Learning to Deny: Action Denial in Multimodal Large Language Models

Raiyaan Abdullah, Shehreen Azad, and Yogesh Singh Rawat

Institute of Artificial Intelligence, University of Central Florida, USA
raiyaanabdullah@gmail.com, shehreen.azad@ucf.edu, yogesh@crcv.ucf.edu
Project page: https://raiyaan-abdullah.github.io/Learn-to-Deny-webpage

Abstract. Multimodal large language models (MLLMs) have rapidly advanced video understanding, achieving strong zero-shot and few-shot recognition across standard benchmarks. Yet their ability to **deny** an action by recognizing when an activity is **not** happening despite strong contextual cues remains largely unexplored. We introduce **UCF101-AD**, a large-scale benchmark consisting of paired *Action-Presence* and *Action-Denial* clips, designed to evaluate this capacity for denial. Each negative video in UCF101-AD preserves the same contextual and motion cues (persons, objects, locations) as its positive counterpart, but the defining action itself is explicitly absent. Evaluating 20 state-of-the-art MLLMs reveals a consistent failure: models that exceed 85% accuracy on the positive action classes collapse below 50% on its action-denial counterpart, indicating a strong inclination to affirm plausible actions rather than verify that they truly occur. This exposes a critical blind spot in modern video understanding: the inability to reason causally about whether a motion actually happens. To probe this issue, we explore a causal graph formulation, **CausalAct**, which expresses scene structure through natural-language prompts linking context, interaction, and motion. Incorporating such causal cues substantially reduces false positives, demonstrating that denial is a learnable reasoning skill. UCF101-AD provides a new lens for diagnosing and improving causal reasoning in multimodal models. Dataset and relevant code: https://github.com/raiyaan-abdullah/Learn-to-Deny.

Keywords: Benchmark Dataset · Action Denial · Causal Graph · MLLMs

1 Introduction

Recent Multimodal Large Language Models (MLLMs) [3,29,56,61,71,72] achieve strong results on standard action recognition benchmarks, but closer analysis shows that much of this performance can arise from spurious correlations rather than genuine temporal reasoning [1, 18, 19, 28, 40]. Instead of verifying the motion that defines an activity, models often rely on contextual cues such as background or object presence. This aligns with evidence that MLLMs display an agreeableness or affirmative bias, tending to endorse a premise rather than negate it [16, 51, 64]. In action recognition, this bias appears as predicting that

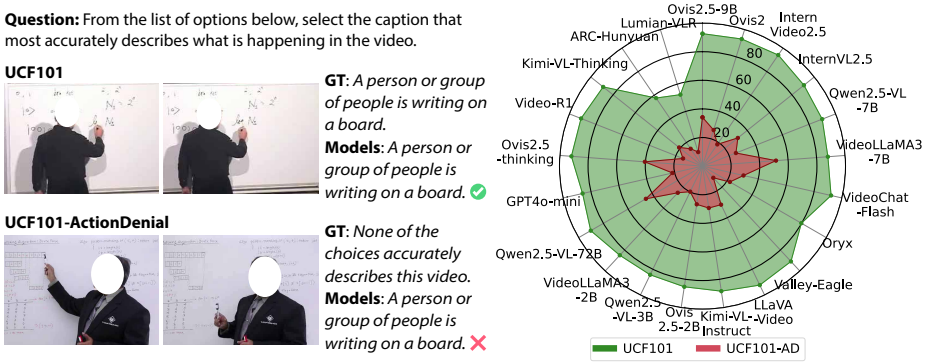

Fig. 1: ***Model failures in denying an action:*** *(left)* Models are capable of recognizing when the action occurs, but seeing a person holding a marker in front of a board but not writing, they still predict "Writing On Board". *(right)* Radar plot of accuracy on the original UCF101 test set (green) vs. the UCF101-AD *Action-Denial* videos (red). Most models exceed 85% on UCF101 but fall below 50% on videos where the action is absent, highlighting a large gap between action recognition and denial.

an action is occurring whenever context is suggestive, even if the defining motion is missing, leading to false positives and potential failures in settings like surveillance, autonomous systems, and sports analytics. We formalize this challenge as **action denial**: the ability of a model to determine that an action is not occurring when its defining motion is absent, despite compelling contextual evidence (Fig. 1, *left*). Robust video understanding therefore requires models to confirm actions through their defining motion and to reliably deny them when that motion is not observed.

Despite the scale and diversity of benchmarks such as Kinetics [6], UCF101 [66], HMDB51 [36], and ActivityNet [27], they primarily contain positive examples where the labeled action truly occurs and its defining motion is visible. They rarely include hard counterfactual negatives in which relevant objects, actors, or scenes are present but the action does not take place. This reinforces object-action correlations and allows models to succeed without explicitly verifying motion. Since models are neither trained nor evaluated on contextually similar yet motion-absent clips, current benchmarks offer limited insight into whether a system can separate genuine action evidence from misleading context and reliably deny an action when its defining motion is missing.

To address this gap and directly evaluate action denial, we introduce **UCF101-ActionDenial (UCF101-AD)**, a dataset of 11,283 videos built on UCF101 [66] action categories, with paired positive and negative classes for each action. Each negative clip preserves the typical context of its positive counterpart while ensuring the target action does **not** occur, and includes realistic motion that often co-occurs with the action. For instance, in "Not Basketball Dunk," players dribble on a court with a ball present, but no dunk happens. This controlled design

isolates defining motion as the key discriminative signal and provides a focused benchmark for evaluating action denial.

Our experiments reveal that the **capacity for denial**, a fundamental reasoning skill, is exceptionally challenging for multimodal action models and general-purpose MLLMs. We observe that several state-of-the-art architectures struggle to differentiate these negatives, most achieving performance below 50%, often confusing with the positive action (Fig. 1, *right*). This failure highlights that models are learning spurious correlations rather than the true causal relationship between interaction, motion and target action. They lack a reliable mechanism to verify if the defining interaction is present.

We hypothesize that models can more reliably decide whether an action is occurring, including its absence, if they are guided to reason about how contextual evidence (persons, objects, location) supports the defining motion. Existing debiasing methods that weaken correlations with background or object cues [10, 13, 18, 82] operate mainly in positive-action regimes and reward correct labels, not explicit denial when motion is missing. Robust denial instead calls for a structured representation that organizes scene components and checks whether the preconditions of an action are actually instantiated. Prior work on causal structure and spatio-temporal scene graphs shows that modeling dependencies among entities, relations, and events can improve robustness [8, 31, 44], but is typically used for prediction or explanation. In contrast, we ask whether an explicit structure can help models decide when to deny a target action despite strong contextual cues.

Motivated by this, we propose **CausalAct**, a causal framework that organizes scene components and their dependencies with a focus on verification. Building on graph-based prompting and structured scene descriptions that improve multimodal video reasoning [35, 42, 50], we express the causal graph in natural language and ask the models to check whether the necessary conditions for the target action hold. We find that this improves action-denial performance for models with strong language backbones, indicating that the language module must be able to parse such graph-style prompts. We further finetune smaller MLLMs on auxiliary graph-related VQA tasks that teach graph concepts without exposing target action labels, avoiding label leakage, and observe similar gains, showing that causal graphs can aid denial once the model learns to interpret them. Our contributions are:

- We introduce **UCF101-AD**, a benchmark for action denial reasoning where the scene context and motion cues remain intact but the target action is absent, comprising 11,283 video samples.
- We evaluate **20 state-of-the-art MLLMs**, demonstrating that their strong action recognition capabilities fail to translate to robust action denial when misleading contextual cues are present.
- We propose **CausalAct**, a graph-based finetuning framework that represents scene structure as natural language graphs and trains models through graph-centric VQA tasks to improve reasoning about action absence.

2 Related Work

Video Action Recognition Benchmarks: Action recognition datasets span trimmed clips [6,23,36,54,66], untrimmed videos [27,32,86], and domain-specific settings such as sports, instructional content, and driving [15, 21, 34, 43, 53, 63], as well as egocentric views [12,24], multi-label action understanding [2,25,37,65, 79], and skeleton/depth/RGB-based benchmarks [45,46,81]. Most of these focus on recognizing positive actions; negatives are usually implicit, e.g., background or unlabeled frames [32, 86], or are used to probe related failure modes such as similar same-scene actions, missing context, or failed executions [14, 73, 76]. [78] addresses related co-occurrence ambiguities in still images. However, there remains a gap to test denial of a target action in videos when familiar context, objects, or motion cues are present but the defining action is absent.

