VISTA: Video Interaction Spatio-Temporal Analysis Benchmark

Alejandro Aparcedo[1] Akash Kumar[1] Aaryan Garg[2] Dalton Pham[3]
Wen-Kai Chen[1] Anirudh Bharadwaj[1] Aman Chadha[4*] Yogesh Rawat[1]

[1]University of Central Florida [2]BITS Pilani
[3]Ho Chi Minh City University of Science [4]Google DeepMind
Project Page: https://aaparcedo.github.io/VISTA/

Abstract

*Existing benchmarks for Vision-Language Models (VLMs) primarily evaluate spatio-temporal understanding on simple single-action videos, closed attribute sets and restricted entity types, failing to capture the freeform, multi-action interactions between diverse entities which characterize real-world video understanding. Furthermore, the lack of a systematic framework for analyzing model failures across complementary spatio-temporal axes hinders comprehensive evaluation. To address these gaps, we introduce **VISTA**, a **V**ideo **I**nteraction **S**patio-**T**emporal **A**nalysis benchmark designed for open-set, multi-entity and multi-action spatio-temporal understanding in VLMs. VISTA decomposes videos into interpretable entities, their associated actions, and relational dynamics, enabling multi-axis diagnostics and unified assessment of relational, spatial, and temporal understanding. Our benchmark integrates multiple datasets into a single interaction-aware taxonomy and comprises ~12K curated video-query pairs spanning diverse scenes and complexities. We systematically evaluate 11 state-of-the-art VLMs on VISTA, and break down aggregate performance across our taxonomy to reveal shortcomings and pronounced spatio-temporal biases obscured by traditional metrics. By providing detailed, taxonomy-driven diagnostics on a challenging dataset, VISTA offers a nuanced framework to guide advances in model design, pretraining strategies, and evaluation protocols. Overall, VISTA is the first large-scale, interaction-aware diagnostic benchmark for spatio-temporal understanding in VLMs.*

1. Introduction

Real-world video understanding requires reasoning about complex interactions among entities over time. From pedestrian–vehicle dynamics in autonomous driving to human–human and human–object interactions in surveillance. To achieve this, intelligent visual systems must determine *which* entities interact, *how* they interact, and *where* and *when* these interactions occur. This capability, broadly referred to as spatio-temporal understanding [45, 55], extends beyond traditional object detection and motion analysis, requiring the joint modeling of spatial structure, temporal evolution, and inter-entity relationships.

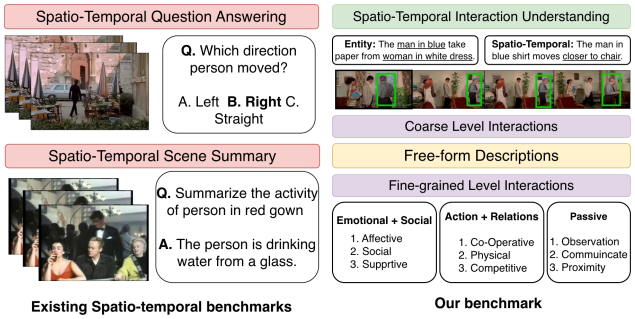


Figure 1. **VISTA vs. Existing Spatio-Temporal Benchmarks.** Existing benchmarks focus on coarse, single-step spatio-temporal understanding without localization. VISTA utilizes grounded evaluation and enables detailed analysis of multi-entity, multi-action dynamics through coarse-to-fine categorization.

Vision-Language Models (VLMs) [2, 6, 12, 37, 38, 43, 44] have significantly advanced spatio-temporal understanding by scaling architectures capable of jointly modeling visual and linguistic information. Early evaluation of these models relied on high-level VQA-style benchmarks, which were instrumental in measuring general capabilities. However, subsequent analyses [9, 22, 26, 28, 54] indicate that performance on such benchmarks can be confounded by linguistic priors limiting their ability to faithfully assess visual understanding. In response, the community has shifted toward grounded benchmarks that validate visual understanding through localization. Early efforts centered on tasks such as object tracking and action recognition, while more recent

*Work done outside role at Google DeepMind.

works [1, 29, 64, 67, 75] have introduced increasingly complex tasks involving multi-entity tracking and 4D reasoning reflecting a growing emphasis on capturing the relational dynamics underlying real-world spatio-temporal understanding. Despite this rapid progress, key limitations remain: existing benchmarks largely reduce performance to aggregate metrics, providing little insight into *where* and *why* models fail. Moreover, as model families expand, the lack of a structured evaluation framework renders consistent, fine-grained cross-model analysis increasingly intractable.

To address these limitations, we introduce **interaction** as a unifying lens for structured evaluation. Through a systematic dataset aggregation and annotation pipeline, VISTA transforms video-query pairs into a coarse-to-fine interaction-centric representation, factorized into involved entities, spatio-temporal type, and fine-grained interaction type. An overview of the differences between previous work and ours is presented in Figure 1. Our interaction-centric framework enables three diagnostic capabilities: (a) exposing hidden failure modes, as interaction-level evaluation surfaces systematic limitations masked by aggregate metrics; (b) characterizing generalization patterns, revealing how model behavior stratifies across interaction types, entity configurations, and query formulations; and (c) uncovering directional biases and tendencies, identifying consistent spatial, temporal, and semantic preferences embedded in modern VLMs.

In summary, VISTA provides the first large-scale, interaction-focused diagnostic benchmark for spatio-temporal understanding in VLMs. Our contributions are threefold:

1. **Interaction-centric diagnostic framework:** We introduce a unified coarse-to-fine evaluation taxonomy that decomposes spatio-temporal grounding into interpretable interaction types enabling principled diagnostics across ~12K video-query pairs and 11 diverse models.
2. **Systematic cross-model analysis:** By aggregating and reorganizing multiple datasets under a common interaction-aware structure, we reveal consistent stratification patterns across model families, exposing how architecture, pretraining breadth, and instruction tuning shape understanding.
3. **Bias and failure-mode characterization:** Our analysis uncovers prominent failure modes - same-entity disambiguation, linguistic template preferences, and semantic-intent inflation - offering the first interaction-grounded view of systematic reasoning failures in modern VLMs.

