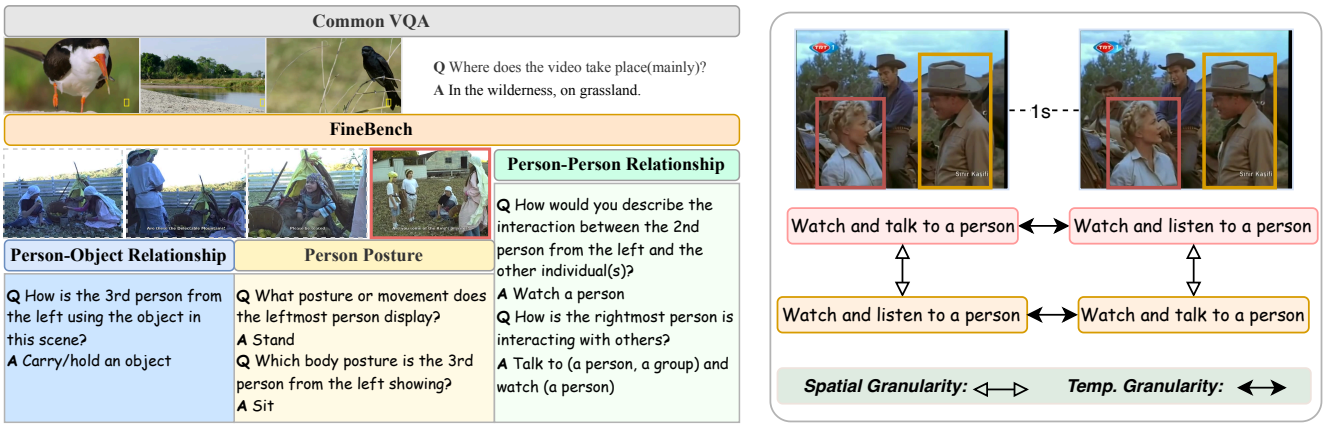
FineBench: Benchmarking and Enhancing Vision-Language Models for Fine-grained Human Activity Understanding

Gueter Josmy Faure[1,2], **Min-Hung Chen**[3], **Jia-Fong Yeh**[1], **Hung-Ting Su**[1], **Winston H. Hsu**[1]

[1]National Taiwan University, [2]Google, [3]Independent Researcher

Project Page — Code — Dataset



(a) FineBench versus regular VQAs

(b) FineBench's Levels of Granularity

Figure 1. **(a)** Examples of question types in FineBench which go beyond summarization to cover person posture, person-object interaction, and person-person interaction. **(b)** The capture of temporal evolution of interaction labels across frames, emphasizing spatial granularity (e.g., distinguish individuals in the same frame) and temporal granularity (e.g., resolving transitions between similar but distinct actions).

Abstract

Vision-Language Models (VLMs) have demonstrated remarkable capabilities in general video understanding, yet they often struggle with the fine-grained comprehension crucial for real-world applications requiring nuanced interpretation of human actions and interactions. While some recent human-centric benchmarks evaluate aspects of model behaviour such as fairness/ethics, emotion perception, and broader human-centric metrics, they do not combine long-form videos, very dense QA coverage, and frame-level spatial/temporal grounding at scale. To bridge this gap, we introduce FineBench, a human-centric video question answering (VQA) benchmark specifically designed to assess fine-grained understanding. FineBench comprises 199,420 multiple-choice QA pairs densely annotated across 64 long-form videos (15 minutes each), focusing on detailed person movement, person interaction, and object manipulation, including compositional actions. Our extensive evaluation reveals that while proprietary models like GPT-5 achieve respectable performance, current open-source VLMs significantly underperform, struggling particularly with spatial reasoning in multi-person scenes and distinguishing subtle differences in human movements and interactions. To address these identified weaknesses, we propose FineAgent, a modular framework that enhances VLMs by leveraging a Localizer and a Descriptor. Experiments show that FineAgent consistently improves the performance of various open VLMs on FineBench. FineBench provides a rigorous testbed for future research into fine-grained human-centric video understanding, while FineAgent offers a practical approach to enhance such reasoning in current VLMs.

