DynaVis: A Benchmark for Dynamic Object Understanding in Video

Kaleb S. Newman
Brown University

1. Introduction

Understanding the physical world through visual input is one of the hallmarks of intelligence. The visual world is indexed by a range of objects and we humans attend to them to understand environments and their dynamic changes. Recent advances in Multimodal Large Language Models (MLLMs) and Vision-Language Models (VLMs) have brought us closer to emulating this capability in AI, yet crucial gaps remain in their ability to comprehend how the physical world unfolds over time. This thesis introduces a new benchmark aimed at evaluating the capacity of many MLLMs to understand dynamic, object-centric changes in videos. This work focuses on four core dimensions of visual reasoning: **object state recognition**, **temporal understanding**, **spatial localization**, and **counting-based reasoning**.

While my prior work has studied object states in static images [14], real-world understanding demands the capacity to process continuous events. This motivates the need to extend the work of dynamic object understanding to video. Temporal ordering of transformations (e.g., slicing before boiling), tracking the spatial movement of objects across frames, and enumerating the sequence actions in an event are essential components of physical scene understanding. Without these, AI models will continue to be jagged in their intelligence: good at summarizing video or listing objects of interest, but lacking some of the deeper, fine-grained aspects of visual intelligence. Recent multimodal large language models (MLLMs) have demonstrated impressive general visual understanding, including captioning, retrieval, and question-answering. However, understanding dynamic real-world scenes requires perceptual grounding, temporal reasoning, and spatial precision. This benchmark, which we refer to as the **Dynamic Visual Understanding Benchmark (DynaVis)**, is designed to test these capabilities using natural, egocentric and third person videos from the VOST [17] and VidOSC [21] datasets. We include multiple question formats, including multiple choice and numerical, covering various aspects of visual intelligence.

Most existing benchmarks evaluate visual understanding through image-text alignment or retrieval tasks, which, while valuable, do not capture the deeper reasoning required for the physical understanding of dynamic scenes. For example, LLaVA-Video [10] and Qwen2.5-VL [1] excel at many visual reasoning tasks, but in our benchmark they fail to consistently answer questions correctly on spatial understanding and temporal ordering. Meanwhile, work like "Thinking in Space" [22] and "Hour-Video" [2] shows that even proprietary models struggle when tested on tasks that require spatial, perceptual, and other types of visual reasoning. These insights motivated the creation of a benchmark specifically targeting **dynamic and object-centric reasoning in naturalistic videos**.

To build this benchmark, we developed a structured protocol for question generation. First, we define the core reasoning skills that we wished to evaluate:

- **State Recognition** – Can the model identify the physical states of objects (e.g., raw, cooked, sliced) and how they change?
- **Temporal Reasoning** – Can the model determine the sequence of actions or transformations?
- **Counting and Enumeration** – Can it count objects or track how many actions or state changes occurred?
- **Spatial Understanding** – Can it determine where objects move, end up, or are placed across the course of the video?

For each category, we manually authored questions over a set of videos from two datasets—26 from VOST and 5 from VidOSC—totaling 120 carefully crafted examples. These seeds were then used to guide a GPT-4o based generation pipeline, in which we supplied detailed, manually written descriptions of over 125 videos. These descriptions averaged 163.93 words and included precise accounts of object transformations, spatial configurations, and action sequences. This system produced 317 questions for VOST and 80 for VidOSC. All questions and answer choices were

manually verified for visual grounding, diversity, and correctness.

We evaluate a suite of models including LLaVA models, InternVideo models [3, 20], CogVLM2, Qwen2.5-VL [1], Vamba [16], also we run smaller, open ended, experiments with Gemini 2.5 Pro Experimental (03-25). Despite their impressive capabilities on other benchmarks, we find a consistent underperformance across all categories—especially for spatial localization and counting, which saw accuracy rates as low as 6–14%. Even state recognition and temporal understanding remain challenging in video as we show by our quantitative experiments and qualitative analysis.

We present an error taxonomy categorizing common model failure modes: hallucination, confabulation, undercounting/overcounting, incorrect temporal order. We also supplement our benchmark with qualitative error analysis, deeper insight into spatial understanding, and open-ended evaluations using Gemini Pro.

In summary, our Key Contributions:

Key Contributions

- **DynaVis**, a new benchmark designed to evaluate visual understanding of object dynamics in videos across four axes: state, temporal, spatial, and counting reasoning.
- **A suite of zero-shot evaluations** across many MLLMs.
- **A taxonomy of error types** (e.g., spatial mislocalization, hallucination) and fine-grained breakdowns of spatial understanding question types (containment, trajectory, etc.).
- **Open-ended generation experiments** showing that fluency in generation often masks deep perceptual failures, even in proprietary models like Gemini 2.5 Pro.
- **An empirical analysis of question format** effects, showing that converting counting questions to MCQ helps some models but misleads others.

We hope this work will serve as a stepping stone toward building multimodal models with deeper grounding in the physical world—models that don't just describe what they see but can track, reason, and infer about how objects change and interact over time.

2. Benchmark

In this section we discuss our Dynamic Visual Understanding Benchmark (DynaVis). This benchmark includes 517 question-answer pairs across 125 videos. We pull our videos from two different data sources and evaluate them separately.

2.1. Motivation and Design Goals

Many existing video understanding benchmarks assess models through high-level retrieval, captioning, or question-answering tasks. In another light, we focus on evaluating object-centric, dynamic visual reasoning. Specifically, we set out to rigorously assess a model's ability to reason about object transformations, track and recognize spatial relations, or temporal orderings.

Inspired by the observation that humans can effortlessly recognize when a tomato has been sliced, when liquid has filled a cup, or when a cereal box has been taken from the kitchen to the cupboard, we design our benchmark to test whether models can perform these same types of grounded perceptual inferences from video.

We focus on egocentric and third-person videos depicting natural interactions with objects—often involving cooking, cleaning, organizing, and tool use—where objects undergo state changes and spatial movement. These scenarios require temporal precision, action recognition, understanding causality and object permanence.

Our design goals are as follows :

- Emphasize **visual grounding**, requiring models to base their answers on what is visible in the frames—not just what is probable.
- Include **diverse reasoning tasks**, probing complementary abilities like spatial tracking, temporal ordering, and quantitative estimation.
- Use **naturalistic data**, using real-world videos from VOST and VidOSC instead of synthetic or overcurated clips.

