VELOCITI: Benchmarking Video-Language Compositional Reasoning with Strict Entailment

Darshana Saravanan[1]☯ Varun Gupta[1]☯ Darshan Singh[1]☯ Zeeshan Khan[2]
Vineet Gandhi[1] Makarand Tapaswi[1]

[1]CVIT, IIIT Hyderabad, India [2]Inria, Paris, France

katha-ai.github.io/projects/velociti

Abstract

A fundamental aspect of compositional reasoning in a video is associating people and their actions across time. Recent years have seen great progress in general-purpose vision/video models and a move towards long-video understanding. While exciting, we take a step back and ask: are today's models good at compositional reasoning on short videos? To this end, we introduce VELOCITI, a benchmark to study Video-LLMs by disentangling and assessing the comprehension of agents, actions, and their associations across multiple events. We adopt the Video-Language Entailment setup and propose StrictVLE that requires correct classification (rather than ranking) of the positive and negative caption. We evaluate several models and observe that even the best, LLaVA-OneVision (44.5%) and Gemini-1.5-Pro (49.3%), are far from human accuracy at 93.0%. Results show that action understanding lags behind agents, and negative captions created using entities appearing in the video perform worse than those obtained from pure text manipulation. We also present challenges with ClassicVLE and multiple-choice (MC) evaluation, strengthening our preference for StrictVLE. Finally, we validate that our benchmark requires visual inputs of multiple frames making it ideal to study video-language compositional reasoning.

1. Introduction

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.

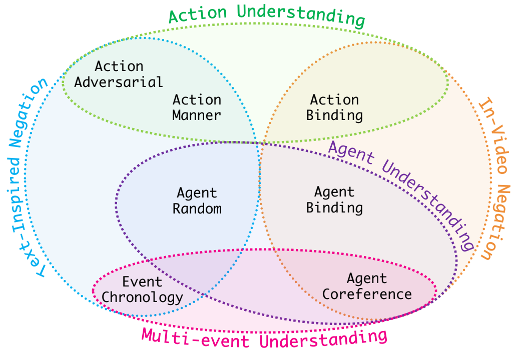


Figure 1. A Venn diagram grouping VELOCITI's seven tests (in black) that evaluate a Video-LLM across different facets: Agent Understanding, Action Understanding, and Multi-event Understanding. The benchmark is formulated as video-language entailment, where negative captions are created by manipulating text (Text-inspired Negation) or from other parts of the same video (In-Video Negation). Best seen in color.

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 [5, 9, 11, 23, 45]. A similar approach is adapted for videos to create Video-LLMs [23, 28, 45, 48]. Keeping pace with the development of new models, there is a flurry of work on evaluating them (Tab. 1). Video researchers are also creating benchmarks to study long video comprehension [7, 13, 15, 37]. 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

☯ Equal contribution

[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).

that studies **V**ideo **e**t **L**anguage **C**ompos**i**tionality through **Ti**me. We adopt the video-language entailment (VLE) evaluation setup [4] where a model is prompted to predict whether a video entails a caption ('Yes' for an aligned or positive caption and 'No' for a misaligned or negative 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 *multiple events* through time. As illustrated in Fig. 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). Note that although the tests (Sec. 3.2) have varying levels of difficulty, they are all important as they shed light on whether a model is able to solve a specific facet of video-language compositional reasoning.

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 [39]. 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.

Strict entailment. In the classic VLE setup, benchmarks typically check if the entailment score for the positive caption is higher than the negative caption [25, 40, 47]. While this traces back to visual-semantic embedding models [12, 14, 35], it is unsuitable for evaluating modern Video-LLMs that generate text (and not similarity scores).

We propose a strict entailment scoring mechanism where Video-LLMs should output 'Yes' for an aligned caption and 'No' for a misaligned one. Our analysis reveals that models produce marginally different entailment scores for the positive and negative captions attaining good performance on ClassicVLE, but predict 'Yes' for both. This is critical as VLE evaluates a model's ability to reject (partially) misaligned descriptions. Poor performance here implies that the model may produce erroneous outputs on other tasks (*e.g*. question-answering) and assumes that partial hallucinations (like negative captions) are acceptable.

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 (Sec. 3). (ii) We propose a strict metric for video-language entailment that requires a model to produce 'Yes' for an aligned caption *and* 'No' for the corresponding misaligned caption (Sec. 4). (iii) We evaluate both open and closed models and show that they struggle with compositional reasoning. While larger models such as LLaVA-OneVision-72B (OV-72B) [23] tend to perform better than smaller ones (OV-7B), even the best commercial model (Gemini-1.5-Pro [10]) achieves 49.3% accuracy, about half that of humans at 93.0%. (iv) Our experiments reveal important findings: a) Understanding actions is harder than agents for open models, and b) tests incorporating in-video negation are more challenging than text-inspired negation (Sec. 5.1). c) Smaller models are predisposed towards 'Yes' for the entailment task (Sec. 5.2). d) ClassicVLE hides information as entailment scores of positive and negative captions are often close to each other, likely due to subtle differences between them (Sec. 5.3). e) Multiple-choice (MC) evaluation is unsuitable due to a choice bias observed even in large and closed models (Sec. 5.4). f) Finally, we show that VELOCITI requires visual inputs and multiple frames and cannot be solved with text-only or single-frame models (Sec. 5.5).

2. Related Work

Several benchmarks exist to evaluate image-language compositionality (Winoground [42], COLA [38], MMVP [44], and others [17, 26, 32, 49, 52]). They require identifying the correct caption among distractors, exposing models failure to bind concepts [20]. We focus on short complex videos.