1. Introduction

Vision-Language Models (VLMs) are rapidly advancing, showing increasing proficiency in interpreting and reasoning about visual content, particularly in the domain of

video understanding. However, much of the focus has been on general comprehension tasks—recognizing overall scenes, identifying high-level activities, or summarizing broad events. While valuable, this often falls short in real-world scenarios demanding a *fine-grained* understanding of video content involving humans. Fine-grained video understanding requires perceiving subtle visual details, precise temporal dynamics of actions, complex spatial relationships, and nuanced interactions, especially concerning human behavior. For instance, distinguishing between a person deliberately sitting versus accidentally falling, or discerning intricate social cues in a conversation, requires a level of detail beyond general scene description. Such capabilities are critical for applications ranging from assistive technologies and healthcare monitoring to autonomous systems and detailed behavior analysis.

Despite its importance, fine-grained, human-centric video understanding remains relatively underexplored and under-evaluated in the current VLM landscape. Existing VQA benchmarks often rely on sparsely annotated clips, focus on object-centric or broad activity recognition, or lack the scale and density needed to probe deep, temporally-grounded comprehension [11, 22, 23, 28]. As highlighted in Table 1, existing benchmarks often lack a specific focus on fine-grained human-centric actions, dense temporal and spatial grounding, or the sheer density of questions required to thoroughly test reasoning over extended video durations. This gap hinders progress, as we lack standardized ways to measure and drive improvements in VLMs' ability to grasp subtle human behavior in videos.

To address this gap, we introduce **FineBench**, a new benchmark specifically designed to evaluate fine-grained, human-centric video understanding. FineBench is formulated as a multiple-choice VQA dataset containing nearly 200,000 QA pairs derived from 64 long-form videos. Uniquely, it features dense annotations, averaging over 3,100 questions and linking to approximately 785 distinct keyframes per video, enabling detailed assessment of model capabilities at a granular temporal level (e.g., seconds). The questions cover three core domains: Person Movement, Person Interaction, and Object Manipulation, with over 20% requiring compositional reasoning about combined actions. FineBench explicitly tests spatial and temporal precision through carefully constructed questions and distractors derived from the rich annotations of the AVA v2.2 dataset [9].

Using FineBench, we conduct a comprehensive evaluation of state-of-the-art VLMs, encompassing both leading proprietary models and a wide array of open-source models. Our findings, detailed in Section 3.3 and summarized in Table 3, reveal a significant performance gap depending on the action type. Rather than a total failure of VLMs, we observe a dramatic performance divide: models excel at *Object Manipulation* tasks (scoring in the high 80s) but perform markedly worse on nuanced *Person Interaction* and *Person Movement* tasks (dropping to the 50s and 60s). While powerful proprietary models achieve a peak accuracy of around 77%, this indicates there is still over 20% room for improvement before the benchmark saturates. Furthermore, further analysis (Section 3.4, Figure 3) pinpoints specific weaknesses: VLMs exhibit a marked decline in accuracy as the number of people in the scene increases, underscoring enduring challenges with spatial reasoning and subject disambiguation in challenging multi-person scenes.

Motivated by these findings, we propose **FineAgent**, a modular framework designed to enhance the fine-grained video understanding capabilities of existing VLMs by directly addressing the identified bottlenecks (Section 4). FineAgent integrates two key components: a *Localizer* that provides explicit bounding box information to aid subject disambiguation in complex scenes, and a *Descriptor* that generates frame summaries, thereby providing richer semantic context. Our main contributions are as follows:

- We introduce FineBench, the first densely annotated, human-centric VQA benchmark targeting fine-grained video understanding, featuring 199,420 QA pairs.
- We provide a comprehensive benchmark of current proprietary and open-source VLMs on FineBench, revealing that while models succeed in object-centric tasks, there remains significant room for improvement (over 20%) in fine-grained reasoning abilities, particularly in spatial reasoning and nuanced action interpretation.
- We conduct an in-depth analysis identifying key failure modes for VLMs: degraded performance in multi-person scenarios (spatial reasoning) and difficulties understanding nuanced human movements and interactions.
- We propose FineAgent, a modular framework leveraging spatial grounding and contextual captioning, demonstrating its effectiveness in improving the fine-grained video understanding performance of various open-source VLMs by targeting their specific weaknesses.

2. Related Work

Our work on FineBench builds upon extensive research in Video Question Answering (VQA) and the rapid advancements in Vision-Language Models (VLMs).