2.2. Reasoning Categories

We define our benchmark based on four core categories of inference. These categories serve both as the organizational backbone of our benchmark and as explicit subskills we hypothesize to be critical for general physical scene understanding:

State Recognition Under this category, we test the model to see if it can identify the general object dynamics in the video. In particular, over the course of the video for one or multiple objects, we investigate if models can infer the current state of an object or recognize that a transformation has occurred. This includes changes in material (e.g., raw $\rightarrow$ cooked),

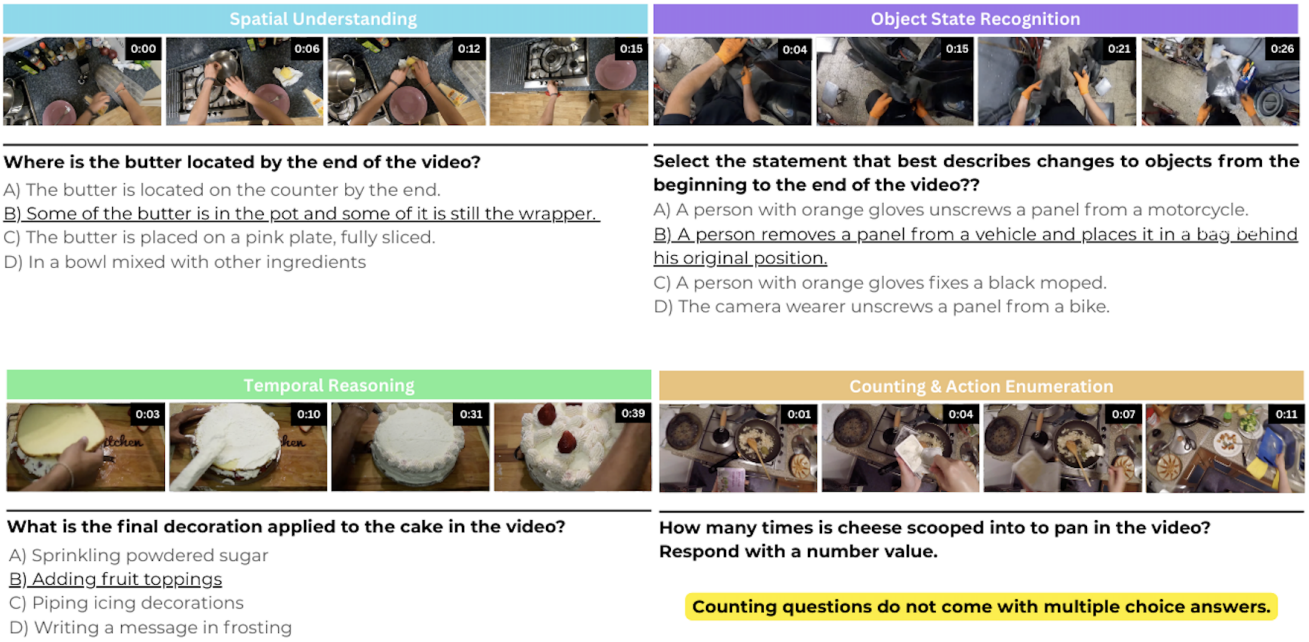

Figure 1. **Example questions from the DynaVis Benchmark** Each row represents one of the four core reasoning categories: **Spatial Understanding**, **Object State Recognition**, **Temporal Reasoning**, and **Counting & Action Enumeration**. Questions are paired with representative video frames to visually ground the task. Counting questions are open-ended and do not include multiple choice options, unlike the other categories.

form (e.g., whole $\rightarrow$ sliced), or condition (e.g., full $\rightarrow$ empty).

Temporal Reasoning For temporal reasoning, we investigate if models can recollect the order of actions or events in the video. Questions in this category require understanding the sequence in which actions and transformations occurred. Typically questions include asking which action occurred first in the video, last in the video, or in relation another action.

Counting and Enumeration This category tests MLLMs' counting abilities. Can the model count objects, actions, or state transitions? These questions measure the model's ability to track repeated interactions and maintain quantitative information across a video sequence. *We acknowledge this has been an exceedingly difficult task for feed-forward networks.*

Spatial Understanding Lastly, we test spatial understanding. Can the model track where objects are located, how they moved, or where they ended up? These questions probe for object permanence, relative positioning, and movement-based reasoning over time.

Each of these categories is designed to isolate a specific form of perceptual reasoning. These categories are also dependent on each other. For instance, correctly answering a temporal reasoning question could require identifying the states involved in a transformation.

2.3. Dataset Selection

To evaluate dynamic visual understanding in realistic conditions, we get our video data from two complementary datasets: **VOST [17]** and **VidOSC [21]**. Both datasets feature object-centric state transformations in video, but they differ in resolution, camera angle, structure, and complexity. This allows us to probe model capabilities across diverse challenges.

VOST The first dataset we use is the Video Object Segmentation under Transformations(VOST) Dataset. This dataset is uniquely suited for testing physical reasoning in video. It contains 713 high-resolution, egocentric videos, each depicting clear object transformations. This dataset was originally designed for video object segmentation through physical transformations, making it a prime candidate to be a testbed for understanding dynamic object changes. VOST captures the temporal extent of each change, and includes objects that undergo visually significant alterations in shape, texture, and physical state.

A significant aspect of VOST videos is that they are naturalistic and require minimal additional cura-

tion. The videos are human-like: egocentric and featuring no jump cuts or text overlays. As a result, the videos preserve the spatial and temporal continuity essential for grounded visual reasoning. These characteristics allow our benchmark to measure recognition, tracking and transformation understanding, under realistic conditions.

VidOSC The second dataset we use is the Video Object State Changes (VidOSC) Dataset. It complements VOST by probing model robustness under open-world, visually cluttered conditions. Sourced from instructional videos (HowTo100M [13]), VidOSC clips are longer in duration, often contain jump cuts, and feature multiple, sometimes simultaneous, object transformations. They frequently include text overlays and are from third-person point of view, introducing noise and occlusions that challenge visual grounding.

Unlike VOST, which emphasizes clarity, VidOSC pushes models to recognize specific visual elements while dealing with distractions. For instance, an object may undergo a transformation while other unrelated objects are manipulated in the same scene. The jump cuts also force models to track objects under non-continuous contexts. The inclusion of such distractors enables us to test whether models can still track an object's state, position, and interactions in spite of visual interference and camera discontinuities.

VOST and VidOSC together offer a robust spectrum of challenges. The combination of both ensures our benchmark captures the breadth and depth of dynamic visual understanding.

2.4. Question Generation Pipeline

To generate high-quality, visually grounded questions, we developed a three-stage protocol: (1) manually writing questions (2) crafting detailed natural language descriptions of each video, and (3) prompting GPT-4o to generate structured questions similar to our manually written ones, by using those descriptions. The goal was to ensure that all questions were explicitly grounded in the visual content of the video. We focused on writing questions that could not be inferred from language priors alone.