2. Related Work

VLM Benchmarks: The rapid progress of VLMs has been paralleled by increasingly sophisticated benchmarks designed to probe spatio-temporal understanding [45, 55]. From early datasets that evaluate general visual understanding [15, 26, 27, 49, 50, 69], to fine-grained spatio-temporal localization tasks [3, 24, 32, 48, 51, 52, 59, 63, 68]. Unlike general video understanding benchmarks [5, 23, 35, 46, 74] that assess abstract comprehension, spatio-temporal benchmarks emphasize grounded reasoning. Within the segmentation community, MOSE [19] introduced crowded, heavily occluded scenes where targets frequently disappear and reappear, revealing that state-of-the-art VOS methods are brittle under such conditions. Its successor MOSEv2 [21] extends this further with adverse weather, low-light environments, camouflaged objects, and non-physical targets. On the language-guided side, MeViSv2 [20] shifts the focus from static-attribute referring expressions to *motion*-based descriptions that require genuine temporal reasoning across frames, supporting multi-target and no-target expressions. In the detection-style grounding setting, Spatio-Temporal Video Grounding (STVG) [60, 73] requires joint localization of entities across space and time from freeform relational queries, with recent efforts [1, 29, 64, 67, 75] further broadening this toward 4D reasoning, multi-object grounding, and grounded captioning. Yet across both settings, benchmarks largely reduce performance to aggregate metrics, providing little insight into where and why models fail. While prior diagnostic efforts [22, 39] shed light on performance across coarse-grained spatial and temporal categories, they neglect the intricate interaction semantics that critically influence spatio-temporal behavior. VISTA complements segmentation benchmarks by adopting STVG as its diagnostic probe, enabling structured evaluation of *how* and *why* models fail across diverse interaction types, entity configurations, and query formulations—dimensions that mask-based benchmarks do not directly expose.

Spatio-temporal understanding in VLMs: Early spatio-temporal understanding modeled space and time independently under closed-set conditions - spatial models [10, 53] handled object detection within fixed categories while temporal models [11, 58] targeted action recognition under constrained label settings. Subsequent work advanced vision-language alignment across both spatial and temporal dimensions - through OVD [47, 68] and REC [16], culminating in strong detectors such as GLIP [36] and Grounding-DINO [43], while parallel progress in Moment Localization [3, 24] enabled language-guided temporal understanding [2]. The integration of LLMs into VLMs [6, 34, 41] further strengthened multimodal grounding, and video-centric extensions [4, 5, 37, 70] introduced joint spatial-temporal understanding. Despite evaluating increasingly complex spatio-temporal tasks, existing benchmarks reduce performance of VLMs to aggregate metrics, leaving the failure modes, biases, and systematic tendencies of modern VLMs largely undiagnosed.

Approach	Encoders		VISTA			Spatio-Temp.		Entity							
	Image	Text	R	F	R&F	S	T	AA	AO	HA	HH	HO	HS	NI	OO
Foundation Model w/o LLMs															
GDINO [42]	Swin-T	BERT	37.79	32.34	34.64	35.0	30.8	12.6	38.8	52.1	29.9	37.0	39.4	41.3	38.3
Generalist MLLMs															
Intern-VL 2.5 [17]	InternViT-300M	InternLM2-7B†	51.11	48.65	49.73	46.3	<u>48.0</u>	37.9	49.8	52.2	<u>49.2</u>	49.5	50.9	48.2	47.7
Mini-GPT-v2 [12]	EVA-CLIP ViT-G/14	LLaMA2-7B†	46.62	45.13	45.78	43.1	44.3	33.6	47.4	48.5	46.0	46.1	46.4	44.3	43.1
Sphinx-v2 [40]	CLIP ViT-L/14	LLaMA2-7B	47.79	44.28	45.82	42.6	45.0	30.4	46.9	51.0	47.0	46.8	48.1	46.0	42.5
Qwen-VL-Chat [7]	ViT-bigG	Qwen-7B	45.56	45.43	45.49	45.7	45.3	33.7	54.8	65.2	48.2	47.9	<u>58.8</u>	42.2	31.7
Qwen3-VL [8]	SigLIP-2	Qwen3-8B	**62.85**	**64.41**	**63.96**	**64.8**	**64.3**	**59.5**	**63.2**	**74.5**	**66.2**	**64.7**	**75.7**	**60.6**	**59.1**
MimoVL [62]	Qwen2.5-ViT	MiMo-7B-Base	43.34	42.13	44.54	36.9	43.5	40.4	38.3	38.3	45.1	36.1	46.0	43.0	27.0
Specialist MLLMs															
Shikra [13]	CLIP-ViT-L/14	Vicuna-1/7B	30.91	31.44	31.21	29.9	32.4	20.0	28.9	36.0	34.0	35.3	38.8	31.4	24.6
Ferret-v1 [66]	CLIP-ViT-L/14	Vicuna-1.3/7B	17.74	22.71	20.53	20.9	23.8	14.9	23.7	23.3	26.4	24.5	33.6	19.7	13.4
CogVLM‡ [61]	EVA2-CLIP-E	Vicuna-1.5/7B	<u>60.56</u>	<u>50.13</u>	<u>54.70</u>	<u>57.5</u>	45.7	<u>48.1</u>	<u>60.3</u>	<u>70.2</u>	44.7	<u>54.0</u>	46.8	<u>50.7</u>	<u>54.0</u>
LLAVA-G [72]	CLIP-ViT-L/14	Vicuna-1.3/7B	22.51	30.47	27.11	28.1	31.9	10.7	37.1	56.1	28.4	37.0	38.9	36.0	28.4

Table 1. Main results on VISTA. Referral and freeform query performance is denoted with R and F, respectively. † and ‡ denotes chat and grounding versions, respectively. **Bold** and <u>underline</u> indicate the best and second best results, respectively.

3. The VISTA Benchmark

Problem Formulation: In VISTA, the input comprises a trimmed video $V = (v_1, v_2, \ldots, v_T)$ with T frames and a descriptive query caption Q that specifies the primary subject and activity within the video. The objective is to accurately localize the mentioned subject (A_R) in all T frames, thereby forming a spatio-temporal tubelet denoted as $A_R = \{a_r\}_{t_1}^{t_T}$, where a_r represents the bounding-box for the subject in the r-th frame.

3.1. VISTA Taxonomy

Motivation: Despite extensive evaluation of VLMs on spatio-temporal tasks, model failure modes remains largely a black box: aggregate metrics conflate failures across fundamentally different understanding demands, making it impossible to distinguish whether a model struggles with entity identification, spatial grounding, or temporal reasoning. Our taxonomy addresses this by breaking down evaluation into structured, interpretable categories across two complementary levels: **coarse-grained**, capturing *who* interacts and *where* interactions unfold across space and time, and **fine-grained**, characterizing the specific relational and behavioral dynamics observed in daily activities. Critically, by stratifying performance across taxonomy categories rather than reporting a single aggregate score, consistent failure patterns and systematic model tendencies become directly visible.

