption for improved language fluency while strictly maintaining all original details and arguments from the dictionary.

Please proceed with generating the Naive Caption first, ensuring it remains comprehensive and accurate based on the provided dictionary entries. Then, if adjustments are needed to enhance fluency, refine the Naive Caption into the Fluent Caption while ensuring that no details are overlooked or omitted.

Few Shot Example 1

```
{'Verb':'walk (walk)',
'Arg0 (walker)':'man in suit',
'ArgM (direction)':'into room',
'ArgM (manner)':'slowly',
'Scene of the Event':'Warehouse'}
```

Naive Caption: In a warehouse, a man in suit is walking slowly into the room.
Fluent Caption: In a warehouse, a man in suit is walking slowly into the room.

Few Shot example 2

```
{'Verb':'burn (cause to be on fire)',
'Arg0 (thing burning)':'Wreckage',
'ArgM (location)':'Wreckage'}
```

Naive Caption: The wreckage is burning on the wreckage.
Fluent Caption: The wreckage is burning.

Figure 3.4. Prompt to generate the positive caption given an SRL dictionary.

System Prompt

Your objective is to generate a contradiction caption using the provided PropBank style "input dictionary" and the 'Verb' labelled as 'source' based on a specific "misalignment scenario" called "verb misalignment". In this scenario, you should suggest an alternative contradictory value for the "source" and label it as "target".

Key Requirements

1. "naive caption + verb misalignment": should be plausible and could theoretically occur in real life.
2. The "fluent caption + verb misalignment": If the "naive caption + verb misalignment" is already fluent and naturally phrased, directly copy it to the "fluent caption + verb misalignment". If necessary, refine the "naive caption + verb misalignment" for improved language fluency while strictly maintaining all original details and arguments from the dictionary

Guidelines

1. The "target" should introduce a contradiction when compared to "source", without being a mere negation.
2. The "naive caption + verb misalignment" should be clearly distinguishable from the scene described by the "input dictionary" and should be visually distinguishable.
3. Your replacements should be creative yet reasonable.
4. If adjustments are needed to enhance fluency, refine the "naive caption + verb misalignment" into the "fluent caption + verb misalignment" while ensuring that no details are overlooked or omitted.

Few Shot Example 1

```
{'Verb': 'speak (speak)'},
'Arg0 (talker)': 'a man with dark hair',
'Arg2 (hearer)': 'old man'
'ArgM (manner)': 'greeting him',
'Scene of the Event': 'warehouse'}
```

Target: Ignore

Naive Caption: On the front porch, a man with dark hair is ignoring an old man, greeting him.

Fluent Caption: On the front porch, a man with dark hair is ignoring an old man.

Few Shot Example 2

```
{'Verb': 'open (open)',
'Arg0 (opener)': 'woman with long hair',
'Arg1 (thing opening)': 'the front door',
'ArgM (manner)': 'slowly',
'Scene of the Event': 'inside a house'}
```

Target: Close

Naive Caption: Inside a house, a woman with long hair is closing the front door slowly.

Fluent Caption: Inside a house, a woman with long hair is closing the front door slowly.

Figure 3.5. Prompt to generate the negative caption for Action Adversarial Test.

System Prompt

Your objective is to generate a contradiction caption using the provided PropBank style "input dictionary" and the 'ArgM (manner)' labeled as 'source' based on a specific "misalignment scenario" called "manner misalignment". In this scenario, you should suggest an alternative contradictory value for the "source" and label it as "target"

Key Requirements

1. "naive caption + manner misalignment": should be plausible and could theoretically occur in real life.
2. The "fluent caption + manner misalignment": If the "naive caption + manner misalignment" is already fluent and naturally phrased, directly copy it to the "fluent caption + manner misalignment". If necessary, refine the "naive caption + manner misalignment" for improved language fluency while strictly maintaining all original details and arguments from the dictionary.

Guidelines

1. The "target" should introduce a contradiction when compared to "source", without being a mere negation.
2. The "naive caption + manner misalignment" should be clearly distinguishable from the scene described by the "input dictionary."
3. Your replacements should be creative yet reasonable.
4. If adjustments are needed to enhance fluency, refine the "naive caption + manner misalignment" into the "fluent caption + manner misalignment" while ensuring that no details are overlooked or omitted

---

Few Shot Example 1

```
{'Verb': 'look (vision)',
'Arg0 (looker)': 'a man wearing all black',
'Arg1 (thing looked at or for or on)': 'a building'
'ArgM (direction)': 'infront of him',
'ArgM (manner)': 'breathing heavily',
'Scene of the Event': 'warehouse'}
```

Target: *Whistling*

Naive Caption: Outside, a man wearing all black is looking in front of him at a building while whistling.

Fluent Caption: Outside, a man wearing all black is looking at a building in front of him while whistling.

Few Shot Example 2

```
{'Verb': 'burn (cause to be on fire)',
'Arg0 (agent, entity causing something to be suspended)': 'climbing ropes',
'Arg1 (thing suspended)': 'woman in pink shirt',
'Arg2 (suspended from)': 'climbing ropes',
'ArgM (location)': 'on the face of the rocks',
'ArgM (manner)': 'precariously'}
```

Target: Securely

Naive Caption: climbing ropes are hanging the woman in a pink shirt securely on the face of the rocks.

Fluent Caption: The woman in a pink shirt is hanging on the face of the rocks from the climbing ropes securely.

Figure 3.6. Prompt to generate the negative caption for Action Manner Test.

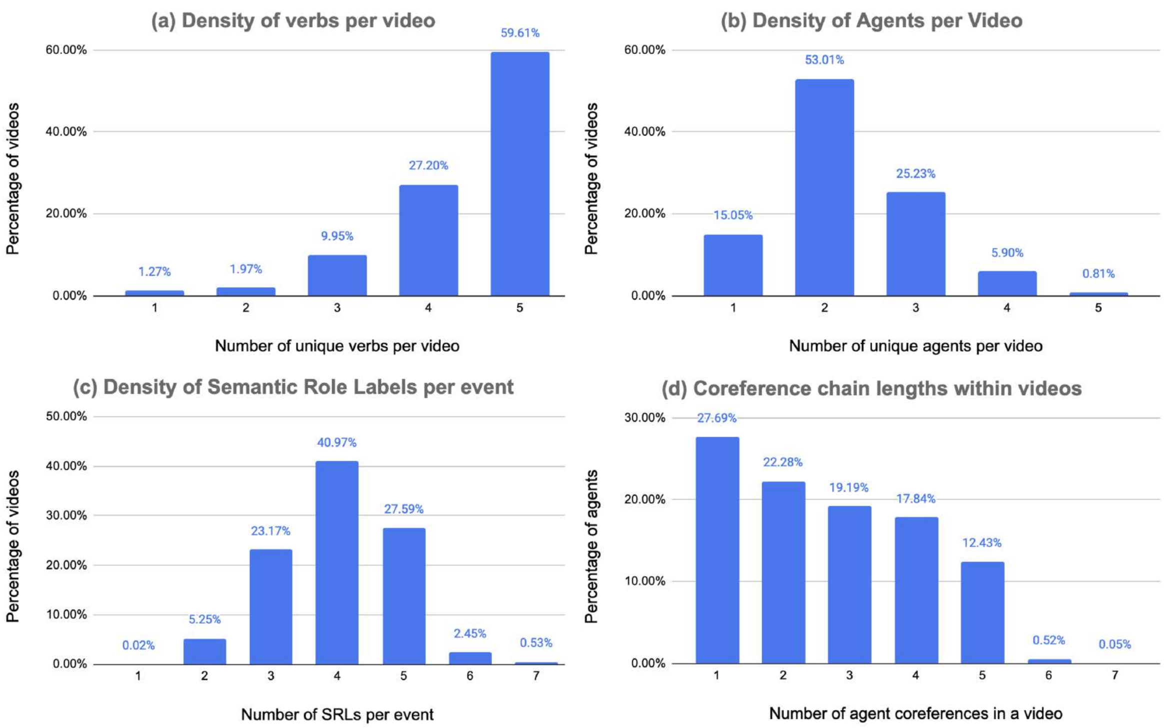


Figure 3.7. Statistics of various features of the VELOCITI benchmark. (a) and (b) show the distribution of verbs and agents per video, respectively. (c) shows the density of SRL annotations per event; and (d) shows the distribution of agent coreference lengths. Even with short videos, the complexity of the VidSitu annotations make the task challenging.