2.4.1 Seed Question Design

First, we manually wrote **120 seed questions** across 31 videos (26 from VOST and 5 from VidOSC). Each seed question was tagged with one of our four visual reasoning categories:**State Recognition (st)**, **Temporal Reasoning (temp)**, **Counting and Enumeration (cnt)**, and **Spatial Understanding (spat)**.

These examples were carefully curated. We ensured that we chose a diverse selection of videos and asked a range of questions for each of our categories.

2.4.2 Manual Description Writing

Next, with the goal of automating the rest of our question generation, we wrote rich descriptions for each of our selected videos.

Example Video Description

Video Description:
9818_sp.mp4 In the first frame someone is in front of a tin pan of two halves of long green papayas. One of the papayas is in their left hand. They use a knife in their right hand to scrape the innards out of the papaya in their hand. They scrape from the top to the middle, then they scrape from the top to the bottom, then it falls out. Then they scrape from the bottom to the side, to get all of the innards out of the papaya. They use their right thumb and pointer finger to get the last bit of papaya out. They bring a pot that is to the north of the pan the papayas are in, they use their right hand to pull it around another bowl closer to the tin pan. They use the knife in their right hand to cut the papaya horizontally. They slice it half way and then pull both pieces apart, one side in each hand. They then use a knife to slice away the little piece connected and the top half (half of the original half) falls into the pot. They pick up that top half that fell into the bowl with their right hand and put both pieces into the pot they pulled over. They then pick up the other long half that was in the tin pan that still has innards. They put it in their left hand and had a scraper in their right hand. They scrape out all the innards in two top to bottom scraping motions. They start to cut this in half horizontally over the pot in front of the pan. The top half falls into the pot and they place the other half into the pot.

These descriptions (exemplified in the above description) served as the primary input for question generation and were crafted with careful attention to four key properties: **Action-level granularity**: Each action or manipulation was individually named and temporally ordered. **Object state transitions:** We explicitly described each object transformations (e.g., *"the onion is first peeled, then diced and placed into a*

pan"). **Spatial positioning**: Descriptions included the positions and movements of objects (e.g., *"the cup is initially to the left of the sink, then placed behind the cutting board"*). For spatial positioning we also included cardinal direction and perspective references (e.g. "*east to the perspective of the person slicing the apple*"). **Counting and timing**: Repetitive and discrete actions (e.g., *"the person flips the pancake 3 times"*) were explicitly enumerated to support numerical question generation.

On average, each description contained **163.93 words**, often structured as a paragraph detailing a list of events. For the VidOSC dataset, when there was a jump cut we included the tag "[transition]". This level of detail ensured that GPT-4o would be grounded in the actual visual content and generate questions very similar to the structure, difficulty, and diversity of the manually.

2.4.3 GPT-4o Generation Pipeline

Once we had our seed questions and our rich video descriptions, we used both of these and ChatGPT-4o to generate our remaining questions. We first asked GPT to analyze our 100 VOST questions. We then directed it on the structure and nature of the video description it would receive and that it needed to produce questions like the previous 100. We then produced 3-6 questions for each selected video from the VOST dataset. This process was repeated for the VidOSC dataset.

This pipeline produced a total of **397 manually verified questions**: **317 questions over 74 videos for VOST** and **80 questions over 20 videos for VidOSC**.

2.5. Verification

All generated questions were manually verified for correctness, diversity, adequate difficulty, and coverage across the four reasoning categories. The result is a benchmark that isolates visual reasoning failure modes with high precision, enabling robust diagnostic evaluations of multimodal model capabilities.

2.6. Benchmark Format and Question Structure

Each question in our benchmark is associated with a single video clip and designed to probe one of our four core categories of visual reasoning. Questions are presented in one of two formats:

- **Multiple-Choice Questions** – These are the majority of the questions, consisting of a prompt and 2-4 answer options labeled (A) through (D). Exactly one answer is correct, and the other three serve as plausible or implausible answers.
- **Numerical Questions** – These require the model to answer with an integer value, in response to a counting or enumeration question (e.g., "How many times is the bag folded in the video?"). No answer choices are provided for this question.

For each question we include the following metadata:

- `video_path` – The location of the video file.
- `question` – The question posed to the model.
- `question_type` – `mc` or `numerical`.
- `options` – A list of choices (if applicable).
- `ground_truth` – The correct answer.
- `question_class` – One of `st`, `temp`, `cnt`, `spat`.

This structure supports both automatic evaluation and fine-grained error analysis.

2.7. Model Evaluation Protocol

We evaluate each model using the same protocol, with slight differences to adapt each model to our benchmark. For numerical questions, we evaluate direct answer generation.

2.7.1 Evaluation Setting

For our quantitative experiments all models were tested in a zero-shot setting:

Each model processes video input (mp4) and a text prompt formatted as:

```
Analyze the following video to
answer the question.
<QUESTION>
Options: (A) ..., (B) ..., (C)
..., (D) ... <POST PROMPT>
```

To encourage our models to return answers in the correct manner, we add **post prompts** such as "*Respond with a single digit value."* or "*Only respond with a single letter: A, B, C, or D."*

2.7.2 Accuracy and Category-wise Scoring

We compute accuracy as the proportion of correctly answered questions over the total. For MCQs, a prediction is considered correct if the model's output exactly matches the ground truth option. For numerical questions, we accept integer matches only.

To facilitate diagnostic insights, we report accuracy by question category. This categorization allows us to assess strengths and weaknesses of each model class, revealing trends across reasoning types (e.g., a

model may perform well on state recognition but fail at spatial tracking).

All predictions, along with metadata (ground truth, model output, is_correct flag, question type), are logged in TSV format for further analysis. These files enable post hoc construction of an error taxonomy and facilitate targeted qualitative inspection.

3. Experimental Design

3.1. Model Selection and Evaluation Setup

To assess the capabilities of modern MLLMs on the DynaVis benchmark, we evaluate a diverse set of multimodal models—each varying in architecture, pretraining data, and integration of visual and language modalities. The selected models include:

- **Qwen2.5-VL (7B, 0.5B)** [1]: An open-source vision-language model based on Qwen2.5 with a language backbone and integrated visual encoder.
- **CogVLM2** [8]: A family of vision-language models designed for image and video understanding, featuring high-resolution input support and automated temporal grounding data construction. *For our evaluation we use their video captioning model**
- **InternVideo (V1 and V2)** [3, 20]: General video foundation models utilizing both generative and discriminative learning approaches to achieve state-of-the-art performance on various video understanding tasks.
- **Vamba-Qwen2-VL-7B** [16]: A hybrid Mamba-Transformer model that leverages cross-attention layers and Mamba-2 blocks for efficient hour-long video understanding.
- **LLaVA-OneVision (7B, 0.5B)** [9]: A specialized variant of LLaVA designed to handle frame-by-frame video inputs with improved alignment between vision and language components.
- **Gemini 2.5 Pro Experimental (03-25)** [7]: Google's most advanced AI model to date, featuring enhanced reasoning and coding capabilities, native multimodality, and a 1 million token context window.