Video-language benchmarks broadly related to our work are presented in Tab. 1. We discuss differences to closely related work here. Among previous benchmarks that study compositional reasoning, our work differs due to the (i) emphasis on a test suite that provides disentangled understanding of agents and actions across multiple events; (ii) task formulation as *strict* video-language entailment (unlike TestOfTime [3], VideoCon [4], VITATECS [25], Vinoground [51]); (iii) explicit use of text-inspired *and* in-video negation (unlike TestOfTime [3], VideoCon [4], MVBench [24], VITATECS [25]); and (iv) use of short complex videos (*e.g*. compared to indoor, single-agent Charades [41] in AGQA [16], STAR [46]).

While contrast captions are a popular strategy [3, 4, 8, 24, 25, 33, 51], the structured SRL annotations used in VELOCITI facilitate evaluating specific aspects of a model's capabilities. Further, different from comprehensive benchmarks that evaluate holistic video understanding (*e.g*. SEED-Bench-2 [21], CVRR-ES [19], Video-MME [15]), we focus on the fundamental ability of compositional understanding and highlight major shortcomings. Importantly, our tests are designed to prevent text-only and single-frame models from solving them (validated empirically), guarding against issues highlighted by ATP [6] and recently TVBench [8].

Video-Language Entailment (VLE) is posed as a binary classification task [40]. Given a premise (the video) and a hypothesis (the caption), a model should determine if the

Benchmark	Task Setup	Comp	In-V Neg	Strict VLE	Test Creation	Human Eval	Video Duration	Domain (Source)
AGQA [16] CVPR'21	OQA, MCQ	✓	✗	*NA*	T, SG	✓	30s	Open (ActionGenome, Charades)
STAR [46] NeurIPSDB'21	MCQ	✓	✓	*NA*	H, T, SG	✗	30s	Indoor (Charades)
ContrastSets [33] NAACL'22	MCQ	✓	✗	*NA*	H, T, LLM	✓	-	Mixed (MSR-VTT, LSMDC)
TestOfTime [3] CVPR'23	E	✗	✗	✗	T	✗	5-30s	Open (TEMPO, ANet Cap., Charades)
Perception Test [34] NeurIPSDB'23	MCQ	✗	✗	*NA*	H	✓	23s	Indoor (Manual)
Cinepile [37] CVPRW'24	MCQ	✗	✗	*NA*	H, GPT-4	✓	2-3m	Movies (MovieClips channel)
VideoCon [4] CVPR'24	E, OQA	✓	✗	✗	H, PaLM-2	✗	10-30s	Open (MSR-VTT, VATEX, TEMPO)
SEED-Bench-2 [21] CVPR'24	MCQ	✗	✗	*NA*	H, GPT-4	✗	-	Open (Charades, SSV2, EK100)
MV-Bench [24] CVPR'24	MCQ	✓	✗	*NA*	T, ChatGPT	✗	5-35s	Mixed (Charades-STA, MoVQA, $_{+9}$)
TempCompass [31] ACLFindings'24	MCQ, E, VC	✓	✗	✗	H, GPT-3.5	✓	30s	Open (ShutterStock)
MMBench-Video [13] NeurIPSDB'24	OQA	✗	✗	*NA*	H	✗	30s-6m	Open (YouTube)
VITATECS [25] ECCV'24	E	✓	✗	✗	H, GPT-3.5	✓	10s	Open (MSRVTT, VATEX)
CVRR-ES [19] arXiv-2405	OQA	✗	✗	*NA*	H, GPT-3.5	✓	2-183s	Open (SSV2, CATER, $_{+5}$)
Video-MME [15] arXiv-2405	MCQ	✗	✓	*NA*	H	✗	11s-1h	Open (YouTube)
VideoVista [27] arXiv-2406	MCQ	✗	✗	*NA*	T, GPT-4, GPT-4o	✗	131s	Mixed (Panda-70M)
Vinoground [51] arXiv-2410	E	✓	✗	✗	H, GPT-4	✓	10s	Open (VATEX)
TVBench [8] arXiv-2410	MCQ	✓	✗	*NA*	T	✗	-	Mixed (STAR, CLEVRER, $_{+6}$)
VELOCITI (Ours)	E	✓	✓	✓	H, T, LLM	✓	10s	Movies (VidSitu)

Table 1. We review video-language benchmarks and highlight key differences to VELOCITI. Benchmarks use various 'Task Setups': Entailment (E), Multiple Choice (MCQ), Open-ended Question-Answering (OQA), and Video Captioning (VC). We compare VELOCITI against benchmarks that test Compositionality ('Comp') or have In-Video Negation ('In-V Neg'). In 'StrictVLE', benchmarks not adopting VLE are marked not applicable (*NA*). Acronyms in the 'Test Creation' column are: template (T), scene graph (SG), open large language model (LLM), and human (H). The 'Domains' are of 3 types: Open (natural videos), Movies, and Mixed (natural & movies). Different from others, VELOCITI introduces StrictVLE and features tests with negative captions created from entities appearing in the same video.

hypothesis logically follows (entails) from the premise. Entailment was first used with images in [47] and adopted by [3, 4, 25] for videos. Given the rise of Vision LLMs, entailment scores are computed using the likelihood over specific words in the vocabulary [4, 22, 29]. However, most works only require that the positive caption scores higher than the negative [25, 47], or with a margin [22]. We propose a more demanding form, *StrictVLE*, that unlike ClassicVLE, is applicable to both open and closed models (without likelihood scores). Specifically, we independently require that the model entails the positive caption *and* does not entail the negative caption. While this looks simple, we find that models do not sufficiently distinguish positive and negative captions and tend to answer 'Yes' for both.

3. 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. From SRL to a Video-Caption Pair

In VidSitu, videos are divided into five 2 s events [39] (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).