Instructions
These instruction can be opened anytime by clicking 'i' on the bottom left of the panel
Your objective is to mark whether the provided positive caption is an "*accurate"* description of the dictionary contents.

What does an "*accurate",* positive-caption mean?

- The generated caption must include all ideas *inferred* from the dictionary, even though it may miss some exact phrases.
- Ideally, the caption should include everything from the dictionary, but if the caption misses some value of an argument (for example, *direction*), then the caption is correct only if the missing value is *implied* from the caption.
- It needs to be grammatically correct, even though it may sound uncommon in conversational English.

Positive, Neutral, Negative
- Select *positive*, when the caption clearly meets the above requirements.
- Select *neutral*, when the caption partially meets the above requirement (not fully correct).
- Select *negative*, when the caption does NOT meet the above requirements.

Other Rules
- You are only required to look at the provided dictionary, and not the video for this task.
- Captions should NOT be marked incorrect because of noisy annotations in the dictionary.
- Captions should NOT be marked incorrect because of abrupt capitalization inside the sentence.

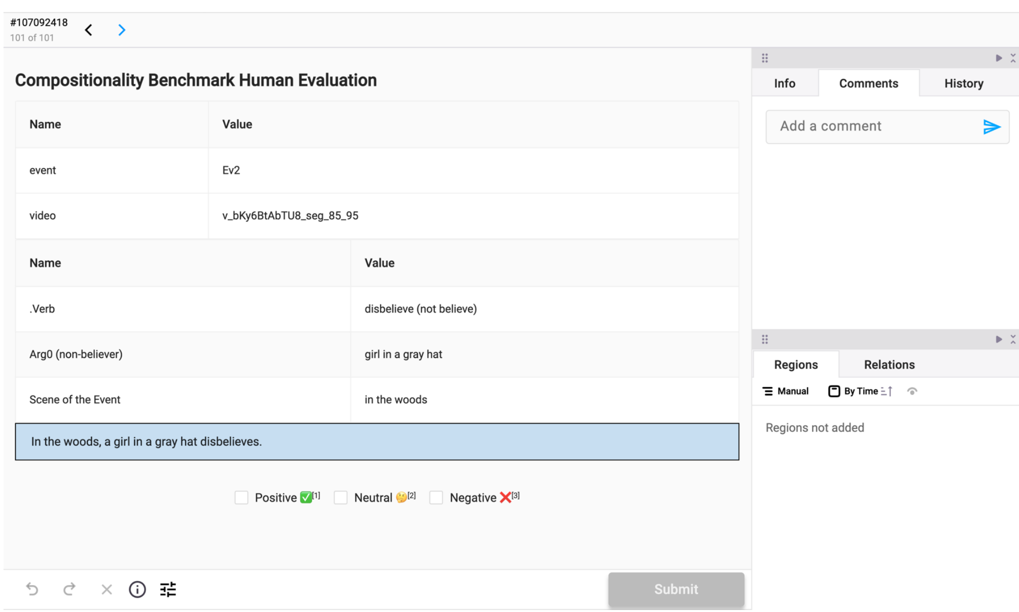


Figure 3.8. Instructions and interface to verify the quality of positive captions generated from LLaMA3 70B.

Instructions

These instructions can be opened anytime by clicking 'i' on the bottom left of the panel

You are given 1 video and 2 captions for each task, one correct caption and a negative caption.

Please watch the video and *verify* if the positive and the negative captions are "*logically correct*".

What is a "*logically-correct*", positive caption?

- Caption that provides a *correct* description of the event in the video.
- It should correctly identify the entities (*humans*, *animals*, *objects, etc.*) and the relationships (*action*) between them.
- Spelling/grammatical errors, if any, shall be ignored.

What is "*logically-correct*", negative caption?

- Caption that provides an *incorrect* description of events in the video.

Note

- You may watch the video multiple times, if required.
- Careful and precise judgement is requirement, as point-of-difference between the positive and the negative caption, may be subtle.

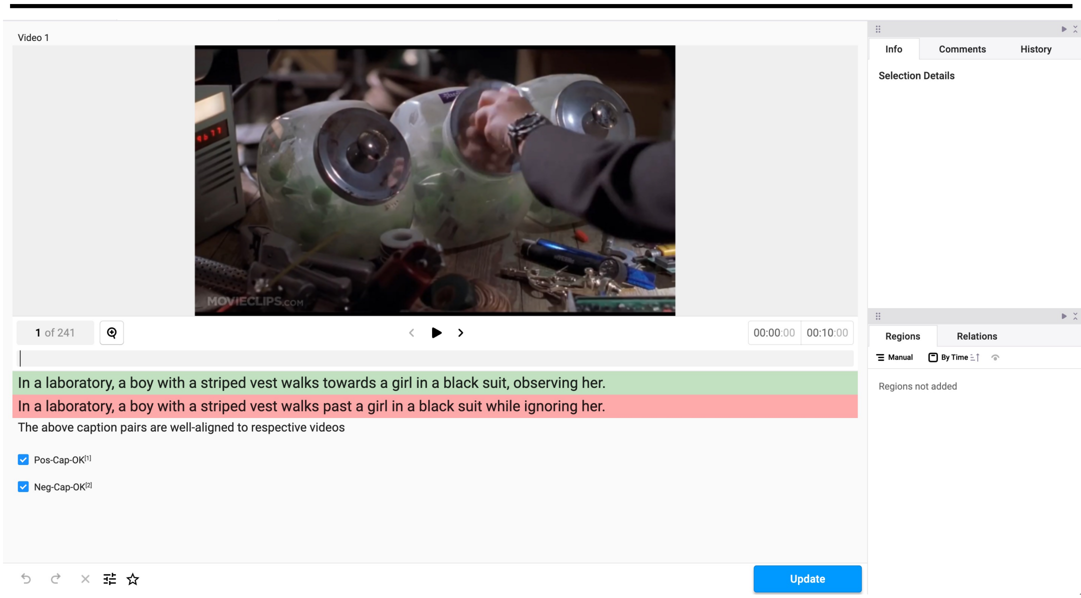


Figure 3.9. Data cleaning instruction for all the tests.

Chapter 4

Investigating Mechanisms for In-Context Vision Language Binding

As Vision-Language models (VLMs) like Gemini [71] and GPT-4o [72] become ubiquitous, it is crucial to understand how they function to determine why they respond the way they do, especially in safety-critical applications. A fundamental ability of VLMs is to associate information across an image and text to reason about a query. For example, given an *image of a furniture store that has a chair with a yellow tag* and the caption *All furniture with a yellow tag have a 30% discount*, a VLM should be able to infer that the chair has a discounted selling price. Our goal is to study this ability to *bind* objects in an image to information in text.

To this end, we propose the *Shapes* task, a controlled synthetic task that requires models to associate 3D objects in an image with their references in the text. In Fig. 4.1, the image contains two 3D objects: a *green sphere* and a *red cube*. The green sphere is referred to as the *green object* in the context. So, to answer the question *'What does the sphere contain?'*, the model needs to internally learn that the sphere corresponds to the phrase green object: *is('green sphere patches', 'green object')*, and that this green object contains item P: *contains('green object', 'item P')*.

The binding ID mechanism proposed in [21] suggests that LLMs' internal activations represent binding information by attaching binding ID vectors to the corresponding entities and attributes. We investigate whether VLMs use a similar mechanism to represent associations between image tokens and text tokens. We study the most commonly used VLM architecture that consists of a visual encoder, a multi-modal projector and a language model.