Evaluation Protocol. Each model is tested under a zero-shot setting using the same TSV file question interface. For each row in the benchmark, a model receives the video and a formatted prompt containing the question. The prompt includes either multiple-choice options (e.g., "(A) ..., (B) ...") or requests a numerical answer (e.g., "Respond with a single number.").

Code Implementation. All evaluations are run using custom inference scripts based on PyTorch and HuggingFace Transformers. For example, LLaVA-OneVision is evaluated using a prompt pipeline that converts each TSV row into a multimodal chat format. Videos are encoded as sequences of frames and processed with vision encoders. The resulting predictions are parsed and compared to ground truth for accuracy scoring. The evaluation pipeline computes per-category accuracy, and logs each result with model outputs, extracted answers, and correctness labels.

3.2. Evaluation Metrics

We report three types of evaluation metrics:

- **Overall Accuracy**: The percentage of correctly answered questions.
- **Per-Category Accuracy**: Accuracy on the four core question classes:
  - *st* – Object State Recognition
  - *temp* – Temporal Reasoning
  - *spat* – Spatial Localization
  - *cnt* – Counting and Enumeration
- **Split-by-Dataset Accuracy**: To better reflect dataset characteristics, we report accuracy separately for VOST (egocentric, high-resolution) and VidOSC (long-form, third-person with visual clutter).

All model predictions were normalized prior to scoring, and multi-choice predictions were reduced to their corresponding letter index (e.g., "B"), while numeric responses were cast as integers. Accuracy was computed by exact match between model response and ground truth label.

3.3. Main Results

We evaluated all models across both datasets. Tables 1 and 2 report the results for VOST and VidOSC, respectively. Models like **Qwen2.5-VL-7B** and **InternVideo** show relatively strong performance on certain categories, but every model struggled on counting questions and in most cases spatial localization. Notably, **InternVideo2** performed significantly worse than its predecessor, indicating variability across model series.

Highlights:

- Qwen2.5-VL-7B and InternVideo lead on overall accuracy on both video sources.
- State recognition (st) results are generally high for some models. These questions are the closest to general summaries of the video, suggesting models at least know the general structure of the video.
- State recognition (st) results are higher in VOST than VidOSC for the majority of models. This is po-

Table 1. Model Accuracy on the **VOST Benchmark** (417 Questions)

Model	Total Accuracy	State (st)	Temporal (temp)	Spatial (spat)	Counting (cnt)
Qwen2.5-VL-7B	50.84% (212/417)	68.60%	52.04%	46.39%	32.67%
InternVideo	50.84% (212/417)	71.90%	53.06%	51.55%	22.77%
InternVideo2	37.17% (155/417)	65.29%	31.63%	34.02%	11.88%
LLaVA-Video-7B	36.93% (154/417)	35.54%	33.67%	44.33%	34.65%
LLaVA-OneVision-7B	41.9% (175/417)	56.67%	48.98%	38.14%	21.78%
LLaVA-OneVision-0.5B	36.21% (151/417)	60.33%	24.49%	38.14%	16.83%
CogVLM2 (captioning)	28.67% (119/415)	46.28%	32.65%	31.96%	0.00%
Vamba	41.25% (172/417)	55.37%	32.65%	41.24%	32.67%

Table 2. Model Accuracy on the **VidOSC Benchmark** (100 Questions)

Model	Total Accuracy	State (st)	Temporal (temp)	Spatial (spat)	Counting (cnt)
Qwen2.5-VL-7B	60.00% (60/100)	60.38%	80.00%	71.43%	20.00%
InternVideo	64.00% (64/100)	73.58%	64.00%	57.14%	33.33%
InternVideo2	39.00% (39/100)	47.17%	48.00%	14.29%	6.67%
LLaVA-Video-7B	38.00% (38/100)	37.74%	40.00%	85.71%	13.33%
LLaVA-OneVision-7B	37.00% (37/100)	43.40%	40.00%	42.86%	6.67%
LLaVA-OneVision-0.5B	38.00% (38/100)	43.40%	44.00%	28.57%	13.33%
CogVLM2 (captioning)	30.00% (30/100)	37.74%	40.00%	0.00%	0.00%
Vamba	34.00% (34/100)	43.40%	24.00%	14.29%	26.67%

tentially the result of the continuity and shorter duration of the VOST videos.

- Counting (*cnt*) questions remain challenging for all models.
- Temporal (temp) and Spatial (spat) understanding scores are intermixed. Generally below 54% with most scores below 50%. There are a few high scores for this category on the VidOSC dataset.

These results highlight weaknesses in many categories. In the next sections, we analyze these dimensions in greater detail through open-ended tasks and error taxonomies.

4. Error Taxonomy and Failure Analysis

4.1. Motivation for Error Analysis

Our overall accuracy results provide a good outlook on model performance, but in this section we want to understand the failure modes of each model on a deeper level. In video understanding, errors can arise from many sources: misinterpreting object movements, losing track of actions, hallucinating actions or events, or confabulating typical actions. As such, we perform a systematic error taxonomy on a sample of 50 common incorrect responses from our models. Our goal is to characterize the nature of model failures and identify patterns that may inform future architectural or training improvements.

4.2. Taxonomy of Common Errors

Based on manual review and annotation, we identify six primary categories of model failure:

- **Spatial Mislocalization** — The model incorrectly describes or predicts where an object is located relative to others or to the environment. For instance, answering that an item is placed behind an object when it was clearly to the left from the camera's point of view. Models can also fail on perspective taking in third person videos (e.g. To what direction did the chef place down the spatula, from his point of view).
- **Temporal Misordering** — The model fails to correctly recall the sequence of actions or state transitions. In this case models often state an action occurred first or last in a video when there was another action that occurred before or after that.
- **Undercounting/Overcounting** — The model provides an incorrect numerical estimate for the number of actions, objects, or repetitions. These errors often stem from poor object tracking or insufficient temporal attention. This has been long noted as a difficult task for multimodal models.
- **Object Identity Confusion** — The model confuses different objects, such as mistaking a cutting board for a countertop, or corn for mushrooms. This typically indicates shallow visual perception.
- **Hallucination** — The model invents events or ac-

tions that do not appear in the video. For instance, stating that an object was placed in a trashcan when a trashcan may not even appear in the video.