Coarse-grained analysis comprises two axes: **(a) Involved Entities** categorizes interactions based on the involved participants among *humans (H)*, *animals (A)*, and *objects (O)*, capturing all six pairwise configurations: *HH, HA, HO, AA, AO, OO*, augmented with *Human-Self (HS)* for solitary actions and *No Interaction (NO)*. **(b) Spatio-Temporal Interaction** classifies samples by their primary understanding demand: ***spatial*** samples focus on positional configurations among entities (e.g., `"the person beside the car"`), while ***temporal*** samples capture entity state transitions over time (e.g., `"the woman sitting down after standing"`).

However, coarse categories alone cannot capture the semantic diversity within each bucket - a Human-Human, spatial query may demand relative-position understanding (e.g., `"the man standing behind the woman"`) or social understanding (e.g., `"the person comforting the other"`), distinctions a flat taxonomy cannot surface. **Fine-grained** analysis addresses this across three thematic groups: **(a) Emotional and Social**: *Affective (AFF)*, *Social (SOC)*, and *Supportive (SUP)* capture emotion, bonding, and assistance; **(b) Physical and Action-Oriented:** *Physical (PHY)*, *Relational Movement (RM)*, *Cooperative (COP)*, *Competitive (CMP)*, and *Antagonistic (ANT)* describe contact, motion, and joint or opposing effort; **(c) Observational and Passive:** *Observation (OBS)*, *Communicative (COM)*, *Proximity (PRX)*, *Body Motion (BM)*, *Provisioning (PRV)*, and *Passive (PAS)* reflect non-contact, attention, and static states - together spanning the spectrum of social, physical, and cognitive behavior. The complete taxonomy class distribution can be seen in Figure 3a.

3.2. Dataset Collection

Motivation: Prior benchmarks [64, 67, 75] have closed object and action vocabularies that restrict evaluation to predefined categories, and templated queries [67, 75] capturing

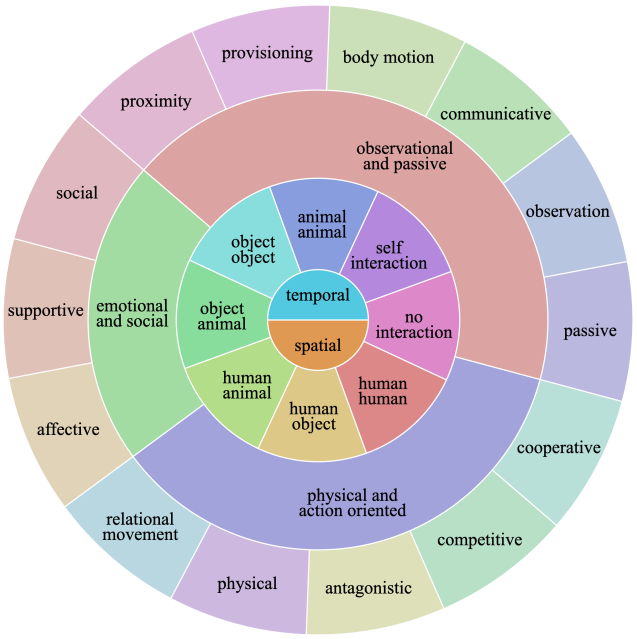
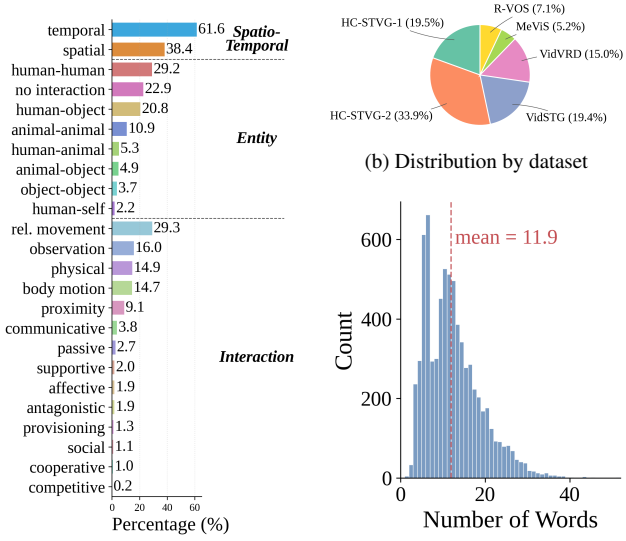


Figure 2. **Taxonomy** of VISTA benchmark. The two inner circles represent coarse-grained categories, while the outermost circle illustrates the distribution of fine-grained categories.



(a) Taxonomy class distribution

(b) Distribution by dataset

(c) Distribution of caption lengths

Figure 3. Statistical analysis of VISTA Benchmark.

only single-step facts that fail to probe the compositional, multi-entity interactions characteristic of real-world video. Our dataset aggregation addresses both - spanning from simple, well-known concepts [56, 57] to fully open-world, complex relational queries [18, 60, 73], with expression styles ranging from template-based to freeform.

Dataset Curation: To build a comprehensive benchmark covering the complexity of spatio-temporal understanding, we aggregate and reformulate six datasets: HCSTVG-v1 and v2 [60], VidVRD [57], VidSTG [73], MeViS [18], and RVOS [56]. From a language perspective, these datasets span diverse query lengths, reasoning complexity, and expression styles (template-based to freeform). Visually, these datasets span a wide variety of scenes encompassing diverse environments, perspectives and visual challenges such as camera motion, occlusion, and complex object interactions.

Query Formulation: A core component of our benchmark is the explicit evaluation of human-style narrative queries (freeform) versus template-based queries (referral). Freeform queries capture open-ended, conversational descriptions, while referral queries focus on concise, object-centric expressions.

- **Freeform Queries (Q_F):** We use freeform captions provided by datasets directly, or reformulate relation triplets (e.g., $\langle$subject, predicate, object$\rangle$) into freeform natural language sentences through LLMs. Freeform queries capture the full activity and relational context, e.g., `"A man in a suit walks into the room and sits down"`.
- **Referral Queries (Q_R):** Derived by prompting an LLM to extract the primary subject and its attributes from a freeform query. Using the same example, `"A man in a suit walks into the room and sits down"` reduces to `"A man in a suit"` - retaining only entity identity and attributes, discarding relational and temporal context entirely.

Sample Annotation Pipeline: For taxonomy classification, we focus exclusively on freeform captions (Q_F), which contain the complex relational and spatio-temporal descriptions necessary to assign meaningful interaction categories. We employ a multi-stage pipeline leveraging `gpt-4o-mini` to classify each caption $qf \in Q_F$ - assigning a single coarse category for involved entities and spatio-temporal interaction, while annotating fine-grained categories exhaustively due to caption complexity. A manual review round was conducted after each classification step to verify and refine labels. Additional implementation details are provided in the supplementary material.