Causal Reasoning: Work on causal reasoning for multimodal models includes synthetic benchmarks with ground-truth graphs for probing interventions and temporal dependencies [44, 52, 80], and large-scale annotated datasets for real-world event-level reasoning [8, 31]. Graph-based approaches use scene graphs or structured prompts to organize visual content for multimodal reasoning [35, 42, 50,55,60,67], while other efforts ask whether LLMs can internalize abstract causal structures and use them for robust inference [4, 7, 33, 70, 84]. Building on these, our work decomposes videos into structured scene components and models their relations, specifically evaluating whether MLLMs can decide if a target action truly occurs and deny it when the proposal conflicts with the causal evidence.

Multimodal Large Language Models (MLLMs): Video-language research has progressed from retrieval and contrastive pretraining [22, 68, 74, 77, 87] to temporal action models [3,29,56,61,71], and more recently to LLM-centric vision-language systems [5,9,49,72] and instruction-tuned video MLLMs [41,75,83,85]. Emerging "thinking" models add explicit reasoning [17, 20, 26, 59, 69], but many are tuned in ways that induce agreeableness or sycophancy [16, 30, 51, 58, 62, 64]. While these works study bias and reasoning in isolation, evaluation of causal reasoning for action denial in videos remains absent, leading to a dual failure: models are biased by misleading context and prone to affirming false positives. Our work provides a first systematic assessment of this denial capacity and a method to guide MLLMs toward robust, non-sycophantic action understanding.

3 Benchmarking Action Denial

To evaluate a model's ability to deny actions under strong contextual cues, we propose **UCF101-ActionDenial (UCF101-AD)**, a benchmark derived from UCF101 action classes [66]. The dataset contains both *Action-Denial* and *Action-Presence* clips. *Action-Denial* videos preserve scenes, objects, and actors associated with a target class, but the defining motion of that action does not occur. These include cases where no defining motion is present or where a different motion occurs despite similar context. This design creates visually plausible clips

Table 1: Comparison of UCF101-AD with prior benchmarks. No target motion indicates classes where the defining motion is absent. Pos-Neg Pair indicates whether each action has an explicit positive and negative counterpart. ✓* denotes applicability to a subset of denial clips.

Benchmark	Negative Composition				Negative Structure			Reasoning Challenge
	Scope	Coverage	% Clips	Semantics	No Target Motion	Pos-Neg Pair	Explicit Annotation	
HACS [86]	Frame	Partial	-	Distractor	×	×	×	Action localization
THUMOS14 [32]	Frame	Partial	-	Distractor	×	×	×	Action localization
CDAD [76]	Frame	Partial	-	Distractor	×	×	×	Action localization
OOPS [14]	Frame	Partial	100	Attempt	×	×	✓	Failure detection
Mimetics [73]	Clip	Subset	-	Contextual	×	×	×	Spatial ambiguity
SSv2 [23]	Clip	Subset	10	Attempt	×	✓*	✓	Outcome detection
UCF101-AD	**Clip**	**Subset**	**94**	**No action**	✓	✓	✓	**Action Denial**

where context suggests an action while the action itself does not take place, directly testing a model's ability to recognize action absence. *Action-Presence* videos contain the defining motion of the target action and serve as a reference for standard action recognition. An overview of the benchmark is shown in Fig. 2 with Tab. 1 showing comparison with prior benchmarks. While prior works primarily treat negatives as localization distractors, failed attempts, or contextually ambiguous clips, UCF101-AD is explicitly designed around action denial: negatives are constructed to remove the defining motion while preserving plausibly misleading context, and are paired with positive counterparts to support structured reasoning about absence.

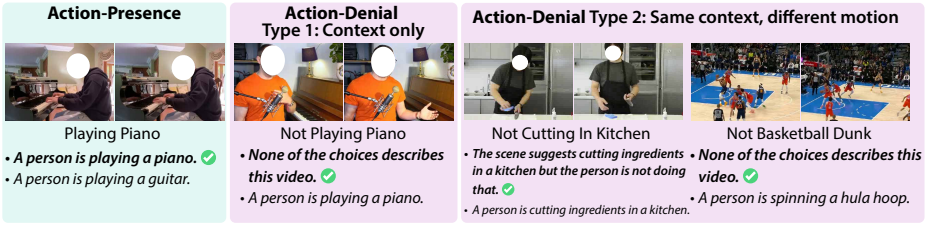

Fig. 2: Overview of UCF101-AD. The dataset contains *Action-Presence* clips and hard *Action-Denial* negatives. Negatives come in two types: ***Type 1 - Context only***, where the usual scene and objects are present but the defining motion is absent, and ***Type 2 - Same context, different motion***, where a different action occurs in the same setting. For the task where the model must select the correct caption for the video, we show examples with different MCQ setups; answer choices are shortened and subsampled for display.

3.1 UCF101-AD Construction

Video Curation: We construct the negative *Action-Denial* videos by collecting clips corresponding to each UCF101 action class where the defining action does

not occur while preserving similar contextual cues such as scene, objects, or actor appearance. These clips resemble the typical setting of the target action but lack the characteristic motion that defines it. Here, the *defining motion* refers to temporally extended activities whose recognition requires modeling motion dynamics over time. Accordingly, we focus on actions characterized by such temporal patterns rather than atomic states such as sitting, standing, or listening. Moreover, the presence of an action requires the full execution of its characteristic motion pattern; partial or incomplete movements are not considered sufficient evidence. For instance, slightly moving a golf club does not constitute a golf swing unless the complete swing motion is performed. Based on how the absence of the target motion manifests, we divide the *Action-Denial* videos into **two types**:

1. **Context only:** Videos preserve the salient objects and environment associated with the target action, but no defining interaction takes place. For example, "Not Playing Piano" shows a person near a piano without actually playing it.
2. **Same context, different motion:** Videos retain the same setting and object interactions as the target action, but the actor performs other plausible actions while never executing the defining motion. For example, "Not Basketball Dunk" shows players dribbling, passing, or shooting on a basketball court, yet no dunk occurs.

Together, these two negative categories disentangle contextual cues from motion evidence, requiring models to verify the defining motion rather than rely on correlated context. In contrast, the positive *Action-Presence* videos contain clips where the defining motion occurs, serving as a reference to verify that models correctly recognize actions when they are actually performed. All videos are sourced from YouTube and temporally segmented into short clips. Following UCF101 [66], clips originating from the same source video are assigned entirely to either the training or test split to avoid information leakage.

QA Generation. For each video, we generate a question asking which description best matches the content of the video, along with several candidate options in a multiple-choice format. We employ this fixed multiple-choice-question based VQA setup to enable objective evaluation by avoiding the variability of free-form generation [38, 39]. For *Action-Denial* videos, the answer choices include a diverse set of candidate actions, including the main distractor corresponding to the original UCF101 target action, additional actions sampled from other categories, and a "None" option. This setup introduces strong visual and linguistic cues toward the target action, allowing us to evaluate whether models can correctly deny the action despite these cues. For *Action-Presence* videos, the correct answer is the target action, and the remaining options are closely related actions plus the same "None" option, forcing the model to distinguish the specific motion rather than rely on coarse context.

Quality control. We employ a fully human verification pipeline to ensure dataset reliability. Annotators first collect candidate videos through keyword searches targeting relevant context without the target action for the *Action-Denial* set. Each clip is manually verified to ensure sufficient contextual cues and

the complete absence of the corresponding action, including partial executions. Low-quality or irrelevant clips are discarded. For the *Action-Presence* set, only clips containing a clear instance of the action are retained. All clips are cross-validated by annotators, removing about 20% of candidates. No MLLMs were used in this process to avoid model-induced bias. After this process, UCF101-AD contains 11,283 clips in total: the *Action-Denial* split includes 7,059 negative clips for training and 3,549 for testing, while the *Action-Presence* split consists of a test-only set of 675 clips with no training clips. In QA generation for every distractor choice, the original UCF101 label is rewritten as a clear descriptive sentence to reduce ambiguity; for example, "HulaHoop" becomes "A person or group of people is spinning a hula hoop."

3.2 Benchmarked Models

We evaluate **20** representative state-of-the-art open-source MLLMs on the test splits of **UCF101-AD** to assess their capacity for action denial. All of our chosen models are trained on very large (public and/or proprietary) datasets.