VLMs and LLMs have some key differences that necessitate careful experimentation. (i) Text tokens have fixed embeddings, while concepts in an image (objects, colors, textures, *etc.*) do not have fixed embeddings; they are represented in the patch tokens obtained from the vision encoder. (ii) Recent powerful VLMs like LLaVA-OneVision [73], Molmo [74], and Qwen2-VL [75] utilize an image encoder

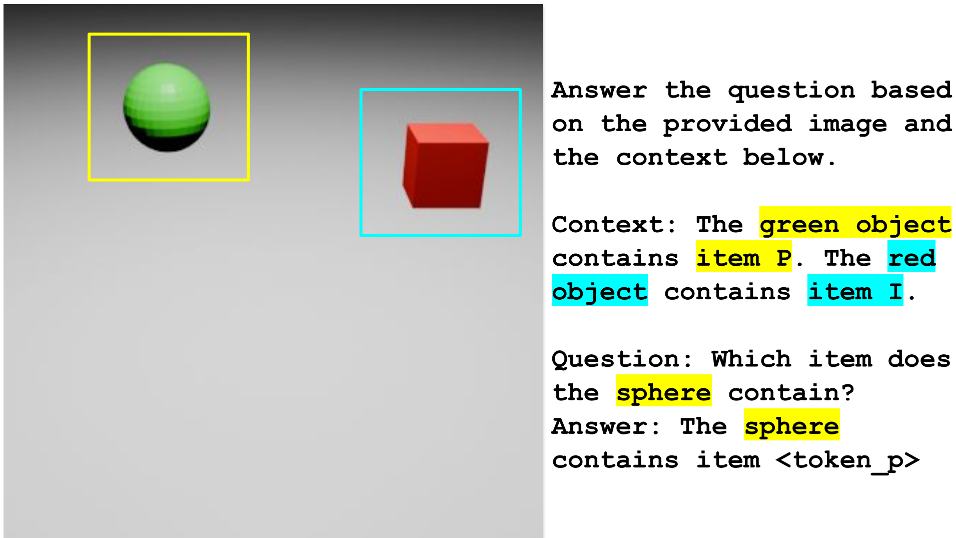


Figure 4.1. **Shapes Task.** Given an image with two 3D objects and a text description (context), the model needs to comprehend the question and identify the correct item (token_p) contained in the queried object. Image and text tokens highlighted with the same color are expected to contain the same binding IDs, allowing the model to predict the correct answer.

that converts the input image into a set of multiscale, multi-crop images and independently maps each of these images into a set of vision tokens. This leads to multiple sets of tokens for the same visual concept. We adapt the causal mediation based experiments from [21] to account for these differences and make the following observations: (i) Image tokens corresponding to the location of the visual concept represent information related to that concept. This is applicable even when there are multiple tokens corresponding to multiple crops from the same image. (ii) VLMs implement the binding ID mechanism. There are binding ID vectors that associate the image tokens corresponding to a visual object and its references in the text tokens.

4.1 Task Definition and Notations

Shapes task. This task consists of images with two 3D objects (O_0, O_1) with distinct shapes and colors. The context refers to both objects using their color (C_0, C_1) and assigns a unique item (I_0, I_1). We use the notation $c = ctxt(O_0 \leftrightarrow C_0 \leftrightarrow I_0, O_1 \leftrightarrow C_1 \leftrightarrow I_1)$ to denote a context where object O_0 of the color C_0 contains item I_0 and object O_1 of the color C_1 contains item I_1. In Fig. 4.1, O_0 and O_1 correspond to the *green sphere* and *red cube* patches in the image, C_0 and C_1 correspond to *green* and *red* in the text and I_0 and I_1 correspond to *item P* and *item I* in text respectively. The question refers to one of the objects using its shape and queries the item assigned to it. Note that '*item P/I*' are randomly chosen uppercase English letters with no inherent meaning.

We generate the images using Blender [76]. We consider four choices for the shape (cone, cube, cylinder and sphere) and six choices for the color (red, blue, green, yellow, cyan and purple). The objects occupy a fixed number of patches and are located in fixed positions.

Notation. Let $\Phi_v(\cdot)$ denote the vision encoder and $g(\cdot)$ denote the multi-modal projector. For an image X_v, the patch embeddings are $t_v = g(\Phi_v(X_v))$. Now, let t_c denote the prompt tokens, comprising image tokens t_v and text tokens up to the context's end, just before the question. Let the LLM have L transformer layers and D-dimensional activation space. For every token position p, $Z_p \in R^{L\times D}$ is the stacked set of residual stream activations. The activations at the object, color, and item positions are denoted as Z_{O_k}, Z_{C_k} and Z_{I_k} respectively where $k \in \{0, 1\}$.

4.2 Do Binding IDs Occur in VLMs?

Binding Id Mechanism. [21] suggests that LLMs associate concepts through binding ID vectors in their activations. Specifically, the activations of an LLM can be decomposed into vectors that encode the concept and those that encode the binding information. Each binding ID consists of similar vector pairs in a subspace, with associated concepts sharing one vector from the same ID.

Extending this, we describe our hypothesis for the existence of binding IDs in VLMs using the Shapes task below:

- Consider 3D objects as visual entities and their colors and items mentioned in the text as their attributes. For the k^{th} visual entity-attributes tuple (I_k, C_k, O_k), the model represents binding vectors in its activations in an abstract form, independent of any particular object, color, or item.

- For object patch tokens, the activations Z_{O_k} can be decomposed as $Z_{O_k} = f_O(O_k) + b_O(k)$. Similarly, $Z_{C_k} = f_C(C_k) + b_C(k)$ and $Z_{I_k} = f_I(I_k) + b_I(k)$. Here $f_O(O_k)$, $f_C(C_k)$, $f_I(I_k)$ are the content vectors and the set of binding vectors $(b_O(k), b_C(k), b_I(k))$ form the binding ID for the k^{th} tuple.

- To answer the question about an object, the model selects the item that shares the same binding ID.