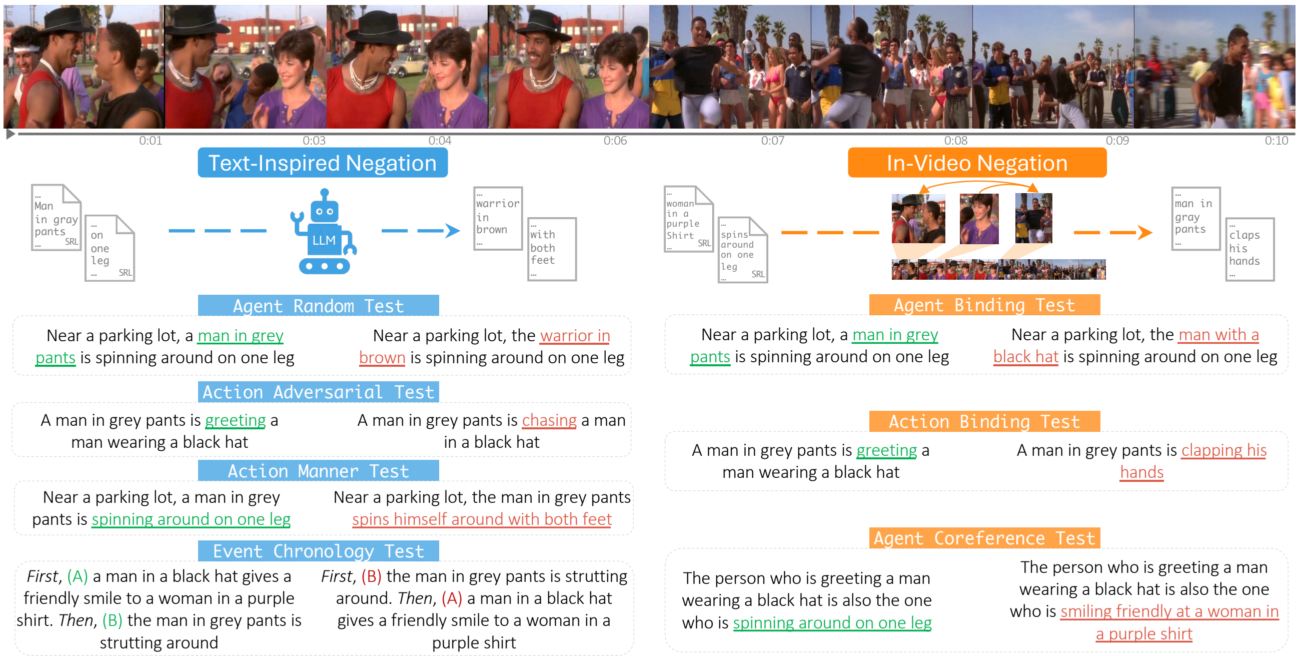
We use an open LLM (LLaMA-3 [2]) to convert the structured SRL dictionary of each event into a caption. The LLM is prompted to combine the atomic concepts into a fluent caption (prompt in the supplement). We filter 864 videos from the validation set and generate 3101 high-quality (V, C) pairs with captions 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 [6].

3.2. VELOCITI Tests

We motivate and describe the seven tests below. Fig. 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



Agent Random Test

Near a parking lot, a man in grey pants is spinning around on one leg | Near a parking lot, the warrior in brown is spinning around on one leg

Action Adversarial Test

A man in grey pants is greeting a man wearing a black hat | A man in grey pants is chasing a man in a black hat

Action Manner Test

Near a parking lot, a man in grey pants is spinning around on one leg | Near a parking lot, the man in grey pants spins himself around with both feet

Event Chronology Test

First, (A) a man in a black hat gives a friendly smile to a woman in a purple shirt. *Then*, (B) the man in grey pants is strutting around | *First*, (B) the man in grey pants is strutting around. *Then*, (A) a man in a black hat gives a friendly smile to a woman in a purple shirt

Agent Binding Test

Near a parking lot, a man in grey pants is spinning around on one leg | Near a parking lot, the man with a black hat is spinning around on one leg

Action Binding Test

A man in grey pants is greeting a man wearing a black hat | A man in grey pants is clapping his hands

Agent Coreference Test

The person who is greeting a man wearing a black hat is also the one who is spinning around on one leg | The person who is greeting a man wearing a black hat is also the one who is smiling friendly at a woman in a purple shirt

Figure 2. VELOCITI evaluates Video-LLMs' video-language entailment capabilities on complex movie clips with dense semantic role label (SRL) annotations from the VidSitu dataset [39]. 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 Sec. 3.2. The 10 s video clip used in this example can be viewed here: `https://www.youtube.com/embed/bt6-F11LZsQ?start=25&end=35`.

baseline understanding. Here, C^+ is as described in Sec. 3.1 and C^- is simply a positive caption of some other random video, making it easily discernible.

1. Agent Random Test. 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 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. 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. Coreference groups two or more phrases that refer to the same entity [18]. In a video, an agent can be referred to by their actions, *e.g*. in Fig. 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. C^+ is as described in Sec. 3.1 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 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. 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. 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* [3]. C^+ is created by concatenating event captions (from Sec. 3.1) 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 s apart to prevent a chance of overlap.

Quality control. All test samples in VELOCITI are verified by humans to ensure that C^+ aligns with the video and C^- is misaligned. The number of samples in each test and other details are presented in the supplement.

4. StrictVLE Evaluation Metric

We adopt Video-Language Entailment (VLE) as the evaluation scheme for VELOCITI. 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'. We define the entailment score similar to [40]:

$$e(V, C) = \frac{p_M(\text{'Yes'}|I(V, C))}{p_M(\text{'Yes'}|I(V, C)) + p_M(\text{'No'}|I(V, C))}, \quad (1)$$

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

ClassicVLE [25, 47] considers that a model is correct when $e(V, C^+) > e(V, C^-)$. The random accuracy is 50%.

Narrative example. Consider a simple video of a red traffic light. Let C^+, "The traffic light is red" score $p(\text{'Yes'}){=}0.7$, $p(\text{'No'}){=}0.3$; and C^-, "The traffic light is green" score $p(\text{'Yes'}){=}0.6$, $p(\text{'No'}){=}0.4$. As $e(V, C^+) > e(V, C^-)$, ClassicVLE considers this as a correct prediction. However, with greedy decoding (constrained to 'Yes' and 'No'), the model will predict 'Yes' for C^-, which is objectively incorrect, and in this example scenario, dangerous.