General-Purpose Models. The evaluated models span major families including Ovis [48,49], Intern [72], Qwen [5], and VideoLLaMA [83], alongside other key architectures [41, 47, 75, 85]; differing in training scale and alignment strategies. We also include the closed-source model GPT-4o mini [57].

Reasoning Models. To examine whether explicit multi-step inference improves action denial, we additionally evaluate recent reasoning-oriented MLLMs, including Ovis 2.5 [49], Video R1 [17], ARC-Hunyuan [20], Kimi-VL [69] and Lumian-VLR [59]. These models generate intermediate reasoning steps before producing a final prediction, enabling more systematic verification.

3.3 Implementation Detail, Evaluation Protocol and Metric

All models are evaluated using their official implementation and released weights. In the multiple-choice (MCQ) setting, the order of the answer options is randomized, and the ground-truth answer is assigned to a random position among them. Each question contains 11 options in total. which ensures that the prompt is reliably processed by all MLLMs while limiting the effectiveness of random guessing to $< 10\%$.

We evaluate all models in a visual question answering (VQA) setup and report several accuracy metrics. For negative clips, we measure *Type 1* and *Type 2* accuracy and their size-weighted average, *Overall Action-Denial (Overall-AD)*. For positive clips, we report *Action-Presence* accuracy. We also provide an *Overall* accuracy as the size-weighted average of *Overall-AD* and *Action-Presence*. To assess a model's balanced ability to both recognize and deny actions, independent of class sizes, we report the *Harmonic Mean (HM)* of *Overall-AD* and *Action-Presence*.

Table 2: Benchmarking models on UCF101-ActionDenial. Reported metric is accuracy (%) for *Type 1* negatives, *Type 2* negatives, *Overall-AD* (Action-Denial) accuracy, *Action-Presence* accuracy, and raw *Overall* dataset accuracy. The last column reports the *Harmonic Mean (HM)* of the *Overall-AD* and *Action-Presence* accuracies. **Best** and second-best performance are highlighted.

Model	Action-Denial			Action-Presence ↑	Overall ↑	HM ↑
	Type 1 ↑	Type 2 ↑	Overall-AD ↑			
Ovis2.5-9B [49]	34.5	33.8	34.1	97.5	43.7	50.5
Ovis2-8B [48]	25.6	27.1	26.4	95.6	36.9	41.4
InternVideo2.5_Chat-8B [72]	19.1	17.7	18.4	96.6	30.3	30.9
InternVL2.5-8B [9]	30.5	32.5	31.6	93.5	41.0	47.2
Qwen2.5-VL-7B-Instruct [5]	22.0	27.9	25.1	95.4	35.8	39.7
VideoLLaMA3-7B [83]	**49.4**	**53.4**	**51.5**	96.0	**58.3**	**67.0**
VideoChat-Flash-Qwen2-7B [41]	27.4	31.1	29.4	94.4	39.3	44.8
Oryx-7B [47]	20.6	23.5	22.1	85.6	31.7	35.1
Valley-Eagle-7B [75]	8.8	13.1	11.1	96.4	24.0	19.9
LLaVA-Video-7B-Qwen2 [85]	28.9	30.7	29.9	96.4	40.0	45.6
Kimi-VL-A3B-Instruct [69]	27.4	31.4	29.6	94.7	39.5	45.1
Ovis2.5-2B [49]	23.1	30.4	27.0	91.1	36.7	41.7
Qwen2.5-VL-3B-Instruct [5]	17.2	22.1	19.8	92.3	30.8	32.6
VideoLLaMA3-2B [83]	20.2	30.0	25.4	93.6	35.8	40.0
Qwen2.5-VL-72B-Instruct [5]	42.6	47.8	45.7	**97.6**	53.6	62.3
GPT4o-mini [57]	20.7	22.3	21.5	90.1	31.9	34.7
Reasoning Models						
Ovis2.5-9B (thinking) [49]	36.7	43.8	40.4	96.7	48.9	57.0
Video-R1-7B [17]	12.7	16.3	14.6	96.0	27.0	25.3
Kimi-VL-A3B-Thinking [69]	14.9	26.1	20.9	96.7	32.4	34.4
ARC-Hunyuan-Video-7B [20]	12.9	16.3	14.7	58.5	21.3	23.5
Lumian-VLR-7B-Thinking [59]	9.8	10.7	10.3	55.3	17.1	17.4

3.4 Results

Tab. 2 summarizes zero-shot performance of all models on UCF101-AD. Across models, *Action-Presence* accuracy is generally high (often above 90%), confirming that current MLLMs recognize actions reliably when the motion is present. In contrast, *Action-Denial* accuracy is much lower: even the strongest model, VideoLLaMA3-7B, reaches only 51.5% *Overall-AD* on negatives, and most models fall in the 20-35% range. Type 1 negatives, which preserve the canonical objects and environments but do not contain the motion, are harder for most models, suggesting that they heavily over-index on context and object presence rather than verifying the critical motion. Type 2 negatives, which keep the same setting but replace the target motion with other plausible activities, yield slightly higher scores for most models, yet performance remains far from robust, implying that models often fail to distinguish fine-grained motion patterns that disambiguate the target action from related activities. As a result, the *Overall* and *Harmonic Mean* between *Overall-AD* and *Action-Presence* accuracy remain modest (maximum 58.3 and 67.0), showing that strong action recognition alone does not translate into robust action denial under strong contextual cues.

Human evaluation of action denial. To assess whether the action-denial examples are reliably verifiable by humans, we conducted a human evaluation on UCF101-AD with 36 participants. Each participant viewed videos from five randomly sampled denial classes and judged whether the target action was absent.

Participants achieved 86.6% *Overall-AD*, indicating that humans can reliably identify action absence even under misleading contextual cues.

3.5 Analysis

Effect of Progressive Reduction of Contextual Confusion. To better understand denial failures, we evaluate two modified MCQ setups that progressively reduce contextual ambiguity. In *Explicit Denial*, the generic "None" option is replaced with a statement that the scene may contain cues for the action but the action itself is not occurring, nudging models toward verification rather than context matching. In *Primary Distractor Removed*, the main competing action label is replaced with a random distractor, removing the strongest alternative. As shown in Fig. 3, accuracy increases from the Standard *Overall-AD* MCQ to Explicit Denial and improves further when the primary distractor is removed, often nearing performance on positive actions. This suggests that models can deny actions when ambiguity is reduced, but still struggle in realistic settings where strong contextual cues and plausible distractors remain present. However, in the real-world we cannot simply remove competing actions or simplify the label space, so models must learn to deny actions even when multiple plausible interpretations remain. Details of setups are in supplementary.

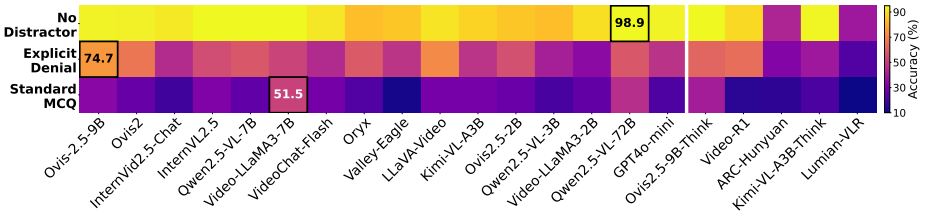

Fig. 3: Effect of progressive hinting on action denial capability. Heatmap of accuracy across three setups: the *Standard MCQ*, *Explicit Denial* with a clarifying verification cue, and *No Distractor* with primary competing action removed. Performance improves as contextual ambiguity is reduced. Best model in each setup is highlighted.

Correlating shortcut learning and sycophancy. To test whether shortcut learning and sycophancy are independent failure modes or manifestations of a single underlying bias, we construct an additional **Binary Yes/No** setup that directly measures sycophancy (e.g., "Is a person or group of people applying makeup to their eyes?"). We then compute a per-video Pearson correlation between shortcut learning and sycophancy, each measured as an error rate: the **MCQ error rate** is the frequency of failing to select the correct "None" option in the Standard MCQ, and the **sycophancy rate** is the frequency of incorrectly answering "yes" in the binary task. As shown in Fig. 4, all models exhibit a positive correlation between these two errors, with a mean correlation of 0.388 and $p < 0.01$ for every model, indicating at least a medium association by Cohen's

conventions [11] for most models. This relationship is slightly stronger on **Type 2** clips (same context, different motion; mean $r = 0.409$) than on **Type 1** clips (context only; mean $r = 0.367$), and most models follow this pattern. Moreover, for every model we observe $P(\text{yes} \mid \text{MCQ wrong}) > P(\text{yes} \mid \text{MCQ correct})$ and $P(\text{yes} \mid \text{primary distractor}) > P(\text{yes} \mid \text{random distractor})$, usually by 0.2-0.6, showing that sycophantic answers concentrate precisely on cases where the model has already latched onto the misleading action hypothesis. This provides evidence that shortcut learning and sycophancy are tightly correlated: models infer actions from contextual cues rather than verifying the defining motion.