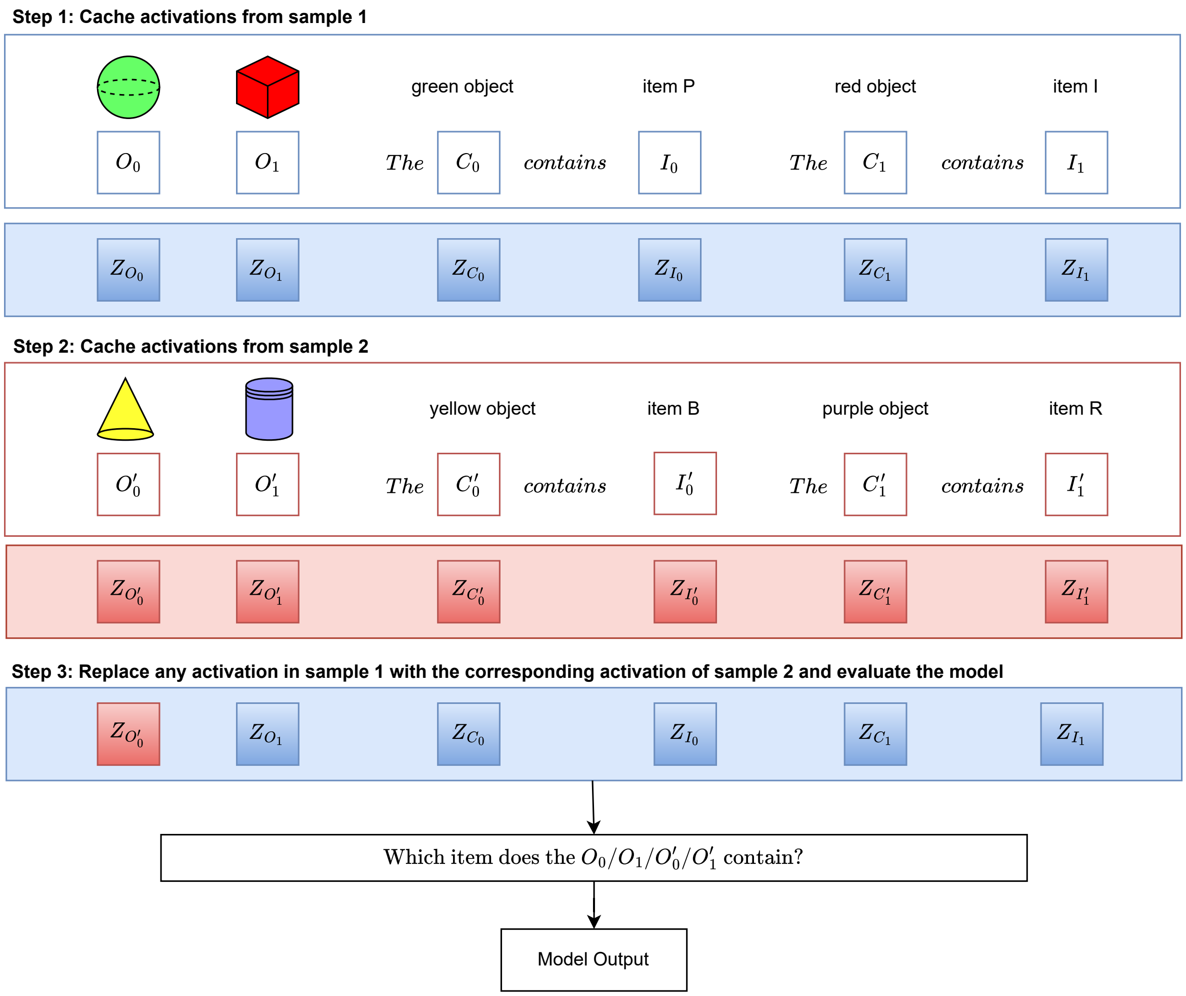
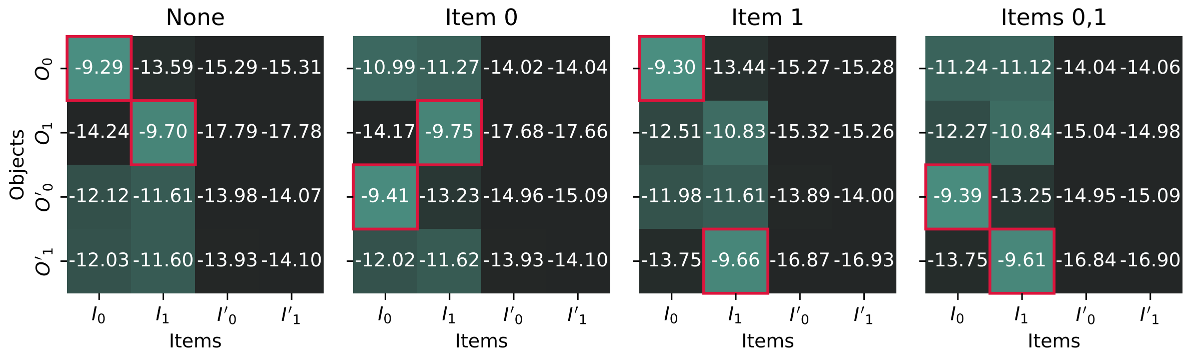
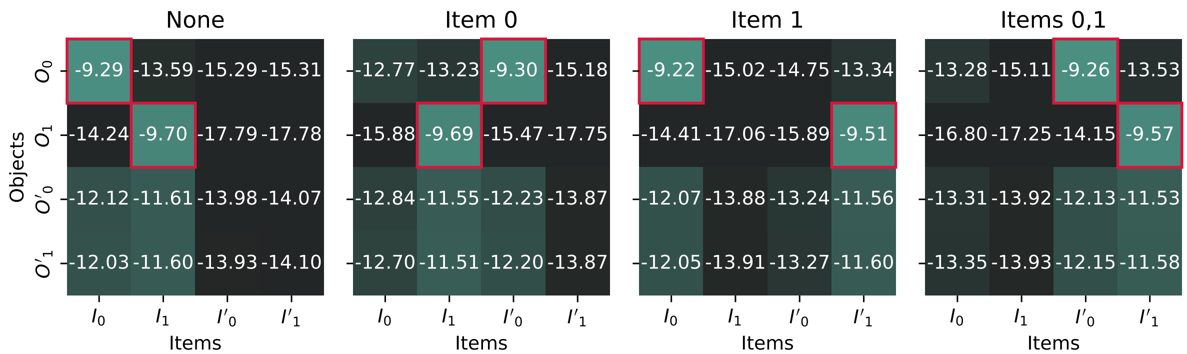


Figure 4.2. Causal intervention. In steps 1 and 2, activations from the first and second samples are saved. In step 3, object/color/item activations in the first sample are replaced with those from the second. This new activation stack is frozen, and the model is queried with all four objects to observe the change in predictions.

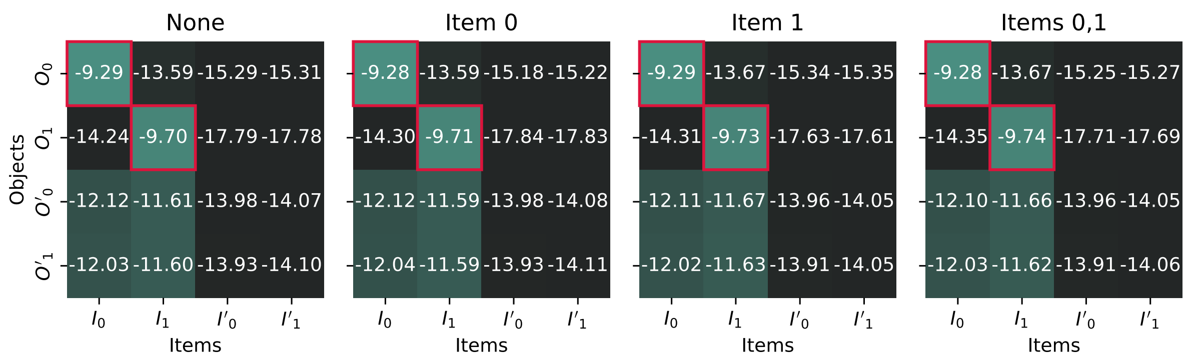
Note that, since binding IDs are independent of the entity/attribute, we can manipulate the associations built by the model by exchanging the binding IDs in the activations as $\hat{Z}_{O_k} := Z_{O_k} - b_O(k) + b_O(k')$ where $k \neq k'$. In the following sections, we assert the existence of the binding ID mechanism by establishing two of its properties: Factorizability (Section 4.2.1) and Position independence (Section 4.2.2). Then, we exchange the associations built by the model using Mean interventions (Section 4.2.3).



(a) Object activation replacements



(b) Item activation replacements



(c) Color activation replacements

Figure 4.3. **Factorizability results.** Each row shows the model's mean log probabilities of an item contained in an object. The first grid in each case shows results with unaltered activations. Squares highlighted in red denote the expected predictions based on our hypothesis. Model outputs match hypothesis suggesting a multimodal binding ID mechanism.

4.2.1 Factorizability

Fig. 4.2 shows two samples from the Shapes task with the contexts $c = ctxt(O_0 \leftrightarrow C_0 \leftrightarrow I_0, O_1 \leftrightarrow C_1 \leftrightarrow I_1)$ and $c' = ctxt(O'_0 \leftrightarrow C'_0 \leftrightarrow I'_0, O'_1 \leftrightarrow C'_1 \leftrightarrow I'_1)$.