Video Question Answering Datasets. VQA evaluates video understanding via question answering. While numerous datasets exist, early influential ones like MSRVTT-QA [23] and ActivityNet-QA [28] often lacked dense spatial or temporal grounding, limiting fine-grained evaluation (Table 1). Subsequent datasets focused on deeper reasoning (e.g., NExT-QA [22], STAR [20]) or specialized domains like egocentric video (EgoSchema [13]). Recent benchmarks (e.g., MovieChat [18], MVBench [11], TemporalBench [2], and MovieCORE [7]) address various aspects like long videos or temporal reasoning. Some bench-

Table 1. Comparison of FineBench with existing VQA datasets across key dimensions. Our dataset is the first to combine fine-grained actions, dense temporal grounding (Temporal G.), dense spatial grounding (Spatial G.), and large-scale QA in a human-centric setting.

	Num. QAs	*Num. Videos*	*Avg. Duration (s)*	*Density*	*Human-Centric*	*Spatial G.*	*Temporal G.*
VideoMME [8]	2,700	900	1017.9	3	✗	✗	✗
EgoSchema [13]	5,063	5,063	–	1	✓	✗	✗
MovieChat-1k [18]	13,000	1000	564	13	✗	✗	✗
ActivityNet-QA [28]	8,000	800	111.4	10	✗	✗	Partial
LongVideoBench [21]	6,678	3,763	473	1.8	✗	✗	Partial
NExT-QA [22]	8,564	1,000	39.5	8.6	✗	✗	✗
MSRVTT-QA [23]	72,821	2,990	15.2	24.4	✗	✗	✗
MSVD-QA [23]	13,157	504	9.8	26.1	✗	✗	✗
STAR [20]	7,098	914	11.9	7.8	✗	✗	✗
MVBench [11]	4,000	3,641	16	1.1	✗	✗	✗
TemporalBench [2]	10,000	2000	–	5	✗	✗	Partial
HV-MMBench [3]	8,700	1,200	–	7.25	✓	✗	✗
FineBench (Ours)	199,420	64	900	3115.94	✓	✓	✓

marks explicitly emphasize human-centric evaluation. HumaniBench [17] focuses on human-centered AI principles such as fairness and empathy through image tasks, whereas HumanVBench [30] explores human-centric video understanding with synthetic data pipelines targeting emotion perception and speech–visual alignment. However, a gap remains for evaluating fine-grained human action understanding with dense grounding, particularly in complex scenes. FineBench addresses this gap by providing large-scale QA with dense *spatial and temporal grounding of human actions and interactions* in long videos (avg. 900s), facilitating rigorous evaluation of precise human behavior localization and comprehension.

Vision-Language Models (VLMs). Vision-Language Models (VLMs), integrating vision encoders and LLMs, have revolutionized cross-modal understanding with early works such as LlaVA [12], MiniCPM-v2.6 [25], and more recently, InternVL-2.5 [4] and Qwen2.5-VL [1]. Extending this to video, recent VLMs like, mPlugOwl-3 [26], and HERMES [6] handle temporal information to perform video tasks, including video captioning and VQA. Despite their capabilities, our analysis (Section 3.3 and Section 3.4) reveals significant challenges for these models in fine-grained video understanding, particularly concerning spatial localization in complex scenes and interpreting nuanced human actions and interactions. This underscores the need for human-centric benchmarks like FineBench.

3. FineBench

To effectively evaluate Vision-Language Models' (VLMs) capacity for understanding nuanced visual content, we first delineate the characteristics of fine-grained video understanding as distinct from the general video understanding typically assessed by existing VQA datasets (Section 3.1) and overview FineBench. Section 3.2 then elaborates on our data creation and annotation process. Subsequently, we present extensive experiments benchmarking current VLMs to assess their proficiency in fine-grained video comprehension (Section 3.3). Finally, Section 3.4 examines the primary reasons these models struggle with such a task, providing insights for performance enhancements.

3.1. Overview of FineBench

Fine-grained video understanding represents a crucial yet relatively underexplored facet of video-language models (VLMs). Unlike general video understanding tasks that focus on broad concepts, scene recognition, or high-level activities, fine-grained understanding requires models to perceive and reason about subtle visual details, momentary actions, and precise object interactions within video frames. A fine-grained *human-centric* VQA dataset, in particular, must offer comprehensive coverage of all observable human behaviors. This includes not only the subject's body pose and movement, but also their interactions with objects (*person-object interactions*) and with other individuals (*person-person interactions*). Figure 1(a) illustrates this diversity by showcasing QA examples across different reasoning types supported by FineBench, from posture recognition to complex social interactions. Figure 1(b) highlights the temporal and spatial granularity required, where action labels evolve across frames and demand fine discrimination between visually similar behaviors.