StrictVLE. As models improve, it is important for our community to hold them to higher standards. We argue that relative ordering of the entailment scores is insufficient and we propose StrictVLE that requires models to predict 'Yes' for C^+ *and* 'No' for the corresponding C^-. Specifically, StrictVLE considers a sample correct iff $e(V, C^+){>}0.5 \wedge e(V, C^-){<}0.5$[2] and has a random chance accuracy of 25%. The threshold 0.5 arises naturally, and is equivalent to greedy decoding of 'Yes'/'No' in the response. This equivalence means StrictVLE also works on closed models without access to p_M.

Relation to multiple-choice (MC). While VLE performs *independent* evaluation of C^+ and C^-, MC provides *both* captions to the model at once. Seeing both captions makes the task easier as the model needs to predict the *more likely* option rather than independently assess each one (similar to ClassicVLE). In Sec. 5.4 we reveal biases of MC evaluation and show that our proposed StrictVLE is preferable.

5. Results and Discussion

We evaluate open and closed Video-LLMs on VELOCITI. First, we present results with StrictVLE, our primary evaluation strategy, and analyze entailment scores. We also compare results with ClassicVLE and MC, discussing some evaluation pitfalls. Finally, we evaluate blind and single-frame models to check for bias in the benchmark.

Models. We evaluate multiple open Video-LLMs: PLLaVA [48], Video-LLaVA (V-LLaVA) [28], Owl-Con [4], Qwen2-VL (QVL) [45], and LLaVA-OneVision (OV) [23]; closed models Gemini-1.5-Flash (Gem-1.5F), Gemini-1.5-Pro (Gem-1.5P) [10], GPT-4o [1]; and humans. Due to compute and cost constraints, we evaluate QVL-72B (at native video resolution), closed models, and humans on a subset of 150 samples from each test (created once by random selection). More details in the supplement.

5.1. Evaluation with StrictVLE

We report model performance in Tab. 2 and discuss various facets of model understanding highlighted in Fig. 1 (agents, actions, multiple events, and negation strategies).

Control *vs*. VELOCITI average. Old open models (P-LLaVA, Owl-Con) struggle on the StrictVLE setup as they have a strong bias to predict 'Yes'. This leads to poor performance on the control test and the benchmark average (first and last column). V-LLaVA and subsequent models, QVL and OV, obtain decent accuracies on the control test, ranging from 65-85%. Compared to the control tests, performance dips strongly on the benchmark average, with the best model, OV-72B obtaining 43.2% accuracy (36.2% lower than control). On the benchmark subset, while GPT-4o posts a 46.2% accuracy on average, it performs poorly on the control tests (analysis in Sec. 5.2). Inversely, Gem-1.5F achieves 91.9% on the control tests, but is close to random on the benchmark (23.9%). The best model, Gem-1.5P,

[2] Note, this is different from the Winoground [42] setup that has: (i) 2 images/videos and 2 captions; and (ii) still uses relative scoring.

Model	Ctrl	Ag Rand	Ag Bind	Ag Cref	Act Adv	Act Man	Act Bind	Ev Chr	Avg
Random	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0
P-LLaVA	1.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Owl-Con	24.3	3.4	0.7	0.0	4.3	2.8	0.6	0.1	1.7
V-LLaVA	65.8	16.4	7.6	0.3	8.7	3.3	10.6	3.9	7.3
QVL-7B	**84.6**	39.1	13.5	6.5	17.8	17.5	16.4	0.4	15.9
OV-7B	81.6	56.7	32.9	8.0	29.7	**30.6**	36.4	30.5	32.1
OV-72B	79.3	**63.7**	**45.4**	**38.6**	**33.1**	29.3	**45.1**	**46.5**	**43.1**
VELOCITI Subset									
Gem-1.5F	**91.9**	56.4	23.8	4.7	32.9	21.6	25.0	2.7	23.9
QVL-72B	82.7	56.0	29.3	35.3	30.0	24.0	35.3	1.3	30.2
OV-72B	81.3	**64.0**	46.7	**41.3**	30.7	32.7	46.0	50.0	44.5
GPT-4o	63.3	54.7	44.7	40.7	**55.0**	42.0	**54.0**	32.2	46.2
Gem-1.5P	74.3	60.1	**49.7**	36.7	52.3	**43.5**	52.3	**50.3**	**49.3**
Human	-	91.5	92.9	92.6	92.9	89.9	91.5	100.	93.0

Table 2. Results on VELOCITI using the **StrictVLE** evaluation strategy. The tests are abbreviated as Ctrl (Control), AgRand (Agent Random), AgBind (Agent Binding), AgCref (Agent Coreference), ActAdv (Action Adversarial), ActMan (Action Manner), ActBind (Action Binding), EvChr (Event Chronology). Avg reports the average accuracy on the 7 tests of VELOCITI. All models show a large gap to human performance.

achieves 49.3%, far from human performance at 93.0%. VELOCITI is a challenging benchmark and exposes lack of reasoning in both open and closed Video-LLMs.

Agent understanding tests include the Agent Random Test (AgRand), Agent Binding Test (AgBind), and Agent Coreference Test (AgCref). Broadly, they evaluate a model's ability to understand the *doer* of the actions in the videos. Compared to AgRand, models show worse performance on AgBind and AgCref. For example, the best performing OV-72B, achieves 63.7%, 45.4%, and 38.6% accuracy respectively. AgRand requires verifying the *presence* of the agent, AgBind requires disambiguating between people present in the video and *binding* the correct person with the event description, and AgCref needs resolving identity across *multiple events*. This proves the difficulty of in-video negation. We also note that OV-7B performs worse than OV-72B on complex tests (AgCref, 8.0% *vs*. 38.6%) indicating that multi-level reasoning is slightly better with larger LLMs.