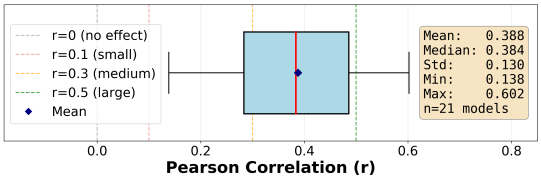

Fig. 4: Distribution of dual-failure correlations. Pearson's r coefficient correlates each model's shortcut learning error with its sycophancy error showing a positive correlation ($r > 0$).

General vs. Thinking models. Our analysis of the negative videos in the UCF101-ActionDenial dataset reveals a counterintuitive pattern: reasoning models generally perform worse than their standard counterparts (except Ovis2.5). Instead of improving robustness, their chains-of-thought often over-index on visual context, confidently selecting the most plausible distractor action or even unrelated options. In many cases they also struggle to properly interpret the denial instruction itself, instead hallucinating descriptions of actions that are not present. This suggests that current "thinking" models are poorly calibrated for negative constraints: tending to explain what might be happening rather than justify that nothing is.

4 Improving Action Denial through Causal Graph

UCF101-AD reveals a fundamental limitation in current MLLMs: they often equate contextual cues with the occurrence of an action. These models operate under a correlational paradigm, implicitly modeling the presence of context with the existence of action and thus lack the capacity for denial. To address this limitation, we move from context-based reasoning to **causal reasoning**. We treat an action as the outcome of a structured combination of factors including participants, environment, interactions, and most importantly, motion dynamics. We therefore introduce **CausalAct**, which organizes these elements into a causal structure and encourages the model to verify the required evidence before predicting an action, rather than relying on contextual shortcuts.

4.1 CausalAct

Causal Framework. We model the evidence required for recognizing an action using a directed acyclic graph (DAG) that organizes scene elements into contextual, relational, and dynamic components. Inspired by structured video representations such as ActionGenome [31], **CausalAct** decomposes an action scene into the following variables: contextual nodes for persons (P), objects (O), and location (L); relational nodes capturing spatial relations (S) and interactions (I) between entities; a dynamic node representing motion (M); and an action node (A) representing the final activity label (Fig. 5). The dependencies between these variables follow a causal hierarchy:

$$S \leftarrow f_S(P, O, L), \quad I \leftarrow f_I(P, O), \quad M \leftarrow f_M(I), \quad A \leftarrow f_A(I, M), \tag{1}$$

where each $f(\cdot)$ denotes an abstract mapping from the video elements that may influence the corresponding variable.

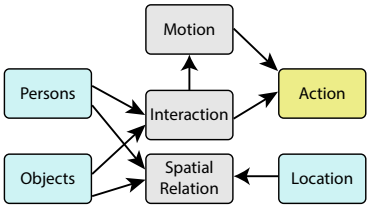

Fig. 5: ***CausalAct showing the components of an action scene:*** Persons (P), Objects (O), and Location (L) are contextual nodes; Spatial Relation (S) and Interaction (I) are relational nodes; Motion (M) is the dynamic node; and Action (A) is the final activity label.

Intuitively, contextual elements (persons, objects, and environment) determine the spatial configuration of the scene, which enables potential interactions. These interactions generate motion patterns, and only when the appropriate motion is observed should the action be inferred. This structure explicitly prevents shortcut reasoning where models predict actions directly from context ($A \leftarrow L$ or $A \leftarrow O$). Instead, the model must verify the presence of interaction and motion before confirming the action. The graph also accommodates different action types: for person-centric actions, reasoning may proceed directly from motion to action, whereas object-centric actions require both interaction and motion as prerequisites.

Operationalizing CausalAct. To operationalize the proposed causal structure within MLLMs, we translate CausalAct into a structured reasoning framework using the CausalAct Prompt, a natural language instantiation of the graph. The prompt describes the causal variables and their dependencies, guiding the model to examine contextual elements (persons, objects, and environment), relational evidence (spatial relations and interactions), and motion dynamics before predicting the action. Rather than acting as a strict rule-based module, CausalAct provides visual-grounding guidance that encourages the model to organize the observed evidence through the causal graph rather than shortcutting from scene/object context. By explicitly structuring this reasoning process, it

encourages the model to verify whether the required causal chain is present and to deny the action when critical evidence, particularly the defining motion, is absent. However, models with weak vision-language grounding may struggle to reliably follow this structured reasoning.

To address this, we introduce an auxiliary finetuning stage that teaches models the dependencies encoded in **CausalAct**. For each video, we construct a graph over the variables (P, O, L, S, I, M, A) and automatically generate graph-based question-answer tasks probing node identities, edge relationships, causal paths, and property consistency. These questions are derived directly from the graph structure and combined with synthetic distractors, forcing the model to reason over the causal dependencies rather than memorizing action labels. Importantly, the tasks **avoid exposing the ground-truth action to prevent label leakage**. *Learning* these graph reasoning tasks encourages models to internalize the causal relationships between context, interaction, motion, and action, resulting in improved ability to correctly deny actions in *Action-Denial* scenarios. Details of finetuning are provided in Supplementary.

4.2 Evaluating CausalAct

Effect of CausalAct in Action Denial. We first evaluate the graph in a zero-shot setting (*CausalAct-0*). As shown in Fig. 6, models with stronger language reasoning (Ovis2.5, Qwen2.5-VL) leverage it to better reject false positives, whereas VideoLLaMA3 often fails to follow the causal instructions and can even degrade, indicating that prompting alone is insufficient under weak vision-language alignment.

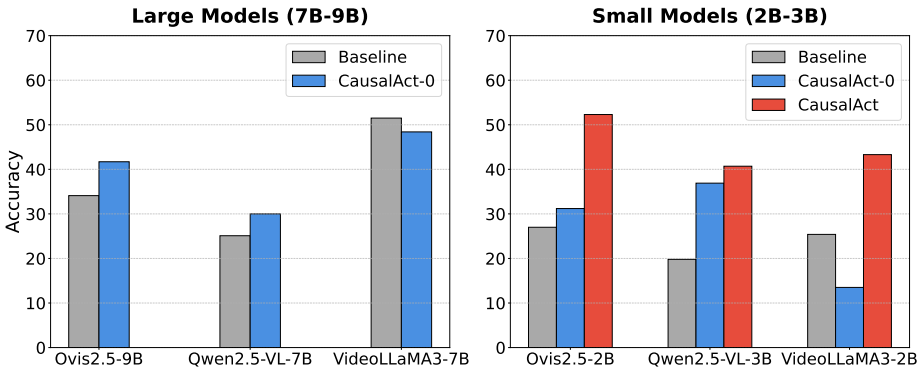

Fig. 6: Comparison of baseline, CausalAct-0, and CausalAct. Left: Larger models with 7B-9B parameters (only zero-shot); right: small models with 2B-3B parameters.

We then finetune the smaller models on graph-based reasoning questions derived from CausalAct, which substantially improves denial accuracy on the UCF101-

AD *Action-Denial* test set and shifts behavior from contextual shortcutting to denying actions when the defining motion is absent. Although trained only on abstract graph-structure questions, not negative labels, the models acquire a transferable skill of checking the causal chain before asserting an action, as illustrated in Fig. 7.

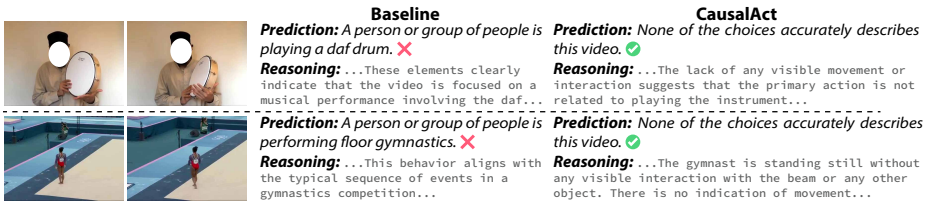

Fig. 7: ***Qualitative comparison of the baseline vs. CausalAct.*** The base Qwen2.5-VL-3B-Instruct takes shortcuts based on contextual cues in the scene, whereas CausalAct correctly leverages the graph structure to verify the defining motion, enabling it to deny the non-existent action for both *Not Playing Daf* and *Not Floor Gymnastics*.