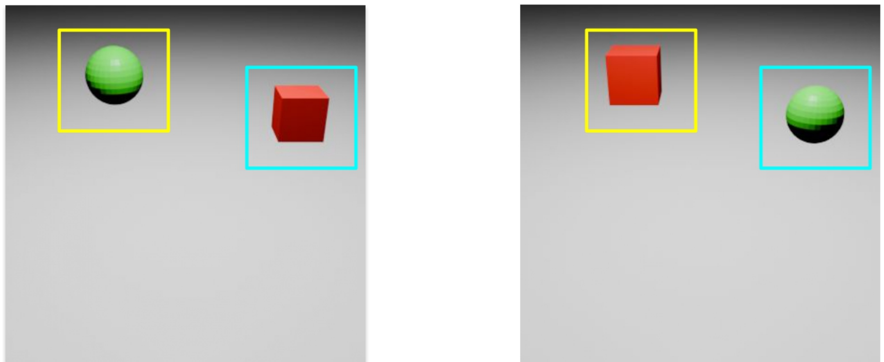
The Binding ID mechanism assumes that the information linking a concept to its attributes is stored locally within the activations at its token positions and is independent of the specific concept itself. This implies that the activations of the *sphere* (O_0) in the first sample and the *cone* (O'_0) in the second sample should contain the same binding vector $b_O(0)$ as they both correspond to the 0^{th} visual entity-attributes tuple in their respective samples. Replacing Z_{O_0} with $Z_{O'_0}$ should now bind the *cone* with the text tokens *green object* and *item P*. We demonstrate this using causal interventions [77] on the activations as described below.

- Cache all activations Z_c from the model run on c.
- Cache activations $Z_{O'_0}$ and $Z_{O'_1}$ from the model run on c'.
- Construct a new stack of activations Z_c^* by modifying Z_c such that Z_{O_k} is replaced with $Z_{O'_k}$ for any $k \in \{0, 1\}$.
- Re-evaluate the model by probing what item each shape (O_0, O_1, O'_0, O'_1) contains by freezing the activation cache as Z_c^*. We expect the model to now associate O'_k with I_k since both Z_{O_k} and $Z_{O'_k}$ contain the same binding ID vector $b_O(k)$.

Results.

Fig. 4.3 shows the mean log probability of choosing an item before and after interventions. We show the factorizability results for object patch tokens, color tokens and item tokens. In Fig. 4.3a, the **first** grid shows the results when the *activations are unaltered*. As expected, for objects O_0 and O_1, items I_0 and I_1 are chosen at a higher rate, respectively and for objects O'_0 and O'_1, items I_0 and I_1 are chosen at a roughly equal rate since these objects do not exist in the image. In the **second** grid, we replace Z_{O_0} with $Z_{O'_0}$. Now, when the model is queried for the item contained by O'_0, the model picks item I_0 over I_1. The **third** grid follows the same pattern, Z_{O_1} is replaced by $Z_{O'_1}$ resulting in O'_1 containing I_1. Finally, both object activations are replaced in the **fourth** grid and we observe that the model chooses $I_{0/1}$ for $O'_{0/1}$ respectively. Note that when the object patches are replaced, the color of the new object no longer matches the color description in the text. Nevertheless, the new object is still associated with the same item as the original object, as they both contain the same binding vector.

We observe a similar behavior for replacing items in Fig. 4.3b. When Z_{I_k} is replaced by $Z_{I'_k}$, the model prefers item I'_k for object O_k. However, when we intervene on the color activations Z_{C_k}, the



`Context: The green object contains item P. The red object contains item I.`

`Context: The green object contains item P. The red object contains item I.`

Figure 4.4. Mean intervention samples.

results are similar to when there are no interventions (Fig. 4.3c). This is expected since both Z_{C_k} and $Z_{C'_k}$ contain the same binding ID vectors.

4.2.2 Position Independence

Next, we hypothesize that the associations formed by the model are invariant to the activation positions of the object, color, or item, as they rely solely on the binding IDs. This implies that swapping the positions of Z_{O_0} and Z_{O_1} should not change items associated with the objects. To validate this, we first obtain the activations of the context tokens Z_c (Section 4.2.1). Then, we compute a new stack of activations Z_c^* wherein the positions of Z_{O_0} and Z_{O_1} are altered, following the procedure described in [21], adapted for models that use Rotary Position Embedding (RoPE) [78]. Unlike absolute position embeddings, RoPE incorporate positional information only through the attention score computations, without injecting it directly into the residual stream activations.

Results. Fig. 4.5 shows the mean log probabilities when the positions of Z_{O_0} and Z_{O_1} are progressively adjusted to get closer and ultimately swapped. We observe that the model answers with the correct item regardless of positions.

4.2.3 Mean Interventions

The factorizability and position independence results show that binding vectors are contained within the activations corresponding to the object, color, and item tokens and cause the model to form associations across image and text. If binding vectors were directly accessible, we could interchange them to observe if the model changes its answer. While this is not feasible, we can approximate the difference in binding vectors from the difference in activations. To estimate $\Delta_O = b_O(1) - b_O(0)$, we consider two instances of the Shapes task as shown in Fig. 4.4. Let O_0, O_1 denote the objects in the first instance and O'_0, O'_1 denote the objects in the second instance. Notice that both O_0 and O'_1 are the same object, a *green sphere*. However, we expect their activations to contain different binding IDs. We can now estimate Δ_O as the difference $Z_{O'_1} - Z_{O_0}$. Concretely, we compute Δ_O as the mean of the difference of activations over multiple pairs of instances ($\Delta_O \approx mean_{O_0,O'_1}[Z_{O'_1} - Z_{O_0}]$). Similarly, we compute $\Delta_C = b_C(1) - b_C(0)$ and $\Delta_I = b_I(1) - b_I(0)$ from the color and item activations.

Using these mean vectors (Δ_O, Δ_C, Δ_I), we can now edit the binding vectors in the activations to alter the model response. For any new instance with the context $c^* = ctxt(O^*_0 \leftrightarrow C^*_0 \leftrightarrow I^*_0, O^*_1 \leftrightarrow C^*_1 \leftrightarrow I^*_1)$, we can alter the binding vector of the objects as $Z_{O^*_0} := Z_{O^*_0} + \Delta_O$ and $Z_{O^*_1} := Z_{O^*_1} - \Delta_O$. This should result in a swap of object-item binding with O^*_0 and O^*_1 being bound to I^*_1 and I^*_0 respectively. Similarly, altering the binding vector of the items as $Z_{I^*_0} := Z_{I^*_0} + \Delta_I$ and $Z_{I^*_1} := Z_{I^*_1} - \Delta_I$ should also exchange the model response. Altering the binding vectors in color token activations will make the model now associate O_k, $C_{k'}$ and I_k where $k \neq k'$. However, O_k is still bound to I_k, and we expect no change in response.

Results. Table 4.1 shows the accuracy, measured as the fraction of samples where the correct item has the highest log probability among the possible items in the context. As expected, both object and item interventions individually change the model's response, while color interventions do not. Further, simultaneously performing object and item interventions restores the model's original response since they now have the same binding IDs. We also repeat these experiments with random vectors that have the same magnitude but different directions. These vectors do not alter the model response, indicating that the specific directions of the mean vectors causally affect the binding.

4.2.4 Experimental Details

Throughout the paper, we report results with LLaVA-OneVision-7B [73], which uses the SigLIP [79] vision encoder and encodes multiple crops from a single image. The Shapes task images are of size 384×384, with each object appearing in two crops and occupying 5×5 patch tokens. Empirically, we found that when intervening on object token activations, a 3-token padding on all sides in both crops yields optimal results. To estimate the difference of binding vectors, we use a separate set that contains

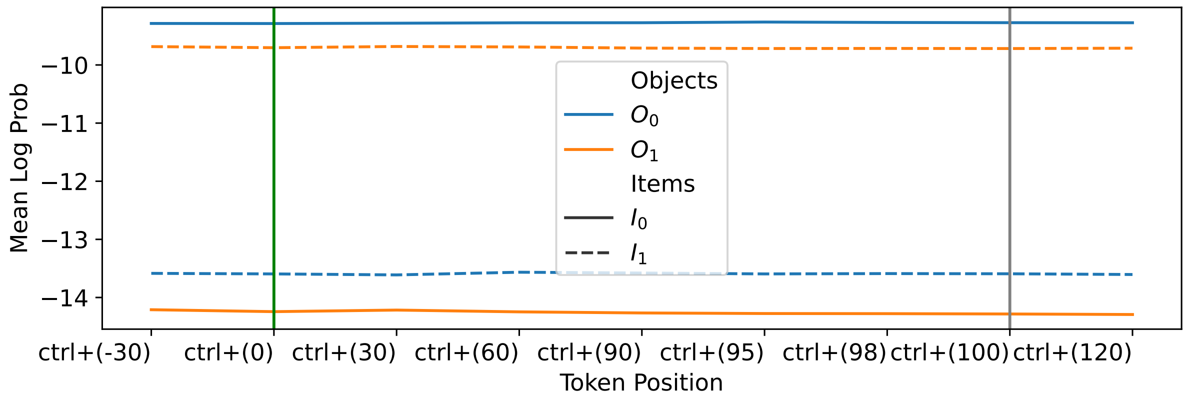
Condition	Mean vectors		Random vectors	
	$O_0 \leftrightarrow I_0$	$O_1 \leftrightarrow I_1$	$O_0 \leftrightarrow I_0$	$O_1 \leftrightarrow I_1$
None	1.00	1.00	-	-
O	0.00	0.05	1.00	1.00
I	0.05	0.00	1.00	1.00
C	1.00	1.00	1.00	1.00
O, I	1.00	0.95	1.00	1.00
O, I, C	1.00	0.95	1.00	1.00

Table 4.1. Mean ablation accuracies: Object (O), Item (I), Color (C).