Annotation Quality: To validate annotation pipeline reliability, we conducted an inter-annotator agreement study on $n = 113$ stratified samples using 2 human annotators and `gpt-4o-mini`. Cohen's κ scores are reported for all three taxonomy levels below.

Level	H-H κ	H-GPT κ
Entity	0.98	0.76
Spatio-Temporal	0.77	0.69
Fine-grained	0.83	0.67

Human-human agreement ($\kappa = 0.77 - 0.98$) indicates substantial to almost perfect agreement [33], con-

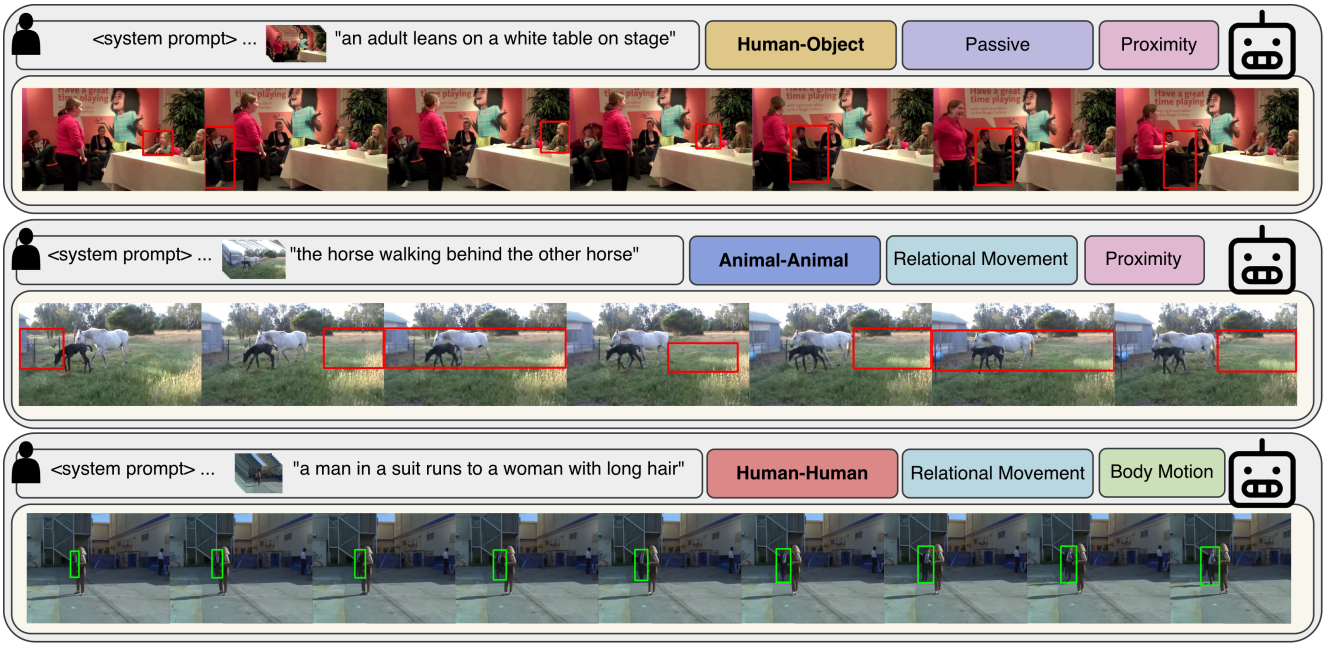


Figure 4. Examples of good (mvIoU > 0.8) and bad (mvIoU < 0.4) spatio-temporal grounding capabilities across VISTA on the best performing model: CogVLM.

firming the taxonomy is well-defined and consistently interpretable across annotators. Human-GPT agreement is moderate ($\kappa = 0.67 - 0.76$), with discrepancies concentrated in visually ambiguous or linguistically underspecified captions - for instance, "`bear cubs in tow, big bear crossing road`" (GPT: Human-Animal, Corrected: Animal-Animal) and "`fat man takes out his gun`" (GPT: No Interaction, Corrected: Human-Object). These errors directly motivate the manual verification step in our annotation pipeline.

Benchmark Stats: Our benchmark comprises 11,814 unique video–caption pairs (V, Q), offering a rich set of fine-grained annotations. Textual descriptions range between 40-60 words on average, reflecting the complexity of the freeform language used. Video resolution and number of frames are approximately 866×544 pixels and 174 frames, respectively. This combination of detailed spatio-temporal annotations, realistic video lengths, and varied scene content distinguishes our benchmark from existing datasets. The distributions by datasets and description length are illustrated in Figure 3b and Figure 3c, respectively. The fine-grained distribution reflects the organic frequency of interaction types in natural video: Competitive (0.2%) and Cooperative (1.0%) are genuinely rare while Relational Movement (29.3%) and Observation (16.0%) are not. Fine-grained analysis spans all categories for diagnostic breadth; quantitative conclusions are restricted to categories with sufficient sample support.

3.3. Evaluation Setup

Benchmark Models: Building on prior work [63], we select a representative set of models capturing diversity in architecture (LLM-based vs. non-LLM-based), training paradigm, and task specialization, as these factors naturally influence grounding capabilities. A fundamental requirement for inclusion is the ability to produce structured bounding-box predictions necessary for IoU-based evaluation - while powerful model families like GPT, Gemini, and VideoLLaMA demonstrate spatio-temporal understanding, they are not explicitly trained for fine-grained localization, making reliable IoU-based assessment infeasible. We organize selected models into three categories: **(1) Foundation Models without LLMs**, **(2) Generalist MLLMs**, and **(3) Specialist MLLMs**. Category (1) includes Grounding-DINO [42] for its strong zero-shot detection generalizability. Category (2) comprises Intern-VL 2.5 [17], Mini-GPT-v2 [12], Sphinx-v2 [40], Qwen-VL-Chat [7], and Qwen3-VL [8] - LLM-backed models trained on diverse tasks including localization. Category (3) represents task-specific models trained exclusively for detection and related localization tasks and includes Shikra [13], Ferret [66], and CogVLM [61] which generate bounding boxes in plain text, and LLaVA-Grounding [71] which combines a LLM with a dedicated detection head. All models are evaluated zero-shot on sub-sampled video frames. Further details are in the supplementary.