Generalizability to other datasets: To evaluate whether CausalAct-0 transfers beyond our benchmark, we apply it to Ovis2.5-9B on additional action datasets by replacing the ground-truth action with "None of the choices...". As shown in Tab. 3, CausalAct-0 improves denial accuracy across all evaluated datasets, including UCF101 [66], K400 [6], HMDB51 [36], and SSv2 [23]. We further evaluate on fine-grained motion datasets, Diving48 [43] and FineGym99 [63], where verifying an action often requires subtle temporal motion, interaction, or ordering cues. These gains suggest that causal reasoning helps the model verify whether the defining motion or outcome of a candidate action truly occurs, improving action denial across diverse dataset distributions.

Table 3: ***Generalization performance on external action datasets.*** We compare Ovis2.5-9B's baseline denial accuracy with CausalAct-0.

Dataset	Base ↑	CausalAct-0 ↑	Δ
UCF101 [66]	78.7	86.1	+7.4
K400 [6]	80.4	91.9	+11.5
HMDB51 [36]	51.2	74.3	+23.1
SSv2 [23]	16.0	37.4	+21.4
Diving48 [43]	48.2	70.9	+22.7
FineGym99 [63]	71.6	87.4	+15.8

4.3 Ablation Studies

Does graph structure matter? To test whether performance gains come from the specific causal structure or just any arbitrary graph, we ablate CausalAct by comparing it to a Pruned Graph (only P, O, I, A nodes) and, a Random Graph (all nodes, but shuffled edges). Fig. 8 shows that in zero-shot setting (CausalAct-0), models show limited sensitivity to these variants, suggesting that without

additional training, they struggle to fully understand the graph's structure. After finetuning, however, accuracy drops in most cases when the causal dependencies are disrupted, indicating that models have learned to utilize the full structure. These trends suggest that the prompt only serves as an interface for expressing the causal graph to the MLLM. Finetuning with structured representations helps models internalize and use the intended causal graph for reasoning, while cases with minimal drops hint that they weakly rely on the graph or do not parse it.

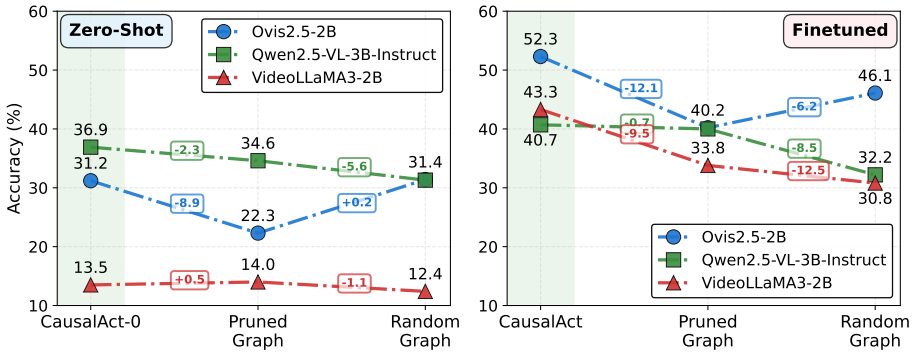

Fig. 8: Ablation study on graph structure: We compare CausalAct to a Pruned Graph (P-O-I-A only) and a Random Graph (shuffled edges). Numeric annotations quantify decline (-) or improvement (+) relative to CausalAct.

Is the object node necessary for person-only actions? To test this, we evaluate a reduced graph that omits object and interaction nodes (mapping only Person, Location, Spatial Relation, Motion, and Action) on a subset of 25 purely body and location-based negatives (e.g., Not Baby Crawling, Not Cliff Diving). As shown in Fig. 9, zero-shot accuracy with both the full and reduced graphs follows the same trend and matches or exceeds the baseline, indicating that CausalAct's gains largely stem from organizing causal relations among core scene elements, even when no explicit object reasoning is required.

Is it just language tuning? We ask whether graph conceptualization finetuning simply teaches new text patterns or requires deeper vision-language alignment. We compare finetuning the **full model** vs. finetuning **only the LLM and projector**, freezing the vision encoder. We observed that restricting updates to the visual stack consistently hurts performance, with accuracy dropping by roughly 15% on average. This systematic degradation indicates that our method is not simply encouraging prompt-following. To effectively ground the causal graph, the vision encoder must also be updated so that visual evidence about the presence or absence of the action is utilized by the MLLM.

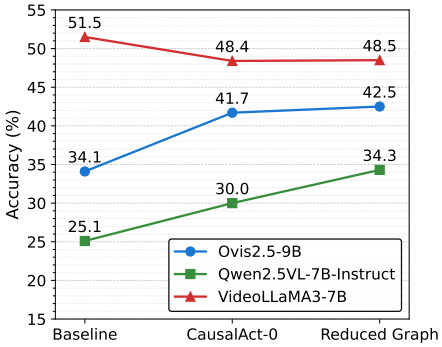

Fig. 9: ***Effect of removing object nodes for person-only actions.*** Baseline and CausalAct-0 are evaluated on all *Action-Denial* videos, while the Reduced Graph is evaluated on body and location-based actions. Comparable performance shows the causal structure remains effective without explicit object-interaction reasoning.

5 Conclusion

We introduce **UCF101-AD**, a benchmark targeting a critical failure of modern MLLMs: denying false actions under strong contextual cues. Our experiments on 20 MLLMs show that models excel at positive action recognition but perform poorly on our *Action-Denial* test set (most scoring $< 50\%$), revealing a dual bias: a perceptual reliance on static shortcuts and a linguistic tendency toward agreeableness. To guide models toward causal reasoning, we propose **CausalAct** and accompanying graph-conceptualization finetuning tasks, which improve their capacity for denial. We intend UCF101-AD to serve as a tool for advancing causal reasoning and teaching models to deny.

Acknowledgements

We gratefully acknowledge Vibhav Vineet [Microsoft Research, USA], Shahzad Ahmad [Norwegian University of Science and Technology, Norway], and Sukalpa Chanda [Østfold University College, Norway] for their valuable support and contributions to this work.

References

1. Abdullah, R., Claypoole, J., Cogswell, M., Divakaran, A., Rawat, Y.: Punching bag vs. punching person: Motion transferability in videos. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 11348–11358 (October 2025)
2. Abdullah, R., Rawat, Y.S., Vyas, S.: isafetybench: A video-language benchmark for safety in industrial environment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops. pp. 1433–1442 (October 2025)
3. Ahmad, S., Chanda, S., Rawat, Y.S.: T2l: Efficient zero-shot action recognition with temporal token learning. Transactions on Machine Learning Research (2025), `https://openreview.net/forum?id=WvgoxpGpuU`

4. Bagheri, A., Alinejad, M., Bello, K., Akhondi-Asl, A.: C^2P: Featuring large language models with causal reasoning (2024), https://arxiv.org/abs/2407.18069
5. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report (2025), https://arxiv.org/abs/2502.13923
6. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 4724–4733 (2017), https://api.semanticscholar.org/CorpusID:206596127
7. Chen, S., Xu, M., Wang, K., Zeng, X., Zhao, R., Zhao, S., Lu, C.: CLEAR: Can language models really understand causal graphs? In: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2024. pp. 6247–6265. Association for Computational Linguistics, Miami, Florida, USA (Nov 2024). https://doi.org/10.18653/v1/2024.findings-emnlp.363, https://aclanthology.org/2024.findings-emnlp.363/
8. Chen, T., Liu, H., He, T., Chen, Y., Ma, X., Zhong, C., Zhang, Y., Wang, Y., Lin, H., Lin, W., et al.: Mecd: Unlocking multi-event causal discovery in video reasoning. Advances in Neural Information Processing Systems **37**, 92554–92580 (2024)
9. Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., et al.: Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. arXiv preprint arXiv:2412.05271 (2024)
10. Choi, J., Gao, C., Messou, J.C.E., Huang, J.B.: Why can't i dance in the mall? learning to mitigate scene bias in action recognition. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 32. Curran Associates, Inc. (2019), https://proceedings.neurips.cc/paper_files/paper/2019/file/ab817c9349cf9c4f6877e1894a1faa00-Paper.pdf
11. Cohen, J.: Statistical Power Analysis for the Behavioral Sciences. Lawrence Erlbaum Associates, Hillsdale, NJ, 2nd edn. (1988)
12. Damen, D., Doughty, H., Farinella, G.M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., Wray, M.: Scaling egocentric vision: The epic-kitchens dataset. In: Proceedings of the European Conference on Computer Vision (ECCV) (September 2018)
13. Duan, H., Zhao, Y., Chen, K., Xiong, Y., Lin, D.: Mitigating representation bias in action recognition: Algorithms and benchmarks. In: Karlinsky, L., Michaeli, T., Nishino, K. (eds.) Computer Vision – ECCV 2022 Workshops. pp. 557–575. Springer Nature Switzerland, Cham (2023)
14. Epstein, D., Chen, B., Vondrick, C.: Oops! predicting unintentional action in video. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 916–926 (2020). https://doi.org/10.1109/CVPR42600.2020.00100
15. Fang, J., Yan, D., Qiao, J., Xue, J., Wang, H., Li, S.: Dada-2000: Can driving accident be predicted by driver attentionf analyzed by a benchmark. In: 2019 IEEE Intelligent Transportation Systems Conference (ITSC). p. 4303–4309. IEEE Press (2019). https://doi.org/10.1109/ITSC.2019.8917218, https://doi.org/10.1109/ITSC.2019.8917218
16. Fanous, A., Goldberg, J., Agarwal, A.A., Lin, J., Zhou, A., Daneshjou, R., Koyejo, S.: Syceval: Evaluating llm sycophancy. CoRR **abs/2502.08177** (Febru-