- **Confabulation** — This is similar to hallucination, but instead the model provides plausible but incorrect inferences based on visual content. We hypothesize this could be from an over-reliance on language priors. Unlike hallucination, which introduces wholly fabricated content, confabulation tends to insert semantically related but unfounded actions (e.g., saying "the onion was peeled" when it was only sliced).

The multiple choice answers in our benchmark allowed us to gain insight into what type of errors occurred per question. Many of these failures are supported by our open-ended analysis later in the paper.

4.3. Error Distribution Across Models

To better understand the frequency of different error types across our evaluated models, we conducted a manual annotation and analysis of 50 incorrect responses sampled from our models. Each error was assigned to one of the categories from our taxonomy (refer to Section 4.2).

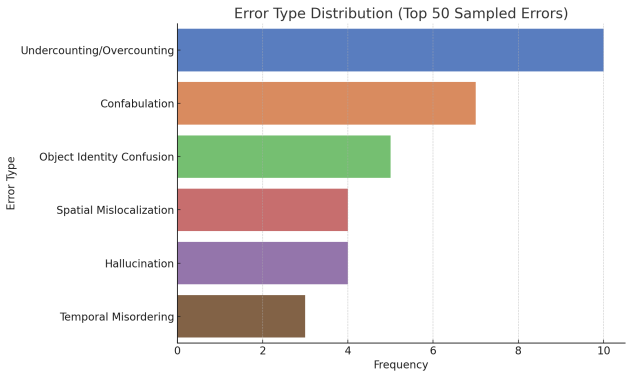

Figure 2. **Frequency of error types across 50 sampled model failures.** Counting errors and Confabulation were among the most common errors, highlighting the challenge of quantitative and spatial reasoning in video-based state understanding.

As shown in Figure 2, **Undercounting** / **Overcounting** and **Confabulation** were the most frequent error types, each accounting for over 20% of observed mistakes. Models frequently underestimated repeated actions (e.g., slicing steps). With **Confabulation** and **Hallucination**, models inventing plausible but ungrounded events—especially when objects exited the frame. With confabulation specifically, models would typically choose answers that were likely to happen or may typically happen, but never occur in the video. This is emphasized in our open-ended experiments later in the paper. **Object Identity Confusion** is our third most common failure mode. Typically in these scenarios models are blind to objects, or mistake certain objects for others. For example, as seen in the image in Figure 3, one of our models mistakes the bowls as pans, and is mislead into an incorrect answer choice. **Spatial Mislocalization** often involves models mistaking left from right. This also includes struggles with spatial relations (e.g. the bowl is on the right side of the knife next to the cutting board).

Figure 3. **Object Identity Confusion Example** One of our models mistakes bowls for pans.

4.4. Qualitative Examples

To further illustrate model failure modes, we present selected qualitative examples showcasing typical reasoning errors.

Example: Confabulation

Video: `pm0_pp.mp4`
Question: What happens to the potato skins after they are peeled?
Model Response: "They are placed in a bowl."
Ground Truth: The skins fall onto the cutting board.
Analysis: The model fabricates a plausible but ungrounded event – possibly based on prior knowledge of cooking routines. No visual evidence in the video supports this claim.

Error Type	Definition	Example Failure
Spatial Mislocalization	Incorrectly predicting where an object is placed or moved relative to the scene or other objects.	The model says the tupperware is "behind the cutting board," when it is clearly to the right from the egocentric camera's perspective.
Temporal Misordering	Failing to identify the correct sequence of events or state transitions.	Predicting that icing was applied before the cake was sliced, when in fact the slicing happened first.
Undercounting / Overcounting	Providing an incorrect count of repeated actions or objects.	Predicting 2 tomato slices were made when the video shows 4.
Object Identity Confusion	Misidentifying visually distinct but semantically similar objects.	Confusing a cutting board for the counter top in a cooking video.
Hallucination	Fabricating events or interactions that do not occur in the video.	Saying "the napkin was placed in the trash" when no trashcan appears in the scene.
Confabulation	Blending visual priors with incomplete evidence to produce plausible but incorrect answers.	Stating that "the onion is peeled" because it is sliced, even though no peeling action occurred.

Table 3. **Color-coded Error Taxonomy of Model Failures in DynaVis.** This table outlines six primary failure types observed across sampled model responses. Each row illustrates a distinct class of visual reasoning error with accompanying definitions and real failure examples.

Example: Temporal Misordering

Video: `8di_cc.mp4`
Question: What is the final decoration applied to the cake?
Model Response: "Piping icing decorations."
Ground Truth: "Adding fruit toppings."
Analysis: The model confuses the order of visual actions in this case, exemplifying a temporal sequencing error.

Our models' reliance on priors rather than visual events and inability to interpret small but critical changes result in incorrect predictions.

4.5. Takeaways from Taxonomy

Across all evaluations, our models exhibit limitations when asked to engage with fine-grained, temporally grounded, and spatially precise reasoning over real-world video. While some models demonstrate moderate competence in object state recognition, no model achieves reliable performance across all four dimensions of the DynaVis Benchmark.

Our error taxonomy presents potential sources of these failures, each detailing gaps in visual understanding. **Spatial mislocalization** and **temporal misordering** errors indicate that models lack a persistent understanding of the relative location of objects, the directions in which they move and the sequence of actions. **Undercounting**/**Overcounting**, meanwhile, points to a struggle of AI models to count, not possessing the ability to tally repeated actions or objects. **Confabulation** and **hallucination** reflect how models sometimes invent contextually plausible but visually ungrounded responses, especially in open-ended or ambiguous scenes.

5. Additional Experiments

5.1. Diagnosing Visual Hallucination with Open-Ended State Prompts

Beyond multiple-choice evaluations, we introduced an open-ended question answering experiment designed to better understand the visual grounding abilities of MLLMs. We aimed to probe whether these models can faithfully perceive and describe object state changes in video without the aid of structured options. In this experiment, we targeted ten videos from our benchmark – five where Qwen2.5-VL-7B correctly answered a state recognition question, and five where

Table 4. Examples of visual hallucination and omission in Gemini 2.5 Pro responses to open-ended prompts.