Action understanding tests include the Action Adversarial Test (ActAdv), Action Manner Test (ActMan), and Action Binding Test (ActBind). These evaluate the model's understanding of actions and/or its modifiers. We observe that OV-72B scores worse on action tests (35.8% average over the 3 tests) as compared to agent tests (49.2%), while GPT-4o achieves a balanced performance of 50.3% and 46.7% respectively. While ActAdv is easier than ActBind for most models, OV shows inverted results. Further, subtle variations in actions are not captured by most models and ActMan is a challenging test with OV-72B at 29.3% and Gem-1.5P posting the highest accuracy of 43.5%.

Multi-event understanding tests. As AgBind and ActBind adopt in-video negation, they require some level of multi-event reasoning, but are ignored in this discussion. Instead, we focus on Agent Coreference Test (AgCref) and Event Chronology Test (EvChr) as they have multiple events in both captions. Time and event order are critical to video comprehension. However, Video-LLMs are still poor at the EvChr test that requires establishing the relative order of two events. Apart from OV-72B (46.5%) and Gem-1.5P (subset, 50.3%), all models are comparable to or worse than random. This is likely as all entities mentioned in both captions are present in the video. AgCref fairs slightly better, with more models showing performance better than random: QVL-72B (35.3%), OV-72B (38.6%), Gem-1.5P (36.7%), GPT-4o (40.7%). However, it is concerning that the smaller OV-7B model collapses on these tests (AgCref 8.0%, EvChr 30.5%). Both tests highlight challenges of reasoning across multiple events in Video-LLMs.

Negation strategies. Finally, we observe that tests adopting in-video negation and requiring associations are harder than text-inspired negation. For GPT-4o that achieves balanced accuracy on agent and action understanding, we observe a 5.5% drop in performance (54.9% AgRand+ActAdv to 49.4% AgBind+ActBind)[3]. Solving tests with in-video negation requires reasoning as it is insufficient to only check presence of entities (all entities from both captions appear in the video). Models need to go beyond detecting the agent and action, and learn to associate them correctly.

Qualitative analysis. Please refer to the supplement for some example predictions of OV-72B for each test.

5.2. Analyzing Entailment Scores for StrictVLE

We analyze whether a model is better at classifying C^+ or C^- in Sec. 5.2. The first number in each table cell corresponds to the accuracy of positive captions, while the second number is the accuracy of negative captions when the positive caption was correct. We see an interesting trend. As the model size increases, the positive caption accuracy decreases (85.9% → 80.4%) and negative caption accuracy increases (38.1% → 53.6%). This holds for both variants: OV-7B to OV-72B and QVL-7B to QVL-72B (although on a subset). Small models are eager to say 'Yes' for both captions, while larger models reason better. A similar trend is seen on the control tests for the 7B and 72B models. However, negative caption accuracies are far higher, confirming why control tests are easier compared to our benchmark.

Somewhat unexpectedly, GPT-4o only achieves 64.5% accuracy on positive captions. But among them, it gets the highest negative caption accuracy of 72.3%. This hesita-

[3] The chosen tests provide a head-to-head comparison of text-inspired *vs*. in-video negation with the same positive caption as seen in Fig. 2.

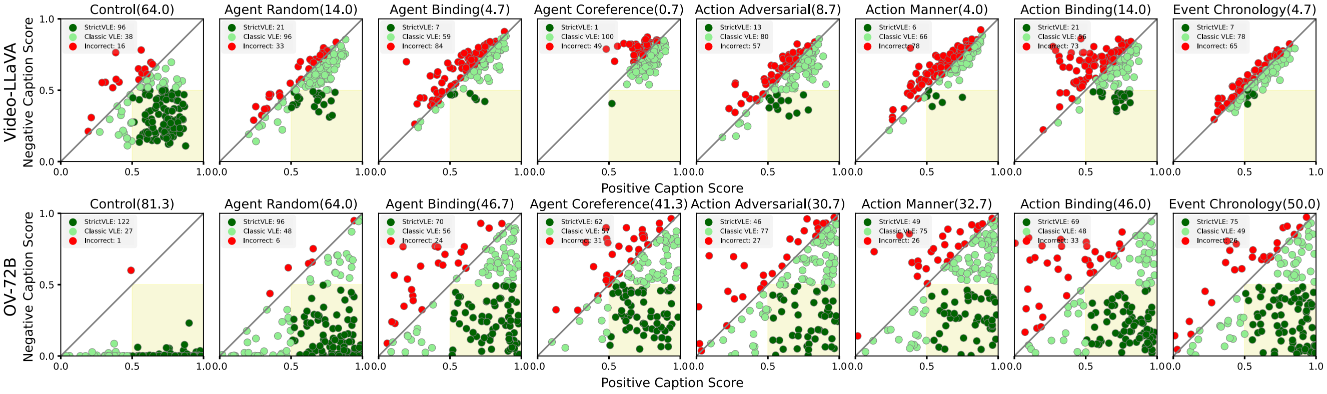


Figure 3. Scatter plot of entailment scores $e(V, C^+)$ (x-axis) and $e(V, C^-)$ (y-axis) for all tests in VELOCITI subset. We visualize the scores for Video-LLaVA (top) and OV-72B (bottom). ClassicVLE calls a sample correct in the region below the diagonal (light green). Instead, StrictVLE requires the dots to lie in the yellow bottom-right quadrant (dark green). Finally, samples whose points are above the diagonal are wrong for both VLE metrics (red). While recent models have improved, older models concentrate near the diagonal and in the top-right 'Yes' quadrant for both captions. The legend includes the actual number of points (please zoom in). Figure is best seen in color.