A "Fine-grained" video understanding system must posess the ability to distinguish between visually similar activities that share common attributes. In our context, this includes disambiguation (between actions such as "carrying" vs. "lifting" an object), temporal precision (identifying when actions start/end), spatial attention (focusing on the relevant regions of a frame), and contextual reasoning

Table 2. Key Statistics of FineBench.

Statistic	Value
Total Questions	199,420
Unique Videos	64
Avg. Annotated Frames/Video	785
Category Distribution:	
Person Movement	94,330 (47.30%)
Person Interaction	70,140 (35.17%)
Object Manipulation	34,950 (17.53%)
Question Composition:	
Single Actions	158,625 (79.54%)
Combined Actions	40,795 (20.46%)

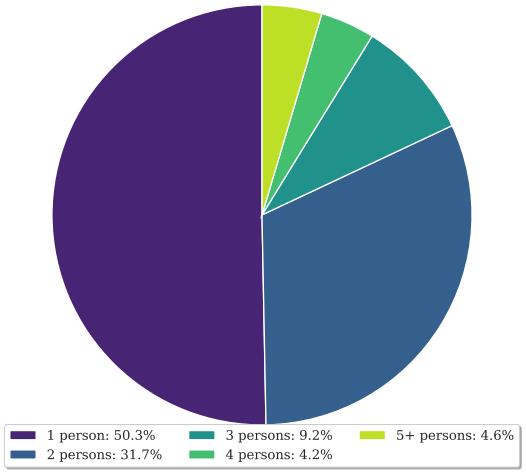


Figure 2. Distribution of Annotated Persons per Keyframe.

(understanding actions in relation to the environment).

The importance of fine-grained understanding for VLMs becomes evident when considering practical applications. In privacy-preserving ambient intelligence, general understanding might merely identify "several people in a room," whereas fine-grained perception can distinguish whether individuals are "standing in conversation," "reaching for objects," or "exhibiting signs of distress." For assisted living monitoring, fine-grained understanding allows systems to differentiate between "a person deliberately sitting down" versus "a person losing balance and falling"—a critical distinction for emergency response. Similar examples exist across human-robot interaction and everyday activities, positioning fine-grained video understanding as a fundamental capability that VLMs must possess to function effectively in complex real-world scenarios where subtle distinctions carry significant meaning.

Our human-centric fine-grained video benchmark, FineBench, is structured as a multiple-choice video question answering (VQA) dataset, where each question is accompanied by four candidate answers, only one of which is correct. It contains a total of **199,420 QA pairs**, making it one of the largest VQA datasets. While the dataset relies on **64 unique videos** derived from the AVA dataset, these are highly dense 15-minute movie clips. Questions are densely linked to an average of **785 unique keyframes** per video, enabling detailed probing of model understanding at the second level. Unlike existing VQA datasets that focus on general comprehension or sparse annotation across many short clips, FineBench offers an average of **3,100 QA pairs per video**. This design choice explicitly prioritizes the depth and density of spatio-temporal grounding over superficial breadth, ensuring the benchmark thoroughly tests nuanced reasoning over extended video durations and fostering both local and holistic reasoning.

Table 2 summarizes the key statistics of FineBench. The dataset spans three broad conceptual domains—*movement*, *human interaction*, and *object manipulation*—which guide the diversity of visual reasoning required. **Over 20% of QA pairs involve combined activities**, testing compositional reasoning where multiple visual cues must be integrated. Figure 2 shows that nearly half the frames contain multiple annotated persons, emphasizing the fine-grained nature of the interactions present in FineBench. These properties (along with those in Table 1) make FineBench the first benchmark explicitly designed to test VLMs' fine-grained human-centric video understanding ability, where success requires precision in space, time, and context.

3.2. Dataset Creation Process

The construction of FineBench leverages the human-annotated action classes and bounding boxes provided by the AVA dataset [9]. Our methodology integrates three core components: (1) systematic question generation using predefined templates, (2) a principled distractor selection strategy, and (3) spatial reasoning for subject disambiguation and subject-specific QA generation.

3.2.1. Question Template Design and Instantiation

We design a structured set of question templates categorized by the nature of the action being queried. Specifically, 23 templates were created for *person movement* actions (e.g., "How would you describe the movement of {*person*}?"), 21 templates for *object manipulation* actions (e.g., "How is {*person*} interacting with the object?"), and 25 templates for *person interaction* actions (e.g., "What social interaction is {*person*} engaged in?"). To anchor these questions visually and ensure clarity, the placeholder {*person*} within each template is instantiated using spatial descriptors de-

rived dynamically from bounding box positions. Phrases such as "the leftmost person" or "the person in the center" are employed to unambiguously refer to the specific individual relevant to the question within the video frame.