Video File	Ground Truth Event	Gemini Description	Error Type
5259_dm.mp4	Panel removed and bagged	Multiple panels removed, no bag mentioned	Confabulation
1082_ct.MP4	Tomato slid into bowl	Salad tossed after dicing	Hallucination
1897_cc.mp4	Napkin torn into quarters	Napkin crumpled into a ball	Visual Misperception
303_fp.mp4	Object disposed of sideways	Direction of motion missed	Omission
1961_eb.MP4	Bag emptied and trashed	Bag emptied and trashed	Correct

it was incorrect. For each video, we prompted both Qwen and Gemini 2.5 Pro Experimental (03-25) with the simple prompt:

"Describe the video in detail."

This prompt removed any guiding context provided by multiple-choice options, forcing the models describe the video off their own assessment. Responses were then manually compared against true visual elements of each video, with particular attention paid to hallucinations, omissions, and overgeneralizations. The same process was applied to Gemini to examine whether more advanced proprietary systems suffer from similar limitations.

Figure 4. **One of the most common failure cases across all models.** Most of the models did not recognize that a panel taken off a motor cycle then placed in a clear plastic bag.

Consistent Hallucinations Across Models. Both models struggled to visually describe scenes accurately in this unconstrained setting. Despite Gemini's scale advantage – hundreds of billions of parameters – and fluency, it repeatedly made perceptual errors similar to Qwen. In one video, a single scooter panel is removed and placed in a plastic bag. Qwen ignored the bag entirely, while Gemini hallucinated the removal of multiple black panels and failed to identify the bagging step. When later asked the associated multiple-choice question, both models answered this question incorrectly, linking descriptive hallucination to benchmark failure. **This example is shown in Figure 4.**

Substituting Likely Events for Observed Ones. Another phenomena our models suffered from confabulation: substituting likely events for observed ones. In another video, a napkin torn into quarters before behind thrown into the trash. Instead Gemini described the napkin as being "crumpled into a ball." Observing multiple occurrences of this kind from **both** Gemini and Qwen, we saw a clear tendency to generate plausible but untrue events based on commonalities in the real world. We posit, when models encountered ambiguous or incomplete visuals, they defaulted to highly probable completions learned from their language priors.

Takeaway While multiple-choice questions can obscure perceptual errors, open-ended prompts force a model to reveal what it actually "sees." Overall, Gemini was more accurate than Qwen and our other smaller models on this diagnostic task. Gemini would analyze many videos with much more visual detail. But many times these descriptions would fall short of being truthful. The fluency would often mask incorrect visual analysis. Across all ten videos, whenever Gemini hallucinated a visual event or missed a key transformation, it also failed the corresponding state recognition question. This duality between descriptive error and benchmark inaccuracy reinforces the importance of grounded perception in the DynaVis Benchmark.

Table 4 summarizes the observed hallucination patterns and how they manifest in representative videos.

5.2. Spatial Consistency Task

As an additional experiment, we wanted to probe whether models understand the relative spatial arrangement of objects over time in a video. Many of the failure cases for spatial understanding stem from an inability to track objects spatially across frames. The degree to which a model can understand the movement (or lack thereof) of an object can elucidate their spatiotemporal grounding abilities.

5.2.1 5.2.1 Multiple Choice Analysis by Spatial Category

To gain deeper insight into our models' spatial understanding, we developed another focused diagnostic task. We manually curated 18 targeted spatial consistency multiple choice questions, along with 12 open ended questions. To quantify model performance across different types of spatial reasoning, we categorized multiple-choice questions into four tags:

- **Containment / Enclosure (`cont`)**: Requires recognition that an object was placed inside or within another object (e.g., a bag or a trash can).
- **Final Location (`fl`)**: Questions about the object's final spatial position, independent of the full trajectory.
- **Relative Position (`rp`)**: Questions about how objects are positioned in relation to others (e.g., to the left of, behind, on top of).
- **Trajectory / Motion (`traj`)**: Questions involving tracking an object's path or change in location across the video.

We evaluated four models on this benchmark. Table 5 reports per-category accuracies.

Observations. Qwen2.5-VL-7B showed strong performance in containment (100%) and relative position reasoning (60%), but completely failed on trajectory questions, showing its struggle with spatial tracking. InternVideo generally had highest accuracy overall, particularly on final location and containment, and trajectory questions, indicating better robustness in following spatial motion. LLaVA-Video and LLaVA-OneVision both struggled with final location questions (29%), suggesting difficulty in retaining spatial details across the course of a video. Notably, LLaVA-Video failed all containment questions, while LLaVA-OneVision succeeded in both.

This breakdown shows that spatial reasoning is not a monolithic capability. Success in one of our spatial categories does not imply competence in others. For example, Qwen is strong in local spatial recognition but weak at tracking, while InternVideo is more balanced.

5.2.2 Open-Ended Spatial Evaluation

For our spatial consistency task, we also went beyond multiple choice questions with 12 targeted open-ended questions. We manually analyzed open-ended responses from three smaller multimodal models: **Qwen2.5-VL-7B**, **InternVideo-8B**, and **LLaVA-Video-7B**. We also ran these open-ended prompts through Gemini 2.5 Pro Experimental (03-25) – this is shown in the next section. An example of one of the spatially oriented open-ended questions is: "Describe the movement of the tomatoes across the video from start to end."

InternVideo-8B was the most visually detailed, often outperforming Qwen in describing directionality, containment, and action continuity. However, this detail sometimes came at the cost of precision, with frequent hallucinations such as fabricating a recycling bin or misreporting the movement of specific object. LLaVA-Video-7B, by contrast, struggled across the board: its responses leaned heavily on generic visual clichés, like "moved to the center of the table," and often misrepresented or fabricated events entirely. Across models, the most common failure modes included the loss of spatial specificity and inconsistent object tracking. These findings reinforce the benchmark's value in surfacing gaps in visual grounding.

5.2.3 Gemini 2.5 Open-Ended Analysis

To complement the earlier evaluations of smaller open-source models, we conducted a focused analysis of Gemini 2.5 Pro's spatial reasoning capabilities. Given its scale and performance across general vision-language benchmarks, we wanted to determine if Gemini could overcome the spatial consistency errors observed in the smaller models we evaluated.

Each prompt asked Gemini to describe the movement or spatial location of an object, and its free-form outputs were evaluated for visual correctness, precision of spatial language, and overall coherence. Follow-up prompts were used in some cases to test Gemini's ability to self-correct or revise earlier outputs based on user feedback.

Table 5. Accuracy by spatial subcategory. Each model is evaluated on the same 18 multiple-choice spatial questions, grouped into containment (`cont`), final location (`fl`), relative position (`rp`), and trajectory (`traj`).