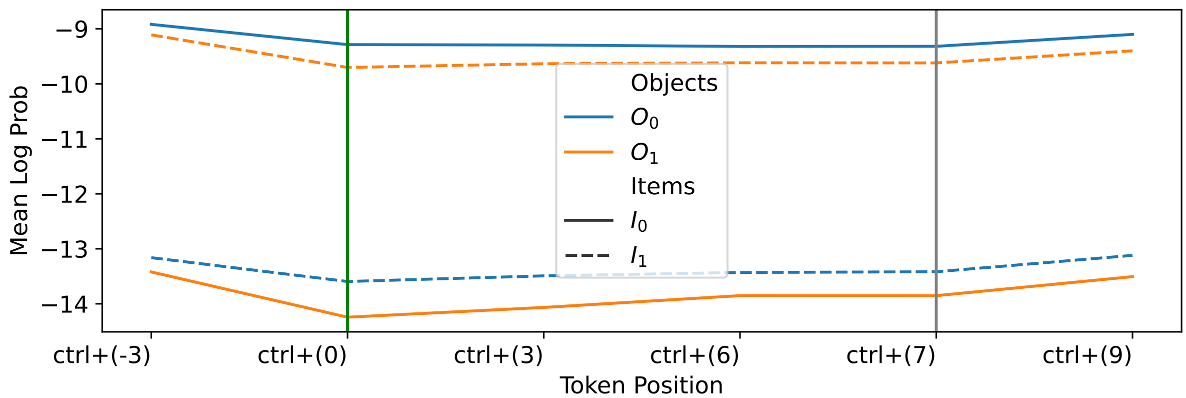
different shapes (frustum, pyramid, prism and toroid), colors (lime, pink, gold, brown, orange and azure) and items (lowercase English alphabet). All colors and items span two text tokens.

4.3 Conclusion

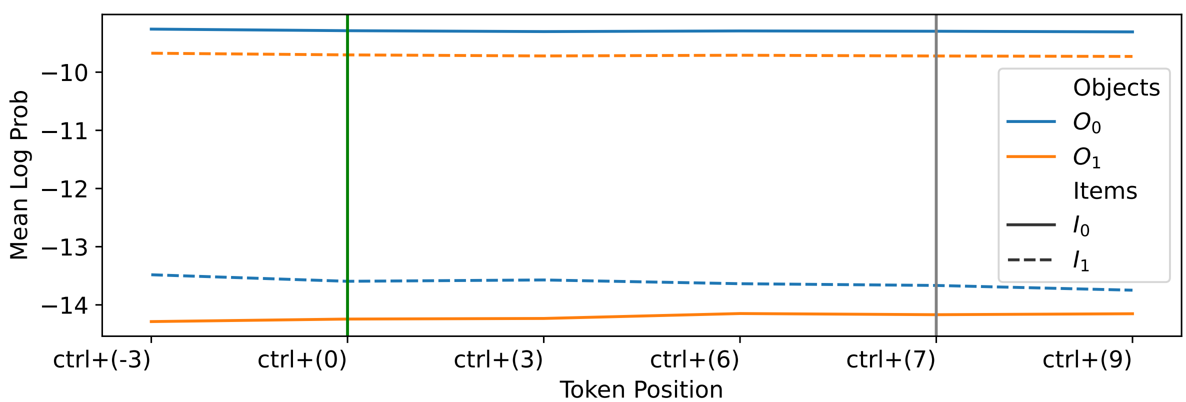
In this work, we explore how in-context associations occur in VLMs. We formulate the Shapes task, a simple and controlled QA task which requires the model to associate 3D objects in an image with their references in the text. Through experiments, we demonstrate that VLMs utilize binding ID vectors to bind concepts across image and text.



(a) Object



(b) Item



(c) Color

Figure 4.5. Position independence results. The integers in the x-axis show how much the position of the first and second objects/items/colors are incremented and decremented respectively. The green line corresponds to no change in positions and the gray line corresponds to swapped positions. In all cases $O_k \leftrightarrow I_k$ (blue solid O_0, I_0 and oranged dashed O_1, I_1) have a higher probability than $O_k \leftrightarrow I'_k$.

Chapter 5

Conclusion

Compositional reasoning, the ability to generalize to novel scenarios by flexibly recombining known concepts lies at the heart of human intelligence. This thesis presents novel insights into the compositional reasoning capabilities of Vision-Language models.

In the first part of the thesis, we introduced VELOCITI, a diagnostic benchmark designed to rigorously evaluate compositionality in Video-Language Models. Unlike prior video benchmarks focused on single-agent scenarios, VELOCITI presents complex, real-world videos with multiple interacting agents and events. Our experiments revealed that, while VLMs show promising zero-shot capabilities, they often struggle with fine-grained compositional distinctions, particularly in associating the correct agents with actions over time. The significant performance gap between text-inspired and in-video negation tests highlights a key challenge: the weakness of current models to reliably bind visual concepts.

In the second part of the thesis, we explored the internal mechanisms that allow VLMs to associate visual and textual concepts. Through probing activations, we found evidence supporting the existence of Binding IDs vectors that facilitate the alignment between visual tokens and their textual descriptions.

Together, these contributions advance our understanding of how VLMs perceive, represent, and reason about multi-modal information. As VLMs continue to evolve, improving their ability to represent and bind concepts compositionally will be essential for achieving robust, generalizable intelligence across tasks and domains.

Future work should further explore architectural and training strategies that explicitly encourage compositional representations, as well as develop more nuanced benchmarks that reflect the complexity of real-world understanding. In doing so, we move closer to building models that reason with the flexibility and depth of human cognition.

Publications

- **Darshana Saravanan***, Varun Gupta*, Darshan Singh*, Zeeshan Khan, Vineet Gandhi, and Makarand Tapaswi, "VELOCITI: Benchmarking Video-Language Compositional Reasoning with Strict Entailment," in Computer Vision and Pattern Recognition (CVPR), 2025.
- **Darshana Saravanan**, Makarand Tapaswi, and Vineet Gandhi, "Investigating mechanisms for in-context vision language binding," in Mechanistic Interpretability for Vision at CVPR 2025 (Proceedings Track), 2025.
- **Darshana Saravanan**, Naresh Manwani, and Vineet Gandhi, "Pseudo-labelling meets Label Smoothing for Noisy Partial Label Learning," in The 12th Workshop on Fine-Grained Visual Categorization at CVPR 2025 (Proceedings Track), 2025.

Bibliography

[1] A. Sadhu, T. Gupta, M. Yatskar, R. Nevatia, and A. Kembhavi, "Visual Semantic Role Labeling for Video Understanding," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[2] M. Grunde-McLaughlin, R. Krishna, and M. Agrawala, "AGQA: A Benchmark for Compositional Spatio-Temporal Reasoning," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[3] B. Bogin, S. Gupta, M. Gardner, and J. Berant, "Covr: A test-bed for visually grounded compositional generalization with real images," *arXiv preprint arXiv:2109.10613*, 2021.