Model Selection Rationale. We prioritize open-weight models for two reasons: (1) *Reproducibility* - proprietary models

such as GPT-4o undergo silent updates that can substantially alter behavior between evaluation runs [14], undermining the diagnostic consistency central to VISTA's contribution; and (2) *Cost* - systematic evaluation across ∼12K video–query pairs with multi-frame sampling is prohibitively expensive through commercial APIs. We note that VISTA's evaluation framework and taxonomy are model-agnostic and directly applicable to proprietary or future models as access constraints evolve.

Data Contamination: We cross-referenced all VISTA video identifiers against the disclosed training splits of all evaluated models, finding no overlaps. Full decontamination remains infeasible given incomplete disclosure of web-scale pretraining corpora; however, our analyses focus on intra-model performance stratification rather than absolute scores. Relative patterns such as the cross-entity vs. same-entity gap are robust to incidental exposure, as contamination inflates scores uniformly across categories.

Evaluation Metrics: We report performance using metrics established in previous studies [25, 30, 31, 65]: mean spatio-temporal IoU (m_vIoU) which is computed as $\frac{1}{|S_u|}\sum_{t\in S_i}\text{IoU}(\hat{b_t}, b_t)$, where S_i and S_u denote the intersection and union, respectively, between the predicted and ground truth timestamps. $\text{IoU}(\hat{b}_t, b_t)$ represents the spatial overlap between the predicted bounding box $\hat{b}_t$ and the ground truth box b_t at frame t.

4. Directional Biases in Interactions

We evaluate and analyze the relative performance of models across **VISTA**'s hierarchical taxonomy, examining differences both intra and inter model families. Our analysis proceeds along three axes: query structure (referral vs. freeform), coarse-grained analysis, fine-grained analysis. Across this taxonomy, several trends emerge. Model performance follows a clear family-level hierarchy, with Generalist MLLMs outperforming both Specialist MLLMs and Foundation Models. More notably, models exhibit a consistent sensitivity to query structure, performing substantially better on referral than freeform queries across all families - indicating continued reliance on syntactic scaffolding over genuine multimodal reasoning. At the interaction level, same-entity interactions reveal systematic symmetry failures, while the relatively balanced performance across spatial and temporal samples stands in contrast to the static, image-based nature of most model training. Beyond these trends, we examine the directional tendencies underlying these failures - reasoning about how pretraining distributions, cross-modal alignment, and architectural choices systematically shape model interpretation. Nearly all reported performance differences are statistically significant ($p < 0.05$); bootstrap confidence intervals and full hypothesis test details are provided in the supplementary.

4.1. Impact of Query Structure

Table 1, reveals a robust and repeatable pattern across all model families: *models perform better on referral (template-like) queries than free-form (natural language) queries*. This gap indicates that models are sensitive to prompt structure - leveraging syntactic cues such as `<subject, verb, object>` ordering when present, but failing to compensate through multimodal reasoning when they are not. This failure mode is exemplified in Figure 4 (top), where, given the query "`an adult leans on a white table on stage`" the model successfully grounds "`an adult`" but fails to leverage the spatial cue "`on the table`" to recognize that the subject is actually a child. This highlights how models prioritize syntactic patterns over spatial reasoning cues embedded in natural language. More broadly, pre-training breadth shapes this gap directly: models trained on heterogeneous, interaction-rich corpora maintain more stable R-F balance, while those fine-tuned on narrow domains or static captions degrade under freeform settings, overfitting to surface co-occurrence statistics rather than learning compositional reasoning. A notable exception is Qwen3-VL, the strongest model overall, which reverses this trend with freeform queries (64.41) outperforming referral queries (62.85), suggesting that sufficient pretraining breadth and instruction diversity can enable models to exploit richer context in freeform descriptions rather than relying on syntactic scaffolding. MimoVL achieves competitive generalist performance but exhibits a pronounced same-entity deficit (OO: 27.0 vs. HO: 46.0), consistent with the broader disambiguation failures identified across models. A subset of Specialist MLLMs exhibit marginal gains on freeform inputs, indicating that *additional linguistic context can be beneficial when it aligns with a model's training distribution*.

4.2. Coarse Grained Category Analysis

Coarse-grained interactions are analyzed along two axes: involved entities and spatio-temporal interaction type as per taxonomy sub-divisions.

Involved Entities: Figure 5(a) reveals a clear pattern across models: *interactions that cross entity categories score substantially higher than same-entity interactions*. Averaging across models, *Human-Animal (HA) interactions are the strongest (51.6% avg.)*, while *Animal-Animal (AA) interactions are the weakest (31.1% avg.)*, and Object-Object (OO) interactions are also relatively low (37.3% avg.). This reflects a prevalent failure mode rooted in *category-level priors*: models can more effectively ground entities when they belong to different semantic classes, but struggle to disambiguate visually similar instances of the same class, instead defaulting to general entity recognition rather than leveraging specific referential cues. This pattern persists even in high-performing Generalist MLLMs, indicating that representational homogeneity, rather than limited capacity, drives

these errors. Figure 4 (middle) illustrates this symmetry failure explicitly: given the query "the horse walking behind the other horse", CogVLM fails to resolve the spatial relationship "behind", defaulting to grounding an entity of the correct class rather than the specific one requested. This contrasts with Figure 4 (bottom), where given "a man in a suit runs to a woman with long hair", the model successfully grounds "the man in a suit". Although this is also a same-entity (Human-Human) interaction, the entities are visually distinct and described by their attributes ("in a suit," "with long hair") rather than a complex spatial relation - confirming that the core failure lies in lack of reasoning about relational and spatial cues when visual distinctiveness between entities is low.

Spatio-Temporal Interaction: Performance across spatial (S) and temporal (T) samples is roughly comparable across all model families, *suggesting no strong global bias for one axis* – a finding that runs counter to expectations, given the predominantly static, image-based training of most architectures. Examining this by model family reveals an interesting split: foundation and specialist models conform to the expected spatial bias, remaining anchored to static appearances consistent with their training. LLM-based generalist models exhibit near-parity between spatial and temporal performance, suggesting that jointly decoding over language and vision features helps compensate for challenging temporal visual conditions such as motion blur or occlusion.

These coarse-level trends highlight two complementary limitations in current VLMs. First, grounding performance is strongly conditioned on visual and semantic distinctiveness between entities - models succeed when category or appearance differentiates the referent, but fail when disambiguation requires relational or spatial reasoning. Second, while surface-level spatial and temporal performance appears balanced, this masks an underlying preference for static configurations: models struggle with causality, motion sequences, and transitions in visually ambiguous scenes.