Model	Control	Average
OV-7B	83.0 / 98.3	85.9 / 38.1
OV-72B	79.8 / **99.3**	80.4 / **53.6**
QVL-7B	**92.4** / 91.5	**93.9** / 17.1
VELOCITI Subset		
QVL-72B	84.0 / 98.4	85.2 / 36.4
Gem-1.5F	**93.9** / 97.8	**95.8** / 25.4
GPT-4o	63.0 / **100.**	64.5 / **72.3**
Gem-1.5P	75.0 / 99.1	74.0 / 66.2

Table 3. StrictVLE Analysis. We study a model's failure modes via positive and negative caption accuracy. Each cell shows the fraction of: (i) correctly classified positive captions; and (ii) correctly classified negative captions among samples whose positive captions are correct.

tion to say 'Yes' hurts GPT-4o on the control tests as well and even though negative caption accuracy is perfect, it gets many positive captions wrong. Similar analysis of each test (in the supplement) shows that harder tests tend to have lower negative caption accuracies.

5.3. Evaluation with ClassicVLE

While we recommend StrictVLE, we present results on the ClassicVLE setup (Tab. 4) for completeness. First, we present a language only baseline that evaluates if C^+ is more plausible than C^-. VERA [30] scores 58.3% (close to random 50%), confirming that language biases are insufficient to solve the tests. Next, we evaluate CLIP-style models [36, 49, 50] that mean-pool video frames and observe a small improvement (ViFi-C 61.2%). New Video-LLMs such as QVL and OV (OV-72B: 83.6%) show good improvement over older ones (*e.g.* P-LLaVA: 59.5%). However, this score is worse than 99.4% on the easy control tests. Even with a relaxed metric, OV-72B gets every sixth sample wrong. The trends for agent and action understanding are similar: AgRand > AgBind > AgCref, and QVL and OV perform better on agent than action understanding.

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 4. Evaluation with **ClassicVLE**. Random accuracy is 50%. Beyond Video-LLMs, we report results for a plausibility-evaluation model (VERA) and contrastive models (SigLIP: ViT-SO400M-14-SigLIP-384 [50], ViFi-C: VIFICLIP-B16 [36], and Neg-C: NegCLIP-B32 [49]). The performance of contrastive models and older Video-LLMs is close to random. However, recent models (*e.g.* OV) produce better relative entailment scores, even if they generate incorrect 'Yes'/'No' responses.

To further analyze entailment scores, we present scatter plots on the benchmark subset in Fig. 3. While OV-72B is clearly better than Video-LLaVA, it has too many points in the top-right quadrant indicating a bias to say 'Yes' to both captions. For Video-LLaVA, it is concerning that scores are close to the diagonal (*i.e.* both C^+ and C^- get similar entailment scores). In fact, these plots motivate us to propose StrictVLE and reveal problems hidden by ClassicVLE. We visualize such plots for all models in the supplement.

5.4. MC Evaluations and Choice Bias

In this setup, we provide the video and both captions to the Video-LLM and ask it to pick the correct description (A or B, see prompt in supplement). In Tab. 5, we report accuracy of the model where C^+ is option A or option B. We see

Model	Control				Benchmark Average			
	A	B	Bias	A∧B	A	B	Bias	A∧B
Random	50.0	50.0	-	25.0	50.0	50.0	-	25.0
QVL-7B	94.9	98.5	(+3.6)	94.6	38.9	88.0	(+49.1)	38.5
OV-7B	96.0	99.6	(+3.6)	95.9	28.7	96.5	(+67.8)	28.7
OV-72B	99.2	99.4	(+0.2)	**99.0**	78.0	88.3	(+10.3)	**76.0**
VELOCITI Subset								
QVL-72B	100.	100.	(+0.0)	**100.**	73.1	75.7	(+2.6)	65.3
OV-72B	100.	100.	(+0.0)	**100.**	77.1	87.8	(+10.7)	**75.0**
Gem-1.5F	100.	99.3	(-0.7)	99.3	85.1	73.2	(-11.9)	67.7
GPT-4o	100.	100.	(+0.0)	**100.**	83.9	74.8	(–9.1)	68.6

Table 5. Multi-choice (MC) evaluation results. Along with video, we provide the model both captions as A and B and ask it to pick the better aligned one. Column headers A (or B) refer to the accuracy when A (or B) is the positive caption. Bias is B minus A and should be close to 0. A∧B involves evaluating the model twice, once with correct caption as A and again as B. A sample is deemed correct when it picks the correct choice in both cases. While a model's decision should be unaffected by the order in which choices are presented, we see a considerable bias.

that small 7B models have a strong choice bias and pick option B more than A (49.1% QVL-7B or 67.8% OV-7B). While this reduces in larger models (10.7% OV-72B), it is still high. Even closed models exhibit this behavior with Gem-1.5F preferring option A over B (11.9%) and GPT-4o preferring option A over B (9.1%). Interestingly, this bias becomes a major issue when the tests are challenging and is negligible in the control tests that are easier.

While one could report accuracy by running the model twice, once with option A as C^+ and again with B as C^+ (referred as A∧B), this is tedious and the number of evaluations increases as a factorial of the number of choices. If we compare StrictVLE with the MC evaluation's A∧B score (both apply ∧ on binary decisions and have random chance at 25%), we observe that MC is much easier (OV-72B: 76.0%) than StrictVLE (43.1%, Tab. 2). This may be attributed to the MC setup, where a model processes both captions at once and only needs to pick the more likely option; in contrast with the StrictVLE setup that requires independent evaluation of each caption. Even though MC is a popular evaluation setup for many benchmarks (see Tab. 1), the choice bias of Video-LLMs makes results difficult to interpret. For all these reasons, StrictVLE is preferred.

5.5. Validating Benchmark Properties

We highlight some additional properties of our benchmark.

Evaluating blind models. Tab. 6A compares Qwen2-LLM (Q LLM) and OV-72B without the video inputs (OV Blind) against the default OV-72B model (here, OV 👁). We see a dramatic drop (43.1% to 3.7% Q LLM and 8.1% OV Blind). Solving tests in VELOCITI requires visual understanding.