[4] D. Bahdanau, H. de Vries, T. J. O'Donnell, S. Murty, P. Beaudoin, Y. Bengio, and A. Courville, "Closure: Assessing systematic generalization of clevr models," *arXiv preprint arXiv:1912.05783*, 2019.

[5] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein, "Neural module networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 39–48.

[6] I. Higgins, N. Sonnerat, L. Matthey, A. Pal, C. P. Burgess, M. Bosnjak, M. Shanahan, M. Botvinick, D. Hassabis, and A. Lerchner, "Scan: Learning hierarchical compositional visual concepts," *arXiv preprint arXiv:1707.03389*, 2017.

[7] D. A. Hudson and C. D. Manning, "Compositional attention networks for machine reasoning," *arXiv preprint arXiv:1803.03067*, 2018.

[8] T. M. Le, V. Le, S. Venkatesh, and T. Tran, "Hierarchical conditional relation networks for video question answering," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9972–9981.

[9] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning Transferable Visual Models From Natural Language Supervision," in *International Conference on Machine Learning (ICML)*, 2021.

[10] Y. Tang, Y. Yamada, Y. Zhang, and I. Yildirim, "When are lemons purple? the concept association bias of vision-language models," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 14 333–14 348.

[11] M. Lewis, N. Nayak, P. Yu, J. Merullo, Q. Yu, S. Bach, and E. Pavlick, "Does CLIP Bind Concepts? Probing Compositionality in Large Image Models," in *European Chapter of the Association for Computational Linguistics (EACL)*, 2024.

[12] T. Thrush, R. Jiang, M. Bartolo, A. Singh, A. Williams, D. Kiela, and C. Ross, "Winoground: Probing vision and language models for visio-linguistic compositionality," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5238–5248.

[13] M. Yuksekgonul, F. Bianchi, P. Kalluri, D. Jurafsky, and J. Zou, "When and why vision-language models behave like bags-of-words, and what to do about it?" in *The Eleventh International Conference on Learning Representations*, 2023.

[14] C.-Y. Hsieh, J. Zhang, Z. Ma, A. Kembhavi, and R. Krishna, "Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality," *Advances in neural information processing systems*, vol. 36, pp. 31 096–31 116, 2023.

[15] Y. Liu, S. Li, Y. Liu, Y. Wang, S. Ren, L. Li, S. Chen, X. Sun, and L. Hou, "TempCompass: Do Video LLMs Really Understand Videos?" in *Findings of the Association for Computational Linguistics (ACL)*, 2024.

[16] J. Zhang, C. Mu, and Y. J. Lee, "Vinoground: Scrutinizing LMMs over Dense Temporal Reasoning with Short Videos," *arXiv*, 2024. [Online]. Available: https://arxiv.org/abs/2410.02763

[17] A. Sadhu, T. Gupta, M. Yatskar, R. Nevatia, and A. Kembhavi, "Visual semantic role labeling for video understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5589–5600.

[18] A. Talmor, O. Yoran, A. Catav, D. Lahav, Y. Wang, A. Asai, G. Ilharco, H. Hajishirzi, and J. Berant, "Multimodal{qa}: complex question answering over text, tables and images," in *International Conference on Learning Representations*, 2021.

[19] R. G. Reddy *et al.*, "MuMuQA: Multimedia Multi-Hop News Question Answering via Cross-Media Knowledge Extraction and Grounding," in *Association for the Advancement of Artificial Intelligence (AAAI)*, 2022.

[20] K. Chen and X. Wu, "VTQA: Visual Text Question Answering via Entity Alignment and Cross-Media Reasoning," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[21] J. Feng and J. Steinhardt, "How do Language Models Bind Entities in Context?" in *International Conference on Learning Representations (ICLR)*, 2024.

[22] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International journal of computer vision*, vol. 123, pp. 32–73, 2017.

[23] D. A. Hudson and C. D. Manning, "Gqa: A new dataset for real-world visual reasoning and compositional question answering," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 6700–6709.

[24] J. Johnson, B. Hariharan, L. Van Der Maaten, L. Fei-Fei, C. Lawrence Zitnick, and R. Girshick, "Clevr: A diagnostic dataset for compositional language and elementary visual reasoning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2901–2910.

[25] Z. Ma, J. Hong, M. O. Gul, M. Gandhi, I. Gao, and R. Krishna, "Crepe: Can vision-language foundation models reason compositionally?" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 10 910–10 921.

[26] T. Zhao, T. Zhang, M. Zhu, H. Shen, K. Lee, X. Lu, and J. Yin, "Vl-checklist: Evaluating pre-trained vision-language models with objects, attributes and relations," *arXiv preprint arXiv:2207.00221*, 2022.

[27] S. Pratt, M. Yatskar, L. Weihs, A. Farhadi, and A. Kembhavi, "Grounded situation recognition," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*. Springer, 2020, pp. 314–332.

[28] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*. Springer, 2014, pp. 740–755.

[29] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2641–2649.

[30] A. Ray, F. Radenovic, A. Dubey, B. Plummer, R. Krishna, and K. Saenko, "Cola: A benchmark for compositional text-to-image retrieval," *Advances in Neural Information Processing Systems*, vol. 36, pp. 46 433–46 445, 2023.

[31] J. Ji, R. Krishna, L. Fei-Fei, and J. C. Niebles, "Action genome: Actions as compositions of spatio-temporal scene graphs," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 236–10 247.

[32] M. Gandhi, M. O. Gul, E. Prakash, M. Grunde-McLaughlin, R. Krishna, and M. Agrawala, "Measuring compositional consistency for video question answering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5046–5055.

[33] R. Girdhar and D. Ramanan, "CATER: A diagnostic dataset for Compositional Actions and TEmporal Reasoning," *arXiv preprint arXiv:1910.04744*, 2019.

[34] J. S. Park, S. Shen, A. Farhadi, T. Darrell, Y. Choi, and A. Rohrbach, “Exposing the Limits of Video-Text Models through Contrast Sets,” in *North American Chapter of Association of Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2022.

[35] H. Bansal, Y. Bitton, I. Szpektor, K.-W. Chang, and A. Grover, “VideoCon: Robust Video-Language Alignment via Contrast Captions,” in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[36] S. Li, L. Li, S. Ren, Y. Liu, Y. Liu, R. Gao, X. Sun, and L. Hou, “VITATECS: A Diagnostic Dataset for Temporal Concept Understanding of Video-Language Models,” in *European Conference on Computer Vision Workshops (ECCVW)*, 2024.

[37] S. Doveh, A. Arbelle, S. Harary, E. Schwartz, R. Herzig, R. Giryes, R. Feris, R. Panda, S. Ullman, and L. Karlinsky, “Teaching structured vision & language concepts to vision & language models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2657–2668.

[38] J. Singh, I. Shrivastava, M. Vatsa, R. Singh, and A. Bharati, “Learn” no” to say” yes” better: Improving vision-language models via negations,” *arXiv preprint arXiv:2403.20312*, 2024.

[39] S. Castro, A. Ziai, A. Saluja, Z. Yuan, and R. Mihalcea, “Clove: Encoding compositional language in contrastive vision-language models,” *arXiv preprint arXiv:2402.15021*, 2024.

[40] H. Singh, P. Zhang, Q. Wang, M. Wang, W. Xiong, J. Du, and Y. Chen, “Coarse-to-fine contrastive learning in image-text-graph space for improved vision-language compositionality,” *arXiv preprint arXiv:2305.13812*, 2023.

[41] D. Kim, A. Piergiovanni, G. Mallya, and A. Angelova, “Videocomp: Advancing fine-grained compositional and temporal alignment in video-text models,” *arXiv preprint arXiv:2504.03970*, 2025.

[42] S. Doveh, A. Arbelle, S. Harary, R. Herzig, D. Kim, P. Cascante-Bonilla, A. Alfassy, R. Panda, R. Giryes, R. Feris, *et al.*, “Dense and aligned captions (dac) promote compositional reasoning in vl models,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 76 137–76 150, 2023.