Our reliance on template-based generation, as opposed to free-form LLM generation, is a deliberate design choice to prevent LLM hallucinations during dataset construction. By strictly binding questions and answers to rigorously annotated human labels from AVA, we ensure that FineBench strictly measures visual perception and spatial reasoning capabilities rather than VLMs' language priors.

3.2.2. Distractor Generation Strategy

For each annotated action instance in AVA v2.2, we generate a corresponding multiple-choice question. The process begins by classifying the ground truth action into one of the three categories: person movement, object manipulation, or person interaction. A question template is then randomly selected from the pool corresponding to that action category. Plausible distractors (incorrect answer options) are generated using a two-tiered approach. The primary strategy involves selecting actions that are semantically similar to the correct answer, based on a predefined similarity mapping. For example, actions like "hand wave", "hand clap", and "hand shake" are considered semantically close and may serve as distractors for one another, thereby increasing the question's difficulty. If no sufficiently similar actions are found via this mapping, a fallback strategy is employed: distractors are randomly selected from the same broad action category (e.g., other person movement actions) to maintain contextual relevance. In scenarios where an individual is annotated with multiple concurrent actions belonging to the same category, we formulate compound questions (e.g., reflecting simultaneous actions like "listening to and watching a person") and select appropriate distractors.

3.2.3. Spatial Referencing and Disambiguation

To enable precise questioning about specific individuals within a scene, especially when multiple people are present, we implement a dynamic spatial referencing system based on bounding box locations. When only one or two individuals are detected, relative positional terms (e.g., "the person on the left", "the person on the right") are used for disambiguation. For scenes containing three or more individuals, ordinal references (e.g., "the second person from the left") are generated to ensure clarity. This ensures that the generated questions unambiguously target the intended person.

3.3. Do VLMs Exhibit Fine-Grained Video Understanding?

To evaluate whether current Vision-Language Models can perform fine-grained human-centric video understanding, we benchmark a diverse set of proprietary and open-source models using FineBench, integrated into the VLMEvalkit [5] library. Due to the high cost of querying proprietary APIs at scale, we provide results on two tiers: a representative subset (7 videos, 20,143 QAs) and the full dataset for open models only. The results are shown in Table 3.

Proprietary models, notably GPT-5-mini [16] and Gemini-2.0-Flash [19], demonstrate strong performance on the representative subset, substantially outperforming open models evaluated on the same data. This suggests these proprietary models possess stronger spatio-temporal reasoning and fine-grained human activity disambiguation capabilities, likely due to large-scale pretraining and robust multimodal pipelines. Crucially, as highlighted in Table 3, the performance of open-source models on the subset closely matches their performance on the full dataset (e.g., MiniCPM scores 58.4% on the subset vs. 59.2% on the full dataset; similar tight tracking is observed for InternVL-2.5 and Qwen2.5-VL placeholders). This directly proves that the subset serves as a highly representative split, allowing for robust direct comparisons between closed-source APIs and open-source models without requiring the prohibitive costs of full-dataset evaluations for proprietary models.

In contrast, open models exhibit wide variability and underwhelming accuracy on the full dataset. The top open model, Qwen2.5-VL (7B) [1], achieves 68.8%, but most models cluster around 55–60%, and a few perform near chance level on Person Movement-related questions. These gaps indicate that current open VLMs struggle with fine-grained temporal cues, subtle interactions, and compositional reasoning—core challenges posed by FineBench. Such results highlight a critical gap in the open ecosystem and a need for progress in training methods, architectures, and benchmarks tailored for fine-grained human-centric video comprehension.

Finding 1: *Current VLMs exhibit a clear performance divide across action types. They handle object-centric tasks well but fall significantly short on human-centric reasoning, revealing that fine-grained human activity understanding remains an open challenge.*

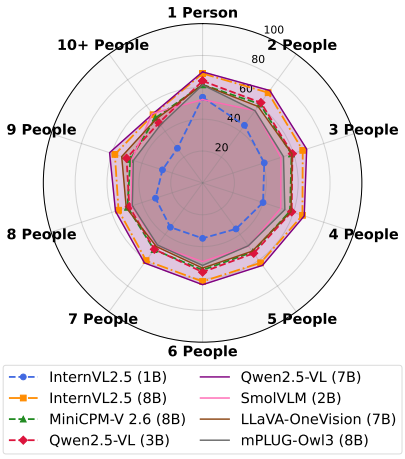
3.4. Why Do VLMs Struggle With Fine-Grained Video Understanding?