Evaluating with a single-frame. Tab. 6B reports results on OV-7B models. The first row is the default setup (Tab. 2) and is compared against: (i) OV-7B with a single frame input (1 Frame) chosen at random from the sampled 1fps frames. (ii) The OneVision team [23] first train an image-only model and extend it to multiple images and videos. We evaluate their single image checkpoint while providing video inputs (OV-7B-SI). The performance drops from 32.1% to about 19.0% in both cases. This confirms that VELOCITI requires video inputs and video models.

FPS and chain-of-thought ablations are in supplement.

Model	Ctrl	Ag Rand	Ag Bind	Ag Cref	Act Adv	Act Man	Act Bind	Ev Chr	Avg
A. Comparing against Blind Models									
OV 👁	79.3	63.7	45.4	38.6	33.1	29.3	45.1	46.5	**43.1**
Q LLM	2.2	2.5	2.2	5.3	4.1	1.3	2.7	7.9	3.7
OV Blind	6.0	9.3	6.7	12.7	10.3	3.3	6.9	7.4	8.1
B. Impact of Single Frame Input or Model									
OV-7B	81.6	56.7	32.9	8.0	29.7	30.6	36.4	30.5	**32.1**
1 Frame	39.6	31.6	22.3	18.1	15.5	13.4	22.1	10.3	19.0
OV-7B-SI	78.9	35.7	15.6	2.9	27.6	21.4	22.0	8.8	19.1

Table 6. VELOCITI benchmark validation. **Part A.** We confirm that the model requires visual inputs. The base LLM Qwen2-72B (Q LLM) or OV-72B without providing the video (OV Blind) perform poorly compared to OV-72B provided with video frames (OV 👁). **Part B.** We also confirm that providing multiple video frames is necessary. When OV-7B is provided a single frame chosen randomly (1 Frame) or when the video is fed to a model trained only on single images (LLaVA-OneVision-SingleImage, OV-7B-SI), the performance dips compared to showing the video at 1fps (our default strategy, OV-7B).

6. Conclusion

We introduced VELOCITI, a benchmark to evaluate the compositional capabilities of Video-LLMs by disentangling and assessing the comprehension of agents, actions, and their associations across multiple events. We improved over the classic Video-Language Entailment setup that relies on relative scoring by proposing StrictVLE that requires models to answer 'Yes' for the positive caption *and* 'No' for the negative caption. All evaluated models, open and closed, performed poorly with a large gap to human performance. Our experiments showed that action understanding is harder than agent understanding, and solving tests with in-video negation is harder than text-inspired ones. We also analyzed limitations of ClassicVLE and the choice bias in multiple-choice evaluations. Overall, our work established that compositional reasoning on short videos is still unsolved and remains challenging for Video-LLMs.

Acknowledgment. MT and ViG thank Adobe Research for supporting this project and travel. MT thanks SERB SRG/2023/002544 for compute server and Google Cloud credits. We thank volunteers for human evaluations.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*, 2023. 5

[2] Meta AI. Llama3. `https://llama.meta.com/llama3/`, 2024. 3

[3] Piyush Bagad, Makarand Tapaswi, and Cees GM Snoek. Test of Time: Instilling Video-Language Models With a Sense of Time. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 3, 5

[4] Hritik Bansal, Yonatan Bitton, Idan Szpektor, Kai-Wei Chang, and Aditya Grover. VideoCon: Robust Video-Language Alignment via Contrast Captions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 3, 5

[5] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. PaliGemma: A versatile 3B VLM for transfer. *arXiv preprint arXiv:2407.07726*, 2024. 1

[6] Shyamal Buch, Cristóbal Eyzaguirre, Adrien Gaidon, Jiajun Wu, Li Fei-Fei, and Juan Carlos Niebles. Revisiting the "Video" in Video-Language Understanding. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 3

[7] Mu Cai, Reuben Tan, Jianrui Zhang, Bocheng Zou, Kai Zhang, Feng Yao, Fangrui Zhu, Jing Gu, Yiwu Zhong, Yuzhang Shang, Yao Dou, Jaden Park, Jianfeng Gao, Yong Jae Lee, and Jianwei Yang. TemporalBench: Towards Fine-grained Temporal Understanding for Multimodal Video Models. *arXiv preprint arXiv:2410.10818*, 2024. 1

[8] Daniel Cores, Michael Dorkenwald, Manuel Mucientes, Cees GM Snoek, and Yuki M Asano. TVBench: Redesigning Video-Language Evaluation. *arXiv preprint arXiv:2410.07752*, 2024. 2, 3

[9] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 1

[10] Google Deepmind. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint, arXiv: 2403.05530*, 2024. 2, 5

[11] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and PixMo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*, 2024. 1

[12] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. In *British Machine Vision Conference (BMVC)*, 2018. 2

[13] Xinyu Fang, Kangrui Mao, Haodong Duan, Xiangyu Zhao, Yining Li, Dahua Lin, and Kai 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. 1, 3, 13

[14] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2013. 2

[15] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Rongrong Ji, and Xing Sun. Video-MME: The First-Ever Comprehensive Evaluation of Multi-modal LLMs in Video Analysis. *arXiv preprint*, 2024. 1, 2, 3, 13

[16] Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. AGQA: A Benchmark for Compositional Spatio-Temporal Reasoning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 3

[17] Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. SugarCrepe: Fixing Hackable Benchmarks for Vision-Language Compositionality. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 2

[18] Daniel Jurafsky and James H. Martin. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models, 2024. 4

[19] Muhammad Uzair Khattak, Muhammad Ferjad Naeem, Jameel Hassan, Muzammal Naseer, Federico Tombari, Fahad Shahbaz Khan, and Salman Khan. How Good is my Video LMM? Complex Video Reasoning and Robustness Evaluation Suite for Video-LMMs. *arXiv preprint arXiv:2405.03690*, 2024. 2, 3