ary 2025), http://dblp.uni-trier.de/db/journals/corr/corr2502.html#abs-2502-08177
17. Feng, K., Gong, K., Li, B., Guo, Z., Wang, Y., Peng, T., Wang, B., Yue, X.: Video-r1: Reinforcing video reasoning in mllms. arXiv preprint arXiv:2503.21776 (2025)
18. Fioresi, J., Dave, I.R., Shah, M.: Albar: Adversarial learning approach to mitigate biases in action recognition. In: The Thirteenth International Conference on Learning Representations (2025)
19. Fukuzawa, T., Hara, K., Kataoka, H., Tamaki, T.: Can masking background and object reduce static bias for zero-shot action recognition? In: Ide, I., Kompatsiaris, I., Xu, C., Yanai, K., Chu, W.T., Nitta, N., Riegler, M., Yamasaki, T. (eds.) MultiMedia Modeling. pp. 366–379. Springer Nature Singapore, Singapore (2025)
20. Ge, Y., Ge, Y., Li, C., Wang, T., Pu, J., Li, Y., Qiu, L., Ma, J., Duan, L., Zuo, X., et al.: Arc-hunyuan-video-7b: Structured video comprehension of real-world shorts. arXiv preprint arXiv:2507.20939 (2025)
21. Giancola, S., Amine, M., Dghaily, T., Ghanem, B.: SoccerNet: A Scalable Dataset for Action Spotting in Soccer Videos . In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 1792–179210. IEEE Computer Society, Los Alamitos, CA, USA (Jun 2018). https://doi.org/10.1109/CVPRW.2018.00223, https://doi.ieeecomputersociety.org/10.1109/CVPRW.2018.00223
22. Gorti, S.K., Vouitsis, N., Ma, J., Golestan, K., Volkovs, M., Garg, A., Yu, G.: X-pool: Cross-modal language-video attention for text-video retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022)
23. Goyal, R., Kahou, S.E., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., Hoppe, F., Thurau, C., Bax, I., Memisevic, R.: The "Something Something" Video Database for Learning and Evaluating Visual Common Sense . In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 5843–5851. IEEE Computer Society, Los Alamitos, CA, USA (Oct 2017). https://doi.org/10.1109/ICCV.2017.622, https://doi.ieeecomputersociety.org/10.1109/ICCV.2017.622
24. Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., Martin, M., Nagarajan, T., Radosavovic, I., Ramakrishnan, S.K., Ryan, F., Sharma, J., Wray, M., Xu, M., Xu, E.Z., Zhao, C., Bansal, S., Batra, D., Cartillier, V., Crane, S., Do, T., Doulaty, M., Erapalli, A., Feichtenhofer, C., Fragomeni, A., Fu, Q., Gebreselasie, A., González, C., Hillis, J., Huang, X., Huang, Y., Jia, W., Khoo, W., Kolář, J., Kottur, S., Kumar, A., Landini, F., Li, C., Li, Y., Li, Z., Mangalam, K., Modhugu, R., Munro, J., Murrell, T., Nishiyasu, T., Price, W., Puentes, P.R., Ramazanova, M., Sari, L., Somasundaram, K., Southerland, A., Sugano, Y., Tao, R., Vo, M., Wang, Y., Wu, X., Yagi, T., Zhao, Z., Zhu, Y., Arbeláez, P., Crandall, D., Damen, D., Farinella, G.M., Fuegen, C., Ghanem, B., Ithapu, V.K., Jawahar, C.V., Joo, H., Kitani, K., Li, H., Newcombe, R., Oliva, A., Park, H.S., Rehg, J.M., Sato, Y., Shi, J., Shou, M.Z., Torralba, A., Torresani, L., Yan, M., Malik, J.: Ego4d: Around the world in 3,000 hours of egocentric video. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18973–18990 (2022). https://doi.org/10.1109/CVPR52688.2022.01842
25. Gu, C., Sun, C., Ross, D.A., Vondrick, C., Pantofaru, C., Li, Y., Vijayanarasimhan, S., Toderici, G., Ricco, S., Sukthankar, R., Schmid, C., Malik, J.: Ava: A video dataset of spatio-temporally localized atomic visual actions. In: 2018 IEEE/CVF

Conference on Computer Vision and Pattern Recognition. pp. 6047–6056 (2018). https://doi.org/10.1109/CVPR.2018.00633
26. Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., Xu, R., Zhang, R., Ma, S., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z.F., Gou, Z., Shao, Z., Li, Z., Gao, Z., Liu, A., Xue, B., Wang, B., Wu, B., Feng, B., Lu, C., Zhao, C., Deng, C., Ruan, C., Dai, D., Chen, D., Ji, D., Li, E., Lin, F., Dai, F., Luo, F., Hao, G., Chen, G., Li, G., Zhang, H., Xu, H., Ding, H., Gao, H., Qu, H., Li, H., Guo, J., Li, J., Chen, J., Yuan, J., Tu, J., Qiu, J., Li, J., Cai, J.L., Ni, J., Liang, J., Chen, J., Dong, K., Hu, K., You, K., Gao, K., Guan, K., Huang, K., Yu, K., Wang, L., Zhang, L., Zhao, L., Wang, L., Zhang, L., Xu, L., Xia, L., Zhang, M., Zhang, M., Tang, M., Zhou, M., Li, M., Wang, M., Li, M., Tian, N., Huang, Y., et al.: Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. Nature **645**, 633–638 (2025). https://doi.org/10.1038/s41586-025-09422-z, https://www.nature.com/articles/s41586-025-09422-z
27. Heilbron, F.C., Escorcia, V., Ghanem, B., Niebles, J.C.: Activitynet: A large-scale video benchmark for human activity understanding. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 961–970 (2015). https://doi.org/10.1109/CVPR.2015.7298698
28. Huang, D.A., Ramanathan, V., Mahajan, D., Torresani, L., Paluri, M., Fei-Fei, L., Niebles, J.C.: What makes a video a video: Analyzing temporal information in video understanding models and datasets. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7366–7375 (2018). https://doi.org/10.1109/CVPR.2018.00769
29. Huang, X., Zhou, H., Yao, K., Han, K.: FROSTER: Frozen CLIP is a strong teacher for open-vocabulary action recognition. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=zYXFMeHRtO
30. Jain, S., Ahmed, U.Z., Sahai, S., Leong, B.: Beyond consensus: Mitigating the agreeableness bias in llm judge evaluations (2025), https://arxiv.org/abs/2510.11822
31. Ji, J., Krishna, R., Fei-Fei, L., Niebles, J.C.: Action genome: Actions as compositions of spatio-temporal scene graphs. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10233–10244 (2020). https://doi.org/10.1109/CVPR42600.2020.01025
32. Jiang, Y.G., Liu, J., Roshan Zamir, A., Toderici, G., Laptev, I., Shah, M., Sukthankar, R.: THUMOS challenge: Action recognition with a large number of classes. http://crcv.ucf.edu/THUMOS14/ (2014)
33. Jin, Z., Chen, Y., Leeb, F., Gresele, L., Kamal, O., LYU, Z., Blin, K., Adauto, F.G., Kleiman-Weiner, M., Sachan, M., Schölkopf, B.: CLadder: A benchmark to assess causal reasoning capabilities of language models. In: Thirty-seventh Conference on Neural Information Processing Systems (2023), https://openreview.net/forum?id=e2wtjx0Yqu
34. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1725–1732 (2014). https://doi.org/10.1109/CVPR.2014.223
35. Koupaee, M., Bai, X., Chen, M., Durrett, G., Chambers, N., Balasubramanian, N.: Causal graph based event reasoning using semantic relation experts. In: Che, W., Nabende, J., Shutova, E., Pilehvar, M.T. (eds.) Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1:

Long Papers). pp. 26169–26199. Association for Computational Linguistics, Vienna, Austria (Jul 2025). https://doi.org/10.18653/v1/2025.acl-long.1269, https://aclanthology.org/2025.acl-long.1269/

36. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T.A., Serre, T.: Hmdb: A large video database for human motion recognition. 2011 International Conference on Computer Vision pp. 2556–2563 (2011), https://api.semanticscholar.org/CorpusID:206769852
37. Li, A., Thotakuri, M., Ross, D.A., Carreira, J., Vostrikov, A., Zisserman, A.: The ava-kinetics localized human actions video dataset. ArXiv **abs/2005.00214** (2020), https://api.semanticscholar.org/CorpusID:218470050
38. Li, B., Ge, Y., Ge, Y., Wang, G., Wang, R., Zhang, R., Shan, Y.: Seed-bench: Benchmarking multimodal large language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13299–13308 (June 2024)
39. Li, K., Wang, Y., He, Y., Li, Y., Wang, Y., Liu, Y., Wang, Z., Xu, J., Chen, G., Lou, P., Wang, L., Qiao, Y.: Mvbench: A comprehensive multi-modal video understanding benchmark. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 22195–22206 (2024). https://doi.org/10.1109/CVPR52733.2024.02095
40. Li, T., Foo, L.G., Ke, Q., Rahmani, H., Wang, A., Wang, J., Liu, J.: Dynamic spatio-temporal specialization learning for fine-grained action recognition. In: Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV. p. 386–403. Springer-Verlag, Berlin, Heidelberg (2022). https://doi.org/10.1007/978-3-031-19772-7_23, https://doi.org/10.1007/978-3-031-19772-7_23
41. Li, X., Wang, Y., Yu, J., Zeng, X., Zhu, Y., Huang, H., Gao, J., Li, K., He, Y., Wang, C., Qiao, Y., Wang, Y., Wang, L.: Videochat-flash: Hierarchical compression for long-context video modeling. arXiv preprint arXiv:2501.00574 (2024)
42. Li, Y., Yang, X., Bao, B.K., Xu, C.: Graph prompts: Adapting video graph for video question answering. In: Kwok, J. (ed.) Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25. pp. 1485–1493. International Joint Conferences on Artificial Intelligence Organization (8 2025). https://doi.org/10.24963/ijcai.2025/166, https://doi.org/10.24963/ijcai.2025/166, main Track
43. Li, Y., Li, Y., Vasconcelos, N.: Resound: Towards action recognition without representation bias. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision – ECCV 2018. pp. 520–535. Springer International Publishing, Cham (2018)
44. Li, Y., Torralba, A., Anandkumar, A., Fox, D., Garg, A.: Causal discovery in physical systems from videos. Advances in Neural Information Processing Systems **33** (2020)
45. Liu, C., Hu, Y., Li, Y., Song, S., Liu, J.: Pku-mmd: A large scale benchmark for skeleton-based human action understanding. In: Proceedings of the Workshop on Visual Analysis in Smart and Connected Communities. p. 1–8. VSCC '17, Association for Computing Machinery, New York, NY, USA (2017). https://doi.org/10.1145/3132734.3132739, https://doi.org/10.1145/3132734.3132739
46. Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L.Y., Kot, A.C.: Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence **42**(10), 2684–2701 (2020). https://doi.org/10.1109/TPAMI.2019.2916873

47. Liu, Z., Dong, Y., Liu, Z., Hu, W., Lu, J., Rao, Y.: Oryx MLLM: On-demand spatial-temporal understanding at arbitrary resolution. In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=ODiY6pbHZQ
48. Lu, S., Li, Y., Chen, Q.G., Xu, Z., Luo, W., Zhang, K., Ye, H.J.: Ovis: Structural embedding alignment for multimodal large language model. arXiv:2405.20797 (2024)
49. Lu, S., Li, Y., Xia, Y., Hu, Y., Zhao, S., Ma, Y., Wei, Z., Li, Y., Duan, L., Zhao, J., Han, Y., Li, H., Chen, W., Tang, J., Hou, C., Du, Z., Zhou, T., Zhang, W., Ding, H., Li, J., Li, W., Hu, G., Gu, Y., Yang, S., Wang, J., Sun, H., Wang, Y., Sun, H., Huang, J., He, Y., Shi, S., Zhang, W., Zheng, G., Jiang, J., Gao, S., Wu, Y.F., Chen, S., Chen, Y., Chen, Q.G., Xu, Z., Luo, W., Zhang, K.: Ovis2.5 technical report. arXiv:2508.11737 (2025)
50. Ma, H., Pathak, V., Wang, D.Z.: Bridging vision language models and symbolic grounding for video question answering (2025), https://arxiv.org/abs/2509.11862
51. Malmqvist, L.: Sycophancy in large language models: Causes and mitigations. In: Arai, K. (ed.) Intelligent Computing. pp. 61–74. Springer Nature Switzerland, Cham (2025)
52. McDuff, D., Song, Y., Lee, J., Vineet, V., Vemprala, S., Gyde, N., Salman, H., Ma, S., Sohn, K., Kapoor, A.: Causalcity: Complex simulations with agency for causal discovery and reasoning (June 2021), https://www.microsoft.com/en-us/research/publication/causalcity-complex-simulations-with-agency-for-causal-discovery-and-reasoning/, preprint under review
53. Miech, A., Zhukov, D., Alayrac, J.B., Tapaswi, M., Laptev, I., Sivic, J.: Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2630–2640 (2019). https://doi.org/10.1109/ICCV.2019.00272
54. Monfort, M., Vondrick, C., Oliva, A., Andonian, A., Zhou, B., Ramakrishnan, K., Bargal, S.A., Yan, T., Brown, L., Fan, Q., Gutfreund, D.: Moments in Time Dataset: One Million Videos for Event Understanding . IEEE Transactions on Pattern Analysis & Machine Intelligence **42**(02), 502–508 (Feb 2020). https://doi.org/10.1109/TPAMI.2019.2901464, https://doi.ieeecomputersociety.org/10.1109/TPAMI.2019.2901464
55. Nguyen, T.T., Nguyen, P., Cothren, J., Yilmaz, A., Luu, K.: Hyperglm: Hypergraph for video scene graph generation and anticipation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 29150–29160 (June 2025)
56. Ni, B., Peng, H., Chen, M., Zhang, S., Meng, G., Fu, J., Xiang, S., Ling, H.: Expanding language-image pretrained models for general video recognition. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision – ECCV 2022. pp. 1–18. Springer Nature Switzerland, Cham (2022)
57. OpenAI: Gpt-4o mini: Advancing cost-efficient intelligence. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/ (Jul 2024), accessed 2026-06-27
58. OpenAI: Sycophancy in gpt-4o: what happened and what we're doing about it. https://openai.com/index/sycophancy-in-gpt-4o/ (Apr 2025), accessed 2025-11-09
59. prithivMLmods: Lumian-vlr-7b-thinking. https://huggingface.co/prithivMLmods/LumianâĂŚVLRâĂŚ7BâĂŚThinking (2025), hugging Face model card