Having established that current Vision-Language Models (VLMs) underperform on fine-grained video understanding tasks (Section 3.3), we investigate the underlying reasons by dissecting their performance. Our analysis focuses on two key aspects: the impact of scene complexity (number of persons) and the variation in performance across different action categories, visualized through radar charts in Figure 3a and Figure 3b, respectively. Additionally, we investigate the influence of input context length to ascertain if insufficient visual information is a bottleneck, as shown in Figure 3c.

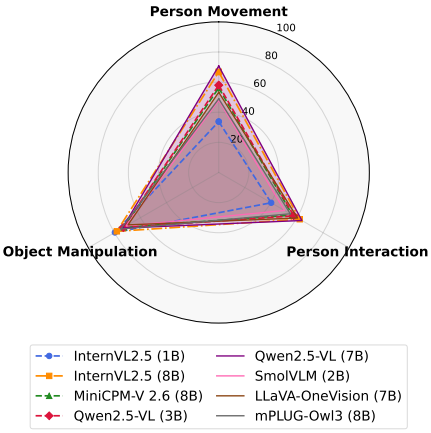
First, analyzing the accuracy relative to the number of people present (Figure 3a) reveals a significant and consis-

Table 3. **Performance of 15 Vision-Language Models (VLMs) on FineBench**. Proprietary models evaluated on a representative subset–comprising 7 representative videos and totaling 20,143 questions–are shown at the top. Open models are evaluated on both the subset and the full dataset. The best full-dataset open score is **bolded** and the second-best underlined. [P.: Person; Obj.: Object]

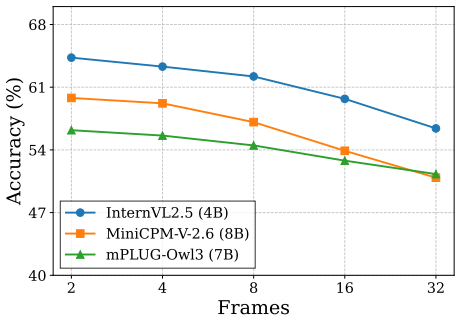
	Size	P. Movement	P. Interaction	Obj. Manipulation	Avg.
Random Choice	–	25.0	25.0	25.0	25.0
Subset Evaluation					
GPT-4o (2024/08/26) [15]	–	70.9	73.9	84.4	74.3
GPT-5-mini (2025/08/07) [16]	–	**75.9**	**75.3**	85.3	**77.4**
Gemini-1.5-Flash [19]	–	71.2	66.8	81.9	71.6
Gemini-2.0-Flash [19]	–	**75.9**	68.7	**86.3**	75.2
SmolVLM [14]	2B	48.5	48.0	80.0	53.9
MiniCPM-2.6 [25]	8B	49.5	57.4	84.8	58.4
mPlugOwl-3 [26]	7B	47.9	55.8	84.0	56.6
Full Dataset Evaluation					
InternVL-2.5 [4]	1B	33.8	40.2	**79.6**	44.1
SmolVLM [14]	2B	47.9	50.5	71.0	52.9
Qwen2.5-VL [1]	3B	58.0	57.5	73.2	60.5
BLIP-3 [24]	4B	34.3	58.6	64.9	48.2
InternVL-2.5 [4]	4B	61.4	58.6	78.1	63.3
mPlugOwl-2 [27]	7B	57.6	49.2	78.5	58.3
mPlugOwl-3 [26]	7B	48.9	54.8	75.2	55.6
MiniCPM-2.6 [25]	8B	56.2	56.5	72.8	59.2
LLaVA-OV [10]	7B	53.3	60.4	69.6	58.6
InternVL-2.5 [4]	8B	66.8	62.1	78.1	67.1
Qwen2.5-VL [1]	7B	**70.7**	**63.8**	73.9	**68.8**



(a) Acc. per number of person



(b) Acc. per question category



(c) Accuracy per the number of frames

Figure 3. VLM performance analysis on FineBench detailing accuracy variations. (a) Performance degradation with increasing number of persons in the scene. (b) Performance differences across action categories, with Person Movement being consistently lower. (c) Performance degradation with increasing number of frames.

tent challenge for all evaluated VLMs. There is a clear trend of performance degradation as the number of individuals in the frame increases. For example, Qwen2.5-VL (7B), the top-performing model overall, has a peak accuracy of 71.7% in scenes with 2 persons, but this accuracy drops to 53.4% when 10 or more people are present.