[20] Martha Lewis, Nihal Nayak, Peilin Yu, Jack Merullo, Qinan Yu, Stephen Bach, and Ellie Pavlick. Does CLIP Bind Concepts? Probing Compositionality in Large Image Models. In *European Chapter of the Association for Computational Linguistics (EACL)*, 2024. 2

[21] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. SEED-Bench-2: Benchmarking Multimodal Large Language Models. *arXiv preprint arXiv:2311.17092*, 2023. 2, 3

[22] Baiqi Li, Zhiqiu Lin, Wenxuan Peng, Jean de Dieu Nyandwi, Daniel Jiang, Zixian Ma, Simran Khanuja, Ranjay Krishna, Graham Neubig, and Deva Ramanan. NaturalBench: Evaluating Vision-Language Models on Natural Adversarial Samples. *arXiv preprint arXiv:2410.14669*, 2024. 3

[23] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy Visual Task Transfer. *arXiv preprint arXiv:2408.03326*, 2024. 1, 2, 5, 8

[24] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. MVBench: A Comprehensive Multi-modal Video Understanding Benchmark. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 3

[25] Shicheng Li, Lei Li, Shuhuai Ren, Yuanxin Liu, Yi Liu, Rundong Gao, Xu Sun, and Lu Hou. VITATECS: A Diagnostic Dataset for Temporal Concept Understanding of Video-Language Models. In *European Conference on Computer Vision (ECCV)*, 2024. 2, 3, 5, 15

[26] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating Object Hallucination in Large Vision-Language Models. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2023. 2

[27] Yunxin Li, Xinyu Chen, Baotian Hu, Longyue Wang, Haoyuan Shi, and Min Zhang. VideoVista: A Versatile Benchmark for Video Understanding and Reasoning. *arXiv preprint arXiv:2406.11303*, 2024. 3

[28] Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-LLaVA: Learning United Visual Representation by Alignment Before Projection. *arXiv preprint arXiv:2311.10122*, 2023. 1, 5

[29] Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating Text-to-Visual Generation with Image-to-Text Generation. In *European Conference on Computer Vision (ECCV)*, 2024. 3

[30] Jiacheng Liu, Wenya Wang, Dianzhuo Wang, Noah A Smith, Yejin Choi, and Hannaneh Hajishirzi. Vera: A General-Purpose Plausibility Estimation Model for Commonsense Statements. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2023. 7

[31] Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. TempCompass: Do Video LLMs Really Understand Videos? In *Findings of the Association for Computational Linguistics*, 2024. 3

[32] Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. CREPE: Can Vision-Language Foundation Models Reason Compositionally? In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[33] Jae Sung Park, Sheng Shen, Ali Farhadi, Trevor Darrell, Yejin Choi, and Anna 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. 2, 3

[34] Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Mateusz Malinowski, Yi Yang, Carl Doersch, et al. Perception Test: A Diagnostic Benchmark for Multimodal Video Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 3

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning (ICML)*, 2021. 2

[36] Hanoona Rasheed, Muhammad Uzair Khattak, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Fine-Tuned CLIP Models Are Efficient Video Learners. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 7

[37] Ruchit Rawal, Khalid Saifullah, Ronen Basri, David Jacobs, Gowthami Somepalli, and Tom Goldstein. CinePile: A Long Video Question Answering Dataset and Benchmark. *arXiv preprint arXiv:2405.08813*, 2024. 1, 3

[38] Arijit Ray, Filip Radenovic, Abhimanyu Dubey, Bryan Plummer, Ranjay Krishna, and Kate Saenko. Cola: A Benchmark for Compositional Text-to-image Retrieval. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 2

[39] Arka Sadhu, Tanmay Gupta, Mark Yatskar, Ram Nevatia, and Aniruddha Kembhavi. Visual Semantic Role Labeling for Video Understanding. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 3, 4, 12, 15

[40] 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*, 2024. 2, 5

[41] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding. In *European Conference on Computer Vision (ECCV)*, 2016. 2

[42] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing Vision and Language Models for Visio-Linguistic Compositionality. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 5

[43] Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. Label Studio: Data labeling software, 2020-2022. Open source software available from https://github.com/heartexlabs/label-studio. 5

[44] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[45] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. *arXiv preprint arXiv:2409.12191*, 2024. 1, 5

[46] Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B. Tenenbaum, and Chuang Gan. STAR: A Benchmark for Situated Reasoning in Real-World Videos. In *Advances in Neural Information Processing Systems (NeurIPS): Track on Datasets and Benchmarks*, 2021. 2, 3

[47] Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual Entailment Task for Visually-grounded Language Learning. In *Visually Grounded Interaction and Language (ViGIL) Workshop at NeurIPS*, 2018. 2, 3, 5

[48] Lin Xu, Yilin Zhao, Daquan Zhou, Zhijie Lin, See Kiong Ng, and Jiashi Feng. PLLaVA: Parameter-free LLaVA Extension

from Images to Videos for Video Dense Captioning. *arXiv preprint arXiv:2404.16994*, 2024. 1, 5

[49] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and Why Vision-Language Models Behave like Bags-Of-Words, and What to Do About It? In *International Conference on Learning Representations (ICLR)*, 2022. 2, 7

[50] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid Loss for Language Image Pre-Training . In *International Conference on Computer Vision (ICCV)*, 2023. 7

[51] Jianrui Zhang, Cai Mu, and Yong Jae Lee. Vinoground: Scrutinizing LMMs over Dense Temporal Reasoning with Short Videos. *arXiv*, 2024. 2, 3

[52] Tiancheng Zhao, Tianqi Zhang, Mingwei Zhu, Haozhan Shen, Kyusong Lee, Xiaopeng Lu, and Jianwei Yin. VL-Checklist: Evaluating Pre-trained Vision-Language Models with Objects, Attributes and Relations. *arXiv preprint arXiv:2207.00221*, 2022. 2