60. Ranasinghe, K., Li, X., Kahatapitiya, K., Ryoo, M.S.: Understanding long videos with multimodal language models. In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=OxKi02I29I
61. Rasheed, H., Khattak, M.U., Maaz, M., Khan, S., Khan, F.S.: Fine-tuned clip models are efficient video learners. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6545–6554 (2023). https://doi.org/10.1109/CVPR52729.2023.00633
62. Rrv, A., Tyagi, N., Uddin, M.N., Varshney, N., Baral, C.: Chaos with keywords: Exposing large language models sycophancy to misleading keywords and evaluating defense strategies. In: Ku, L.W., Martins, A., Srikumar, V. (eds.) Findings of the Association for Computational Linguistics: ACL 2024. pp. 12717–12733. Association for Computational Linguistics, Bangkok, Thailand (Aug 2024). https://doi.org/10.18653/v1/2024.findings-acl.755, https://aclanthology.org/2024.findings-acl.755/
63. Shao, D., Zhao, Y., Dai, B., Lin, D.: Finegym: A hierarchical video dataset for fine-grained action understanding. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2613–2622 (2020). https://doi.org/10.1109/CVPR42600.2020.00269
64. Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S.R., DURMUS, E., Hatfield-Dodds, Z., Johnston, S.R., Kravec, S.M., Maxwell, T., McCandlish, S., Ndousse, K., Rausch, O., Schiefer, N., Yan, D., Zhang, M., Perez, E.: Towards understanding sycophancy in language models. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=tvhaxkMKAn
65. Sigurdsson, G.A., Varol, G., Wang, X., Farhadi, A., Laptev, I., Gupta, A.: Hollywood in homes: Crowdsourcing data collection for activity understanding. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) Computer Vision – ECCV 2016. pp. 510–526. Springer International Publishing, Cham (2016)
66. Soomro, K., Roshan Zamir, A., Shah, M.: UCF101: A dataset of 101 human actions classes from videos in the wild. In: CRCV-TR-12-01 (2012)
67. Taluzzi, A., Gesualdi, D., Santambrogio, R., Plizzari, C., Palermo, F., Mentasti, S., Matteucci, M.: From pixels to graphs: using scene and knowledge graphs for hd-epic vqa challenge. CoRR **abs/2506.08553** (June 2025), https://doi.org/10.48550/arXiv.2506.08553
68. Tang, M., Wang, Z., LIU, Z., Rao, F., Li, D., Li, X.: Clip4caption: Clip for video caption. In: Proceedings of the 29th ACM International Conference on Multimedia. p. 4858–4862. MM '21, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3474085.3479207, https://doi.org/10.1145/3474085.3479207
69. Team, K., Du, A., Yin, B., Xing, B., Qu, B., Wang, B., Chen, C., Zhang, C., Du, C., Wei, C., Wang, C., Zhang, D., Du, D., Wang, D., Yuan, E., Lu, E., Li, F., Sung, F., Wei, G., Lai, G., Zhu, H., Ding, H., Hu, H., Yang, H., Zhang, H., Wu, H., Yao, H., Lu, H., Wang, H., Gao, H., Zheng, H., Li, J., Su, J., Wang, J., Deng, J., Qiu, J., Xie, J., Wang, J., Liu, J., Yan, J., Ouyang, K., Chen, L., Sui, L., Yu, L., Dong, M., Dong, M., Xu, N., Cheng, P., Gu, Q., Zhou, R., Liu, S., Cao, S., Yu, T., Song, T., Bai, T., Song, W., He, W., Huang, W., Xu, W., Yuan, X., Yao, X., Wu, X., Zu, X., Zhou, X., Wang, X., Charles, Y., Zhong, Y., Li, Y., Hu, Y., Chen, Y., Wang, Y., Liu, Y., Miao, Y., Qin, Y., Chen, Y., Bao, Y., Wang, Y., Kang, Y., Liu, Y., Du, Y., Wu, Y., Wang, Y., Yan, Y., Zhou, Z., Li, Z., Jiang, Z., Zhang,

Z., Yang, Z., Huang, Z., Huang, Z., Zhao, Z., Chen, Z.: Kimi-VL technical report (2025), https://arxiv.org/abs/2504.07491
70. Vashishtha, A., Kumar, A., Pandey, A., Reddy, A.G., Balasubramanian, V.N., Sharma, A.: Teaching transformers causal reasoning through axiomatic training. In: Causality and Large Models @NeurIPS 2024 (2024), https://openreview.net/forum?id=vnFtU3fO9h
71. Wang, M., Xing, J., Mei, J., Liu, Y., Jiang, Y.: Actionclip: Adapting language-image pretrained models for video action recognition. IEEE Transactions on Neural Networks and Learning Systems **36**(1), 625–637 (2025). https://doi.org/10.1109/TNNLS.2023.3331841
72. Wang, Y., Li, X., Yan, Z., He, Y., Yu, J., Zeng, X., Wang, C., Ma, C., Huang, H., Gao, J., Dou, M., Chen, K., Wang, W., Qiao, Y., Wang, Y., Wang, L.: Internvideo2.5: Empowering video mllms with long and rich context modeling. arXiv preprint arXiv:2501.12386 (2025)
73. Weinzaepfel, P., Rogez, G.: Mimetics: Towards understanding human actions out of context. International Journal of Computer Vision **129**, 1675 – 1690 (2019), https://api.semanticscholar.org/CorpusID:209376248
74. Weng, Z., Yang, X., Li, A., Wu, Z., Jiang, Y.G.: Open-vclip: Transforming clip to an open-vocabulary video model via interpolated weight optimization. In: ICML (2023)
75. Wu, Z., Chen, Z., Luo, R., Zhang, C., Gao, Y., He, Z., Wang, X., Lin, H., Qiu, M.: Valley2: Exploring multimodal models with scalable vision-language design. arXiv preprint arXiv:2501.05901 (2025)
76. Xiang, W., Li, C., Li, K., Wang, B., Hua, X.S., Zhang, L.: Cdad: A common daily action dataset with collected hard negative samples. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 3920–3929 (2022). https://doi.org/10.1109/CVPRW56347.2022.00437
77. Xu, H., Ghosh, G., Huang, P.Y., Okhonko, D., Aghajanyan, A., Metze, F., Zettlemoyer, L., Feichtenhofer, C.: Videoclip: Contrastive pre-training for zero-shot video-text understanding (2021)
78. Yao, B., Fei-Fei, L.: Grouplet: A structured image representation for recognizing human and object interactions. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 9–16 (2010). https://doi.org/10.1109/CVPR.2010.5540234
79. Yeung, S., Russakovsky, O., Jin, N., Andriluka, M., Mori, G., Fei-Fei, L.: Every moment counts: Dense detailed labeling of actions in complex videos. Int. J. Comput. Vision **126**(2–4), 375–389 (Apr 2018). https://doi.org/10.1007/s11263-017-1013-y, https://doi.org/10.1007/s11263-017-1013-y
80. Yi, K., Gan, C., Li, Y., Kohli, P., Wu, J., Torralba, A., Tenenbaum, J.B.: CLEVRER: collision events for video representation and reasoning. In: ICLR (2020)
81. Yun, K., Honorio, J., Chattopadhyay, D., Berg, T.L., Samaras, D.: Two-person interaction detection using body-pose features and multiple instance learning. In: 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. pp. 28–35 (2012). https://doi.org/10.1109/CVPRW.2012.6239234
82. Zhai, Y., Liu, Z., Wu, Z., Wu, Y., Zhou, C., Doermann, D., Yuan, J., Hua, G.: Soar: Scene-debiasing open-set action recognition. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 10210–10220 (2023). https://doi.org/10.1109/ICCV51070.2023.00940

83. Zhang, B., Li, K., Cheng, Z., Hu, Z., Yuan, Y., Chen, G., Leng, S., Jiang, Y., Zhang, H., Li, X., Jin, P., Zhang, W., Wang, F., Bing, L., Zhao, D.: Videollama 3: Frontier multimodal foundation models for image and video understanding (2025), https://arxiv.org/abs/2501.13106
84. Zhang, C., Zhang, L., Wu, J., He, Y., Zhou, D.: Causal prompting: debiasing large language model prompting based on front-door adjustment. In: Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence. AAAI'25/IAAI'25/EAAI'25, AAAI Press (2025). https://doi.org/10.1609/aaai.v39i24.34777, https://doi.org/10.1609/aaai.v39i24.34777
85. Zhang, Y., Wu, J., Li, W., Li, B., Ma, Z., Liu, Z., Li, C.: Llava-video: Video instruction tuning with synthetic data (2025), https://arxiv.org/abs/2410.02713
86. Zhao, H., Yan, Z., Torresani, L., Torralba, A.: Hacs: Human action clips and segments dataset for recognition and temporal localization. arXiv preprint arXiv:1712.09374 (2019)
87. Zhao, S., Zhu, L., Wang, X., Yang, Y.: Centerclip: Token clustering for efficient text-video retrieval. In: SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, July 11–15, 2022, Madrid, Spain (2022)