[43] L. Zhang, R. Awal, and A. Agrawal, “Contrasting intra-modal and ranking cross-modal hard negatives to enhance visio-linguistic compositional understanding,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 13 774–13 784.

[44] Y. Oh, J. W. Cho, D.-J. Kim, I. S. Kweon, and J. Kim, “Preserving multi-modal capabilities of pre-trained vlms for improving vision-linguistic compositionality,” *arXiv preprint arXiv:2410.05210*, 2024.

[45] C. Zheng, J. Zhang, A. Kembhavi, and R. Krishna, “Iterated learning improves compositionality in large vision-language models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 13 785–13 795.

[46] C. Lai, S. Song, S. Yan, and G. Hu, "Improving vision and language concepts understanding with multimodal counterfactual samples," in *European Conference on Computer Vision*. Springer, 2024, pp. 174–191.

[47] W. Peng, S. Xie, Z. You, S. Lan, and Z. Wu, "Synthesize diagnose and optimize: Towards fine-grained vision-language understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 13 279–13 288.

[48] J. Zhang, M. Cai, T. Xie, and Y. J. Lee, "Countercurate: Enhancing physical and semantic visio-linguistic compositional reasoning via counterfactual examples," *arXiv preprint arXiv:2402.13254*, 2024.

[49] U. Sahin, H. Li, Q. Khan, D. Cremers, and V. Tresp, "Enhancing multimodal compositional reasoning of visual language models with generative negative mining," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 5563–5573.

[50] M. Patel, N. S. A. Kusumba, S. Cheng, C. Kim, T. Gokhale, C. Baral, *et al.*, "Tripletclip: Improving compositional reasoning of clip via synthetic vision-language negatives," *Advances in neural information processing systems*, vol. 37, pp. 32 731–32 760, 2024.

[51] H. Li and B. Li, "Enhancing vision-language compositional understanding with multimodal synthetic data," *arXiv preprint arXiv:2503.01167*, 2025.

[52] W. Dai, J. Li, D. Li, A. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, and S. Hoi, "InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[53] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge, Y. Fan, K. Dang, M. Du, X. Ren, R. Men, D. Liu, C. Zhou, J. Zhou, and J. Lin, "Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution," *arXiv preprint arXiv:2409.12191*, 2024.

[54] B. Li, Y. Zhang, D. Guo, R. Zhang, F. Li, H. Zhang, K. Zhang, Y. Li, Z. Liu, and C. Li, "LLaVA-OneVision: Easy Visual Task Transfer," *arXiv preprint arXiv:2408.03326*, 2024.

[55] M. Deitke, C. Clark, S. Lee, R. Tripathi, Y. Yang, J. S. Park, M. Salehi, N. Muennighoff, K. Lo, L. Soldaini, *et al.*, "Molmo and PixMo: Open weights and open data for state-of-the-art multimodal models," *arXiv preprint arXiv:2409.17146*, 2024.

[56] L. Beyer, A. Steiner, A. S. Pinto, A. Kolesnikov, X. Wang, D. Salz, M. Neumann, I. Alabdulmohsin, M. Tschannen, E. Bugliarello, *et al.*, "PaliGemma: A versatile 3B VLM for transfer," *arXiv preprint arXiv:2407.07726*, 2024.

[57] L. Xu, Y. Zhao, D. Zhou, Z. Lin, S. K. Ng, and J. Feng, "PLLaVA: Parameter-free LLaVA Extension from Images to Videos for Video Dense Captioning," *arXiv preprint arXiv:2404.16994*, 2024.

[58] B. Lin, B. Zhu, Y. Ye, M. Ning, P. Jin, and L. Yuan, "Video-LLaVA: Learning United Visual Representation by Alignment Before Projection," *arXiv preprint arXiv:2311.10122*, 2023.

[59] R. Rawal, K. Saifullah, R. Basri, D. Jacobs, G. Somepalli, and T. Goldstein, "CinePile: A Long Video Question Answering Dataset and Benchmark," *arXiv preprint arXiv:2405.08813*, 2024.

[60] X. Fang, K. Mao, H. Duan, X. Zhao, Y. Li, D. Lin, and K. Chen, "MMBench-Video: A Long-Form Multi-Shot Benchmark for Holistic Video Understanding," in *Advances in Neural Information Processing Systems (NeurIPS): Track on Datasets and Benchmarks*, 2024.

[61] C. Fu, Y. Dai, Y. Luo, L. Li, S. Ren, R. Zhang, Z. Wang, C. Zhou, Y. Shen, M. Zhang, P. Chen, Y. Li, S. Lin, S. Zhao, K. Li, T. Xu, X. Zheng, E. Chen, R. Ji, and X. Sun, "Video-MME: The First-Ever Comprehensive Evaluation of Multi-modal LLMs in Video Analysis," *arXiv preprint*, 2024.

[62] M. Cai, R. Tan, J. Zhang, B. Zou, K. Zhang, F. Yao, F. Zhu, J. Gu, Y. Zhong, Y. Shang, Y. Dou, J. Park, J. Gao, Y. J. Lee, and J. Yang, "TemporalBench: Towards Fine-grained Temporal Understanding for Multimodal Video Models," *arXiv preprint arXiv:2410.10818*, 2024.

[63] M. AI, "Llama3," https://llama.meta.com/llama3/, 2024.

[64] S. Buch, C. Eyzaguirre, A. Gaidon, J. Wu, L. Fei-Fei, and J. C. Niebles, "Revisiting the "Video" in Video-Language Understanding," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[65] D. Jurafsky and J. H. Martin, "Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models," 2024. [Online]. Available: https://web.stanford.edu/~jurafsky/slp3/

[66] P. Bagad, M. Tapaswi, and C. G. Snoek, "Test of Time: Instilling Video-Language Models With a Sense of Time," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[67] Sanyal, Soumya and Xiao, Tianyi and Liu, Jiacheng and Wang, Wenya and Ren, Xiang, "Are Machines Better at Complex Reasoning? Unveiling Human-Machine Inference Gaps in Entailment Verification," in *Findings of the Association for Computational Linguistics (ACL)*, 2024.

[68] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, " Sigmoid Loss for Language Image Pre-Training ," in *International Conference on Computer Vision (ICCV)*, 2023.

[69] H. Rasheed, M. U. Khattak, M. Maaz, S. Khan, and F. S. Khan, "Fine-Tuned CLIP Models Are Efficient Video Learners," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[70] J. Liu, W. Wang, D. Wang, N. A. Smith, Y. Choi, and H. Hajishirzi, "Vera: A General-Purpose Plausibility Estimation Model for Commonsense Statements," in *Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

[71] G. Team, "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context," *arXiv preprint arXiv:2403.05530*, 2024.

[72] A. Hurst *et al.*, "GPT-4o system card," *arXiv preprint arXiv:2410.21276*, 2024.

[73] B. Li, Y. Zhang, D. Guo, R. Zhang, F. Li, H. Zhang, K. Zhang, Y. Li, Z. Liu, and C. Li, "LLaVA-OneVision: Easy Visual Task Transfer," *arXiv preprint arXiv:2408.03326*, 2024.

[74] M. Deitke *et al.*, "Molmo and PixMo: Open Weights and Open Data for State-of-the-Art Vision-Language Models," *arXiv preprint arXiv:2409.17146*, 2024.

[75] P. Wang *et al.*, "Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution," *arXiv preprint arXiv:2409.12191*, 2024.

[76] B. O. Community, *Blender - A 3D Modelling and Rendering Package*, Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. [Online]. Available: http://www.blender.org

[77] J. Vig, S. Gehrmann, Y. Belinkov, S. Qian, D. Nevo, Y. Singer, and S. Shieber, "Investigating Gender Bias in Language Models Using Causal Mediation Analysis," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[78] J. Su, Y. Lu, S. Pan, A. Murtadha, B. Wen, and Y. Liu, "RoFormer: Enhanced Transformer with Rotary Position Embedding," *arXiv preprint arXiv:2104.09864*, 2021.

[79] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid Loss for Language Image Pre-Training," in *International Conference on Computer Vision (ICCV)*, 2023.