4.3. Fine Grained Category Analysis

Fine-grained interactions reveal deeper insights into how models handle interpersonal, physical, and non-contact nuances beyond coarse entity and space–time reasoning. Figure 5(b) shows that Generalist MLLMs consistently achieve the highest scores, while Specialist MLLMs and Foundation Models exhibit sharp variability depending on the type of interaction. A key trend is that *models perform substantially better on interactions with clear visual anchors* (e.g., physical interaction, supportive, social) and struggle when the interaction requires *implicit cognition, emotional inference, passive states*.

(a) Emotional and Social: Models show moderate performance on affective, social, and supportive interactions, yet these categories remain consistently among the weakest across all model families, indicating that *MLLMs lack robust grounding for subtle emotional or interpersonal behaviors*, particularly when cues are indirect or language-driven. This weakness is compounded by a broader tendency we term *semantic-intent inflation*: instruction-tuned and generalist MLLMs systematically over-interpret scenes through high-intent or affective frames, projecting social and emotional significance onto interactions even when the visual evidence supports simpler physical or positional readings. Figure 4 (top) illustrates this directly: given "an adult leans on a white table on stage," the model grounds "an adult" but fails to leverage the spatial cue "on the table" to recognize that the subject is a child. Rather than parsing the **Proximity (PRX)** and **Passive (PAS)** nature of the interaction, the model defaults to a socially inflated interpretation anchored in the referral term. This pattern reflects pretraining distributions and instruction tuning that emphasize conversational and affective content, systematically biasing models toward semantic over-attribution even at the cost of spatial and relational accuracy.

(b) Physical and Action-Oriented: These interactions yield the strongest performance overall, particularly for generalist models, as they involve motion, contact, or clear physical consequences that provide salient visual anchors. Yet even within this group, important distinctions emerge: cooperative and physical interactions benefit from visually structured cues, while *competitive and antagonistic actions remain harder to disambiguate*, requiring models to distinguish between semantically similar but directionally opposed dynamics. Moreover, kinematic and relational reasoning breaks down even when the broader category is favorable. Figure 4 (middle) illustrates this: given "the horse walking behind the other horse," CogVLM fails to parse "behind" as a **Relational Movement (RM)** relationship, defaulting to entity recognition rather than modeling the directional dynamic between two visually similar entities. This failure reveals that strong aggregate performance on physical interactions masks a specific deficit in directional and motion-based reasoning. *Models can leverage visual salience when interactions produce observable consequences*, but struggle when grounding depends on parsing the spatial trajectory or relative motion between entities rather than identifying the entities themselves.

(c) Observational and Passive: Performance splits sharply within this group. Interactions with explicit visual cues, such as proximity and body motion, are handled reasonably well, whereas passive or cognitive categories such as observation and provisioning remain challenging, as they require inferring intent, attention, or perspective from subtle or absent visual signals. This difficulty exposes a systematic tendency we term *social-first bias*: when an interaction contains both social and physical signals, models interpret it

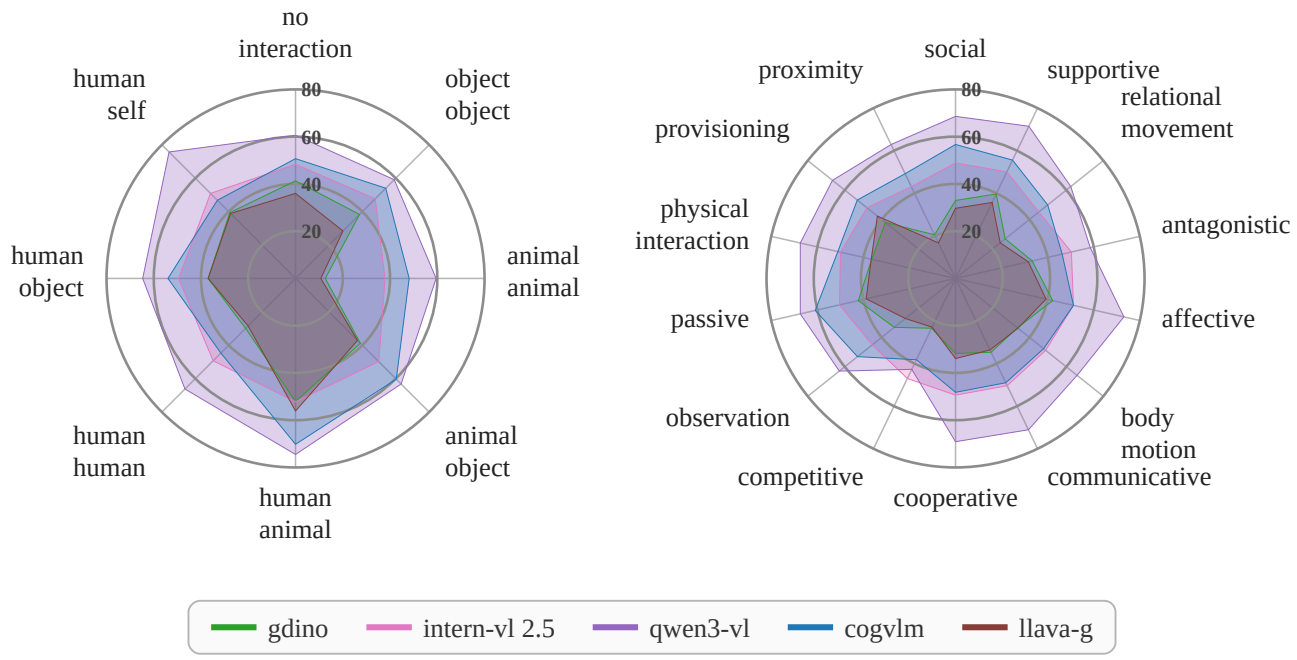


Figure 5. Per-model mvIoU across (left) coarse-grained entity-pair categories and (right) fine-grained interaction types. Cross-entity pairs (e.g., Human-Animal) consistently outperform same-entity pairs (e.g., Animal-Animal), while interactions with salient visual cues (e.g., relational movement) yield stronger performance than those requiring implicit reasoning (e.g., passive, affective).

primarily through the lens of identity and affect, often at the expense of physical dynamics. Figure 4 (bottom) exemplifies this: given "`a man in a suit runs to a woman with long hair,`" the model successfully grounds the correct entity, but its success stems from over-reliance on distinctive textual attributes ("`in a suit,`" "`with long hair`") rather than genuine understanding of the underlying **Body Motion (BM)** and **Relational Movement (RM)**. The model treats the interaction as an identity-matching problem rather than a kinematic one. This shallow grounding strategy generalizes across architectures: *even when models produce correct predictions on observational or passive interactions, the reasoning pathway frequently bypasses the physical and attentional cues* that define the category.