Model	Containment (`cont`)	Final Location (`fl`)	Relative Position (`rp`)	Trajectory (`traj`)
Qwen2.5-VL-7B	1.00 (2/2)	0.57 (4/7)	0.60 (3/5)	0.00 (0/4)
InternVideo-8B	1.00 (2/2)	0.71 (5/7)	0.40 (2/5)	0.50 (2/4)
LLaVA-Video-7B	0.00 (0/2)	0.29 (2/7)	0.40 (2/5)	0.25 (1/4)
LLaVA-OneVision-7B	1.00 (2/2)	0.29 (2/7)	0.20 (1/5)	0.50 (2/4)

Gemini Example: Visual Misperception and Inference Failure

Video: `3568_sh.mp4`
Question: Describe the spatial location(s) of the knife before it is used to spread hummus.
Model Response: "Before the person picks up the knife to spread the hummus, it is resting on the white plate located to the left of the plate holding the two slices of bread."
Ground Truth: The knife begins in the camera wearer's hand and is only later placed on the plate.
Analysis: Even after multiple follow-up prompts, Gemini failed to revise its initial misperception. This reveals an inability to attend to preconditions and transitional frames.

Gemini's open-ended responses were consistently more fluent and detailed than those of smaller models, but this linguistic strength often masked fundamental failures in visual grounding. Errors ranged from miscounting and misidentifying object states, to hallucinating transitions or reversing action sequences. Even when presented with prompts questioning its visual errors, Gemini tended to rationalize rather than revise its outputs, suggesting a reliance on narrative plausibility over grounded perception.

Tracking of Low-Salience Objects Remains Challenging In some of our failure cases, Gemini did a poor job of tracking objects that were not the object of interest. In one instance, Gemini was asked to track the location of a jar of jam throughout a short kitchen sequence. Although the jar remains visible for much of the video, the model incorrectly stated that the jar stayed in place, missing its lateral movement to the left of the cutting board. This was a prime example among others that showed Gemini's tendency to sometimes ignore background or low-movement elements.

5.3. Counting as Multiple Choice

While prior sections focused on spatial reasoning and grounded object tracking, one of the hardest challenges across all models remained counting. In the original benchmark, open-ended numerical questions such as "How many pieces of paper are used in the video?" yielded consistently low scores, even among stronger models. We wanted to investigate whether the low performance was due to the numerical questions being open ended generation questions. If so, converting them into multiple-choice questions might provide cues that help models anchor their predictions.

Method. To evaluate this, we created a new version of the counting questions where the open-ended numerical prompts were transformed into multiple-choice questions with four numerical options. The other answer choices were selected through the following formula:

- **For small values** (1–5), distractors were kept within ± 2 of the correct answer.
- **For medium values** (6–15), distractors varied by ± 3–5.
- **For higher values** (20+), distractors spanned ± 5–10.

Distractors were randomized in position.

Results. Table 6 shows MCQ accuracy across models and compares it with their original open-ended performance.

Interpretation. Results indicate that simply converting numerical queries into MCQs does garner some significant improvement. Qwen2.5-VL-7B, Vamba, and LLaVa-OneVision-0.5B demonstrated a notable gain (+13.33%, +12.33% +20.17%). LLaVA-Video, suffered from a significant accuracy decline (-8.65%), and LLaVA-OneVision-7B showed minimal change. This suggests that access to answer choices may support models that already possess some grounding ability but can confuse models that overly rely on pattern completion or language priors.

These findings support a dual interpretation:

- *Support :* MCQs offer a constrained space in which

Table 6. Comparison of model accuracy on **counting questions** before and after converting open-ended questions into multiple choice format (100 questions).

Model	Original (Open-Ended)	Multiple Choice (Converted)	Change
Qwen2.5-VL-7B	32.67%	46.00%	+13.33%
LLaVA-Video-7B	34.65%	26.00%	–8.65%
LLaVA-OneVision-7B	21.78%	25.00%	+3.22%
LLaVA-OneVision-0.5B	*16.83%*	37.00%	+20.17%
Vamba	32.67%	45.00%	+12.33%

answers near the value of the true answer can serve as scaffolds for partial understanding, allowing models to reason more effectively.

- *Distractor:* On the other hand, choices may act as decoys, reinforcing false visual hypotheses rather than aiding retrieval.

Although converting counting questions to multiple choice may help for many of our models, counting is still a significant challenge. The accuracy does not eclipse 50% for any models. Thus all models still struggle with counting, even with answer choices. While MCQ scaffolding helps in some cases, it does not universally enhance performance.

6. Discussion

This thesis presents a benchmark focused on reasoning about dynamic changes over video—a longstanding challenge for multimodal models in AI. The results point to several critical insights:

6.1. Spatial Reasoning

In Section 5.2, we introduced a taxonomy of spatial reasoning types—`containment`, `final location`, `relative position`, and `trajectory/motion`. Results show that most of our models struggled considerably with trajectory and relative position. Across the models, they failed to perform consistently across all types.

This indicates that spatial reasoning in the MLLMs we tested is inconsistent. In particular, models often confuse similar locations (e.g., “on the counter” vs. “on the cutting board”), misrepresent directionality (e.g., “to the left” vs. “to the right”), or ignore small movements of objects that are not the main focus of the video.

6.2. Open-ended Generation Reveals Deeper Failures

In two of our experiments we turned to open ended prompting to probe model performance without the assistance of multiple choice options. Across multiple questions, Qwen and InternVideo failed to detect fine-grained object transitions, often summarizing with vague phrases like “moved away” or omitting directional movement entirely. In our open ended state recognition experiment, we also saw Qwen deal with visual perception errors. In particular, it would often omit apparent visual events, especially near the beginning or end of the video.

We also saw illuminating results from Gemini 2.5 Pro, a proprietary model known for high-quality generation. It generally performed better than all of our smaller models, often describing videos with more specificity, enumerating more objects and more actions. It was also better in regards to spatial understanding. But, on the other hand, Gemini fell into many of the same pitfalls as our smaller models. Aside its good performance, Gemini also frequently hallucinated or omitted key visual events. This was seen in both the open ended state recognition and spatial understanding tasks. In some cases, it generated plausible but false scenes. These errors suggest a breakdown in visual grounding: the models are not describing what they see, but rather responding with what is likely to fit a scenario.

Open-ended prompts served as a powerful diagnostic tool to supplement our benchmark experimentation.