This decline is even more pronounced for smaller models like InternVL-2.5 (1B), which drops from 53.7% to 26.9%. This consistent decrease suggests that VLMs struggle significantly with spatial reasoning, target disambiguation, and relationship understanding in complex, multi-person scenarios. Identifying and tracking the specific actions of designated individuals becomes substantially harder amidst visual clutter and potential occlusions.

Second, examining performance across action categories (Figure 3b) highlights another area of weakness. Models consistently demonstrate higher proficiency in identifying *Object Manipulation* actions compared to *Person Movement* and *Person Interaction*. Across all tested models, accuracy for Object Manipulation typically ranges from 71% to nearly 80%, whereas accuracies for the other two categories are often considerably lower. For instance, InternVL-2.5 (8B) achieves 78.1% on Object Manipulation but only 66.8% on Person Movement and 62.1% on Person Interaction. This disparity suggests that VLMs find it easier to recognize actions centered around distinct object interactions, which may offer clearer visual cues. Conversely, they appear less capable of interpreting the nuances of human kinematics involved in diverse movements and the complex, often subtle, cues defining social interactions between individuals. These person-centric categories demand a deeper understanding of human pose, gestures, and context that current models do not fully capture. We also isolate the impact of the vision components with a blind evaluation showing that Qwen2.5VL (7B), MiniCPM-v2.6 (8B), and InternVL-2.5 (8B) score only 43.5, 29.9, and 33.0, respectively, when blind.

Our key takeaway is that current open-source VLMs struggle with fine-grained video understanding primarily due to two challenges. First, they exhibit deficiencies in robust spatial reasoning and subject disambiguation, particularly as scene complexity (number of actors) increases. This makes it difficult to correctly attribute actions to the right individuals. Second, they find it harder to interpret and distinguish nuanced human-centric actions, especially subtle body movements and complex social interactions, compared to more visually salient object-related actions. These person-centered tasks require models to pick up on fine-grained visual details and temporal patterns of human behavior, which current architectures and training paradigms are not yet adept at. Addressing these limitations is key for advancing fine-grained human-centric video understanding.
Finding 2: *Scene complexity is a critical bottleneck: The more people present, the harder it becomes for VLMs to correctly attribute actions.*

4. FineAgent

Our error analysis in Section 3.4 identifies two primary obstacles hindering the fine-grained video understanding capabilities of current VLMs: (1) difficulties with spatial reasoning and subject disambiguation in multi-person scenes, and (2) a weaker grasp of nuanced human movements and interactions compared to object-centric actions. To address these limitations, we propose **FineAgent**, a modular framework designed to augment existing VLMs with spatial grounding and contextual information, thereby enhancing their fine-grained reasoning abilities.

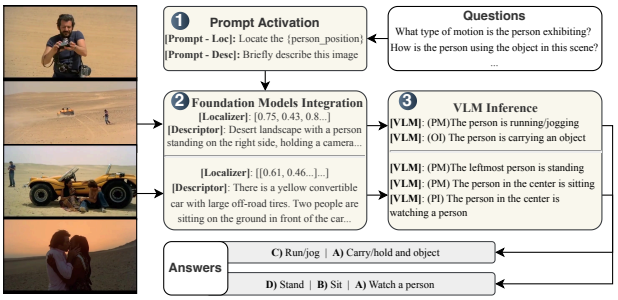


Figure 4. **Workflow of FineAgent**. It begins with (1) prompt activation for the Localizer and Descriptor. (2) The Localizer and Descriptor, both Foundation models, provide bounding box coordinates and textual captions. (3) Finally, the VLM uses this processed information during inference.

4.1. How does FineAgent Enhances Fine-grained Video Understanding?

FineAgent enhances VLMs' fine-grained video understanding capabilities at inference time by integrating two complementary modules, designed to provide information that directly addresses the weaknesses identified in Section 3.4. The workflow of FineAgent is illustrated in Figure 4.

The first module is the **Localizer**, instantiated using EVFSam [29], a foundation model adept at visual grounding and referring segmentation. Given the video frames and the question, the Localizer provides the spatial location of the individual pertinent to the query. By supplying positional information, this module directly tackles the VLM's observed struggle with spatial reasoning and subject disambiguation in multi-person scenes. The Localizer thus assists the base VLM in anchoring its visual analysis to the correct subject, mitigating confusion in crowded environments.