Across all three fine-grained groups, a consistent pattern emerges: current VLMs are more adept at reasoning about *why* interactions occur than *how* they unfold or *where* they are situated. Semantic-intent inflation and social-first bias are complementary manifestations of the same underlying gap. Alignment strategies, instruction tuning, and pretraining distributions have successfully taught models to emphasize semantic content, particularly social and affective aspects, but have not sufficiently reinforced the modeling of physical motion, relational dynamics, or subtle spatial states. Correcting this imbalance will require integrating datasets with complex kinematic cues and multi-agent dynamics, alongside explicit grounding tasks that force models to jointly reason about social intent, temporal evolution, and spatial configuration.

5. Conclusion

In this work, we introduce **VISTA**, a benchmark for evaluating fine-grained, interaction-centric spatio-temporal reasoning in Vision-Language Models (VLMs). By unifying multiple datasets into a single interaction-aware taxonomy and decomposing videos into coarse- and fine-grained interaction types, VISTA enables a critically nuanced evaluation of spatial, temporal, and relational understanding. Our framework not only exposes hidden model weaknesses masked by aggregate metrics but also characterizes generalization patterns and uncovers directional, spatial, and temporal biases across a broad range of state-of-the-art models. Through extensive experiments over several modern MLLMs, we demonstrate that even high-performing models exhibit limitations in multi-entity, multi-action, and temporally compositional reasoning, with same-entity disambiguation and semantic-intent inflation emerging as the two most critical bottlenecks. These findings suggest that targeted training on visually similar multi-instance scenes and kinematic reasoning tasks may yield the largest gains. Overall, VISTA provides a first systematic lens for diagnosing these limitations, bridging the gap between abstract-level assessment and robust, real-world video understanding.

References

[1] Ghazi Shazan Ahmad, Ahmed Heakl, Hanan Gani, Abdelrahman Shaker, Zhiqiang Shen, Fahad Shahbaz Khan, and Salman Khan. Videomolmo: Spatio-temporal grounding meets pointing. *arXiv preprint arXiv:2506.05336*, 2025. 2

[2] Shahzad Ahmad, Sukalpa Chanda, and Yogesh S Rawat. T2l: Efficient zero-shot action recognition with temporal token learning. *Transactions on Machine Learning Research*, 2025. 1, 2

[3] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812, 2017. 2

[4] Shehreen Azad, Vibhav Vineet, and Yogesh Singh Rawat. Hierarq: Task-aware hierarchical q-former for enhanced video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8545–8556, 2025. 2

[5] Shehreen Azad, Vibhav Vineet, and Yogesh Singh Rawat. Streamready: Learning what to answer and when in long streaming videos. *Proceedings of the IEEE/CVF international conference on computer vision*, 2026. 2

[6] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 1, 2

[7] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond, 2024. 3, 5

[8] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025. 3, 5

[9] Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*, 2024. 1

[10] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 2

[11] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 2

[12] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechu Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023. 1, 3, 5

[13] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023. 3, 5

[14] Lingjiao Chen, Matei Zaharia, and James Zou. How is ChatGPT's behavior changing over time? *arXiv preprint arXiv:2307.09009*, 2023. 6

[15] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C. Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server, 2015. 2

[16] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[17] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 3, 5

[18] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Chen Change Loy. Mevis: A large-scale benchmark for video segmentation with motion expressions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2694–2703, 2023. 4

[19] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, Philip HS Torr, and Song Bai. MOSE: A new dataset for video object segmentation in complex scenes. In *ICCV*, 2023. 2

[20] Henghui Ding, Chang Liu, Shuting He, Kaining Ying, Xudong Jiang, Chen Change Loy, and Yu-Gang Jiang. MeViS: A multi-modal dataset for referring motion expression video segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 2

[21] Henghui Ding, Kaining Ying, Chang Liu, Shuting He, Xudong Jiang, Yu-Gang Jiang, Philip HS Torr, and Song Bai. MOSEv2: A more challenging dataset for video object segmentation in complex scenes. *arXiv preprint arXiv:2508.05630*, 2025. 2

[22] Bo Feng, Zhengfeng Lai, Shiyu Li, Zizhen Wang, Simon Wang, Ping Huang, and Meng Cao. Breaking down video llm benchmarks: Knowledge, spatial perception, or true temporal understanding? *arXiv preprint arXiv:2505.14321*, 2025. 1, 2

[23] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang

Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24108–24118, 2025. 2

[24] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, pages 5267–5275, 2017. 2

[25] Aaryan Garg, Akash Kumar, and Yogesh S Rawat. Stpro: Spatial and temporal progressive learning for weakly supervised spatio-temporal grounding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3384–3394, 2025. 6

[26] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 1, 2

[27] Shresth Grover, Vibhav Vineet, and Yogesh Rawat. Revealing the unseen: Benchmarking video action recognition under occlusion. *Advances in Neural Information Processing Systems*, 36:65642–65664, 2023. 2

[28] Shresth Grover, Vibhav Vineet, and Yogesh S Rawat. Navigating hallucinations for reasoning of unintentional activities. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9666–9680, 2024. 1

[29] Tanveer Hannan, Shuaicong Wu, Mark Weber, Suprosanna Shit, Jindong Gu, Rajat Koner, Aljoša Ošep, Laura Leal-Taixé, and Thomas Seidl. Svag-bench: A large-scale benchmark for multi-instance spatio-temporal video action grounding. *arXiv preprint arXiv:2510.13016*, 2025. 2

[30] Yang Jin, Yongzhi Li, Zehuan Yuan, and Yadong Mu. Embracing consistency: A one-stage approach for spatio-temporal video grounding. *ArXiv*, abs/2209.13306, 2022. 6

[31] Akash Kumar, Zsolt Kira, and Yogesh Singh Rawat. Contextual self-paced learning for weakly supervised spatio-temporal video grounding. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025. 6

[32] Akash Kumar, Sirshapan Mitra, and Yogesh Singh Rawat. Stable mean teacher for semi-supervised video action detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4419–4427, 2025. 2

[33] J Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33 1: 159–74, 1977. 4

[34] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 2

[35] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206, 2024. 2

[36] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10965–10975, 2022. 2

[37] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5971–5984, 2024. 1, 2

[38] Fangjian Lin, Jianlong Yuan, Sitong Wu, Fan Wang, and Zhibin Wang. Uninext: Exploring a unified architecture for vision recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*, page 3200–3208, New York, NY, USA, 2023. Association for Computing Machinery. 1