6.3. A Diagnostic Benchmark for Dynamic Understanding

Many video benchmarks focus on ”understanding” for long form video which often includes retrieval of specific events that may be hidden in hours of video. This benchmark was explicitly designed to isolate object dynamics over time. Unlike some prior datasets that focus on static object recognition or spatial reasoning on synthetic data, this benchmark emphasizes realistic object changes in natural videos. Recognizing that an object has moved from the counter to the trash, or that a tomato was sliced before being placed in a

bowl, are not trivial feats—they require persistent internal representations of objects and their properties. This remains an active challenge.

By organizing results into interpretable categories, we highlight where our models struggle. Notably, the failure of models to track objects' spatial trajectories. Also, open-ended generation tasks suggests that they lack persistent memory and robust object representations across frames.

6.4. Limitations and Future Work

This work is not without limitations. The main drawback is scale. For the benchmark, the video and question sample size is small (125 videos and 517 questions). We felt that this was enough to get true quantitative and qualitative results, but in terms of data coverage, there are many more visual scenarios that can be covered. Our videos primarily feature cooking scenarios. Although the cooking setting is frequent with object changes, there are other settings to consider such as office spaces, movies, metropolitan areas, etc. Scaling this benchmark would allow for more robust quantitative conclusions and cross-model comparisons. Additionally, in regards to scale, the models we chose quantitatively evaluated were comparatively small. All of our LLMs had a language model component that was 8 billion parameters or lower. To get an exhaustive grasp of trends across MLLMs, we would need to also test models of sizes 32B, 72B, and above. In hopes to circumvent this, we compared our smaller models against a very large model: Gemini 2.5 pro. We found that this model fell into many of the same pitfalls as our smaller models.

Additionally, the open-ended evaluations were qualitative and manually analyzed. Future work could explore automatic methods for comparing generated text to reference answers.

Lastly, there remains much room to improve upon the models we tested. There is an opportunity to finetune models or propose new architectural or training data implementations that could improve MLLMs' ability to understand object dynamics better in video.

6.5. Implications

The results suggest a broader implication for the development of multimodal systems: the MLLMs we tested are limited in their ability to track object dynamics over time. They lack grounding, persistent memory, and robust spatial understanding. This has consequences for downstream applications like robotics, navigation, and physical reasoning.

A truly "embodied" or perceptually-grounded AI system would need to succeed not just at classification, description, or retrieval, but at understanding how objects move, change, and interact. This benchmark begins to measure that capability.

7. Related Work

We discuss prior work in two key areas:

Video Benchmarks for MLLMs. Prior video benchmarks often focus on long-term event understanding, temporal localization, and content summarization. HourVideo [2] and MMBench-Video [5] provide long-form benchmarks evaluating summarization, perception, and reasoning. Open-ended egocentric question answering benchmarks such as QaEgo4Dv2 [15] and EgoSchema [12] assess MLLMs' capacity for understanding first-person perspectives. We build on this work by emphasizing fine grained analysis. Our work isolates individual aspects of visual intelligence, revealing failure cases in both open-ended and structured tasks. While our setting is narrower than that of long-form or general QA datasets, it provides a precise probe into the visual grounding abilities required for embodied AI and interactive systems.

Spatial Reasoning in MLLMs. Recent studies have explored the visual-spatial capabilities of MLLMs trained on web-scale video and image-text data. While models such as Qwen2-VL [19], Gemini [4], and LLaVA [11] show strong performance in general vision-language tasks, but their spatial consistency and visual grounding capabilites remain to be fully understood. VSI-Bench [22] and SpatialEval [18] evaluate MLLMs on visual-spatial reasoning from images or videos. Other works, including BLINK [6], highlight that even seemingly simple spatial tasks remain challenging for state-of-the-art models. Along with these efforts, our benchmark emphasizes dynamic, object-centric transformations in real-world videos, pushing models to track how objects move, interact, and change over time.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 1, 2, 6

[2] Keshigeyan Chandrasegaran, Agrim Gupta, Lea M. Hadzic, Taran Kota, Jimming He, Cristobal Eyzaguirre,

Zane Durante, Manling Li, Jiajun Wu, and Fei-Fei Li. Hourvideo: 1-hour video-language understanding. In *Advances in Neural Information Processing Systems*, 2024. 1, 14

[3] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024. 2, 6

[4] Google DeepMind. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 14

[5] 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. *arXiv preprint arXiv:2406.14515*, 2024. 14

[6] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. *arXiv preprint arXiv:2404.12390*, 2024. 14

[7] Google DeepMind. Gemini 2.5 pro: Our most intelligent ai model. https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/, 2025. Accessed: 2025-04-15. 6

[8] Wenyi Hong, Weihan Wang, Ming Ding, Wenmeng Yu, Qingsong Lv, Yan Wang, Yean Cheng, Shiyu Huang, Junhui Ji, Zhao Xue, et al. Cogvlm2: Visual language models for image and video understanding. *arXiv preprint arXiv:2408.16500*, 2024. 6

[9] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer, 2024. 6

[10] 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

[11] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 14

[12] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding, 2023. 14

[13] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. *CoRR*, abs/1906.03327, 2019. 4

[14] Kaleb Newman, Shijie Wang, Yuan Zang, David Heffren, and Chen Sun. Do Pre-trained Vision-Language Models Encode Object States? *arXiv preprint arXiv:2409.10488*, 2024. 1

[15] Alkesh Patel, Vibhav Chitalia, and Yinfei Yang. Advancing egocentric video question answering with multimodal large language models. *arXiv preprint arXiv:2504.04550*, 2025. 14

[16] Weiming Ren, Wentao Ma, Huan Yang, Cong Wei, Ge Zhang, and Wenhu Chen. Vamba: Understanding hour-long videos with hybrid mamba-transformers, 2025. 2, 6

[17] Pavel Tokmakov, Jie Li, and Adrien Gaidon. Breaking the "object" in video object segmentation. In *CVPR*, 2023. 1, 3

[18] Jiayu Wang, Yifei Ming, Zhenmei Shi, Vibhav Vineet, Xin Wang, Yixuan Li, and Neel Joshi. Is a picture worth a thousand words? delving into spatial reasoning for vision language models, 2024. 14

[19] 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. 14

[20] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Chenting Wang, Guo Chen, Baoqi Pei, Rongkun Zheng, Jilan Xu, Zun Wang, et al. Internvideo2: Scaling video foundation models for multimodal video understanding. *arXiv preprint arXiv:2403.15377*, 2024. 2, 6

[21] Zihui Xue, Kumar Ashutosh, and Kristen Grauman. Learning object state changes in videos: An open-world perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18493–18503, 2024. 1, 3

[22] Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in Space: How Multimodal Large Language Models See, Remember and Recall Spaces. *arXiv preprint arXiv:2412.14171*, 2024. 1, 14