[39] Yuxiang Lin, Ling Luo, Ying Chen, Xushi Zhang, Zihui Wang, Wenxian Yang, Mengsha Tong, and Rongshan Yu. St-align: A multimodal foundation model for image-gene alignment in spatial transcriptomics. *arXiv preprint arXiv:2411.16793*, 2024. 2

[40] Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, Jiaming Han, Siyuan Huang, Yichi Zhang, Xuming He, Hongsheng Li, and Yu Qiao. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models, 2023. 3, 5

[41] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 2

[42] Siyi Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chun yue Li, Jianwei Yang, Hang Su, Jun-Juan Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *ArXiv*, abs/2303.05499, 2023. 3, 5

[43] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pages 38–55. Springer, 2024. 1, 2

[44] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12585–12602, 2024. 1

[45] Neelu Madan, Andreas Møgelmose, Rajat Modi, Yogesh S Rawat, and Thomas B Moeslund. Foundation models for video understanding: A survey. *arXiv preprint arXiv:2405.03770*, 2024. 1, 2

[46] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244, 2023. 2

[47] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh

Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. Simple open-vocabulary object detection. In *European conference on computer vision*, pages 728–755. Springer, 2022. 2

[48] Rajat Modi, Aayush Jung Rana, Akash Kumar, Praveen Tirupattur, Shruti Vyas, Yogesh Rawat, and Mubarak Shah. Video action detection: Analysing limitations and challenges. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop*, pages 4911–4920, 2022. 2

[49] Rajat Modi, Vibhav Vineet, and Yogesh Rawat. On occlusions in video action detection: Benchmark datasets and training recipes. *Advances in Neural Information Processing Systems*, 36:57306–57335, 2023. 2

[50] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015. 2

[51] Aayush Rana, Akash Kumar, Vibhav Vineet, and Yogesh S Rawat. Omvid: Omni-supervised active learning for video action detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshop*, pages 6911–6921, 2025. 2

[52] Yogesh Singh Rawat and Aayush Jung Bahadur Rana. Active sparse labeling of video frames, 2025. US Patent App. 18/667,244. 2

[53] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2017. 2

[54] Madeline Schiappa, Shruti Vyas, Hamid Palangi, Yogesh Rawat, and Vibhav Vineet. Robustness analysis of video-language models against visual and language perturbations. *Advances in Neural Information Processing Systems*, 35: 34405–34420, 2022. 1

[55] Madeline C Schiappa, Yogesh S Rawat, and Mubarak Shah. Self-supervised learning for videos: A survey. *ACM Computing Surveys*, 55(13s):1–37, 2023. 1, 2

[56] Seonguk Seo, Joon-Young Lee, and Bohyung Han. Urvos: Unified referring video object segmentation network with a large-scale benchmark. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV*, page 208–223, Berlin, Heidelberg, 2020. Springer-Verlag. 4

[57] Xindi Shang, Tongwei Ren, Jingfan Guo, Hanwang Zhang, and Tat-Seng Chua. Video visual relation detection. In *ACM International Conference on Multimedia*, Mountain View, CA USA, 2017. 4

[58] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, 27, 2014. 2

[59] Ayush Singh, Aayush J Rana, Akash Kumar, Shruti Vyas, and Yogesh Singh Rawat. Semi-supervised active learning for video action detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4891–4899, 2024. 2

[60] Zongheng Tang, Yue Liao, Si Liu, Guanbin Li, Xiaojie Jin, Hongxu Jiang, Qian Yu, and Dong Xu. Human-centric spatio-temporal video grounding with visual transformers. *IEEE Transactions on Circuits and Systems for Video Technology*, 32:8238–8249, 2020. 2, 4

[61] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Song XiXuan, et al. Cogvlm: Visual expert for pretrained language models. *Advances in Neural Information Processing Systems*, 37:121475–121499, 2024. 3, 5

[62] LLM-Core-Team Xiaomi. Mimo-vl technical report, 2025. 3

[63] Chi Xie, Zhao Zhang, Yixuan Wu, Feng Zhu, Rui Zhao, and Shuang Liang. Described object detection: Liberating object detection with flexible expressions. *Advances in Neural Information Processing Systems*, 36:79095–79107, 2023. 2, 5

[64] Yunqiu Xu, Linchao Zhu, and Yi Yang. Mc-bench: A benchmark for multi-context visual grounding in the era of mllms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17675–17687, 2025. 2, 3

[65] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Tubedetr: Spatio-temporal video grounding with transformers. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16421–16432, 2022. 6

[66] Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704*, 2023. 3, 5

[67] Yuqian Yuan, Hang Zhang, Wentong Li, Zesen Cheng, Boqiang Zhang, Long Li, Xin Li, Deli Zhao, Wenqiao Zhang, Yueting Zhuang, et al. Videorefer suite: Advancing spatial-temporal object understanding with video llm. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18970–18980, 2025. 2, 3

[68] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14393–14402, 2021. 2

[69] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6720–6731, 2019. 2

[70] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023. 2

[71] Hao Zhang, Hongyang Li, Feng Li, Tianhe Ren, Xueyan Zou, Shilong Liu, Shijia Huang, Jianfeng Gao, Leizhang, Chunyuan Li, et al. Llava-grounding: Grounded visual chat with large multimodal models. In *European Conference on Computer Vision*, pages 19–35. Springer, 2024. 5

[72] Hao Zhang, Hongyang Li, Feng Li, Tianhe Ren, Xueyan Zou, Shilong Liu, Shijia Huang, Jianfeng Gao, Chunyuan Li, Jainwei Yang, et al. Llava-grounding: Grounded visual chat with large multimodal models. In *European Conference on Computer Vision*, pages 19–35. Springer, 2025. 3

[73] Zhu Zhang, Zhou Zhao, Yang Zhao, Qi Wang, Huasheng Liu, and Lianli Gao. Where does it exist: Spatio-temporal video

grounding for multi-form sentences. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10665–10674, 2020. 2, 4

[74] Yilun Zhao, Haowei Zhang, Lujing Xie, Tongyan Hu, Guo Gan, Yitao Long, Zhiyuan Hu, Weiyuan Chen, Chuhan Li, Zhijian Xu, et al. Mmvu: Measuring expert-level multi-discipline video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8475–8489, 2025. 2

[75] Shijie Zhou, Alexander Vilesov, Xuehai He, Ziyu Wan, Shuwang Zhang, Aditya Nagachandra, Di Chang, Dongdong Chen, Xin Eric Wang, and Achuta Kadambi. Vlm4d: Towards spatiotemporal awareness in vision language models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8600–8612, 2025. 2, 3