The second module is the **Descriptor**. This component is responsible for generating captions for the relevant video frames. We utilize Qwen2.5-VL (7B) [1] as the Descriptor, due to its strong performance among open-source VLMs (Table 3). The Descriptor addresses the VLM's weakness in interpreting subtle human-centric actions, particularly those categorized under Person Movement and Person Interaction (Figure 3b). The generated captions provide semantic context and higher-level descriptions of potentially ambiguous activities. This augments the base VLM's understanding beyond raw visual features and aids in the interpretation of

Table 4. Performance gains with FineAgent across different models.

Model	P. Movement	P. Interaction	Obj. Manipulation	Avg.
InternVL-2.5 (1B) [4]	33.8	40.2	79.6	44.1
+ **FineAgent**	**47.9** (+14.1)	**44.2** (+4.0)	**80.6** (+1.0)	**52.4** (+8.3)
Qwen2.5-VL (7B) [1]	70.7	63.8	73.9	68.8
+ **FineAgent**	**71.5** (+0.8)	**64.1** (+0.3)	**76.3** (+2.4)	**69.7** (+0.9)
mPlugOwl-3 (7B) [26]	48.9	54.8	75.2	55.6
+ **FineAgent**	**60.8** (+11.9)	**57.8** (+3.0)	**77.4** (+2.2)	**62.7** (+7.1)
MiniCPM-2.6 (8B) [25]	56.2	56.5	72.8	59.2
+ **FineAgent**	**60.6** (+4.4)	**58.8** (+2.3)	**76.3** (+3.5)	**62.7** (+3.5)

complex kinematics or social cues that might otherwise be missed. These two modules operate synergistically: the Localizer first identifies *who* and *where* the question is focused on, and then the Descriptor provides a textual interpretation of *what* is happening. This structured, auxiliary information is then combined with the question and video input, and fed to the VLM to facilitate a more informed prediction.

The effectiveness of integrating **FineAgent** is demonstrated empirically in Table 4. Augmenting various base VLMs—including InternVL-2.5 (1B) [4], Qwen2.5-VL (7B) [1], mPLUG-Owl-3 (7B) [26], and MiniCPM-2.6 (8B) [25] with **FineAgent** framework consistently yields performance improvements across all models and action categories on FineBench. Notably, the improvements are often most pronounced in the challenging Person Movement and Person Interaction categories, directly addressing the identified weaknesses. For instance, augmenting the InternVL-2.5 (1B) model with **FineAgent** boosts its Person Movement accuracy by a substantial 14.1 percentage points and Person Interaction accuracy by 4.0 points, resulting in an overall 8.3-point increase in average accuracy. Similar positive trends, with varying magnitudes, are observed across the other models. This validates our hypothesis that by specifically targeting spatial grounding and providing richer contextual descriptions for human actions, **FineAgent** can successfully enhance the fine-grained video understanding capabilities of existing VLMs.

Finding 3: *Explicitly providing spatial grounding and contextual descriptions at inference time consistently improves fine-grained video understanding, suggesting that targeted auxiliary information can compensate for architectural weaknesses without retraining.*

4.2. Importance of FineAgent Components

Table 5 ablates the contribution of each module. The Localizer alone yields modest but consistent gains (+2.8% for mPlugOwl-3, +0.5% for Qwen2.5-VL), confirming that explicit spatial grounding helps subject disambiguation in multi-person scenes. The Descriptor contributes more substantially for mPlugOwl-3 (+6.9%), but minimally for Qwen2.5-VL (+0.7%)—an expected result, since the Descriptor itself is powered by Qwen2.5-VL and thus offers little additional signal to the same backbone. Swapping in InternVL-2.5 (8B) as Descriptor (†) recovers this gap (+1.1%), further supporting this explanation. Combined, both modules act synergistically: the total gain exceeds the sum of individual contributions for both models.

Finding 4: *Spatial grounding and semantic context are complementary: combining both yields synergistic gains, underscoring the importance of jointly addressing where and what in activity understanding.*

Table 5. Ablation study on FineAgent components. We report average accuracy (%) on FineBench. Each column corresponds to adding a specific module to the base VLM. Improvements over the base model are shown in green. † means InternVL2.5 (8B) is used as Descriptor.

Model	+ Localizer		+ Descriptor		+ FineAgent	
	Acc.	Δ	Acc.	Δ	Acc.	Δ
mPlugOwl-3 (7B)	58.4	(+2.8)	62.5	(+6.9)	**63.7**	(+8.1)
Qwen2.5-VL (7B)	69.3	(+0.5)	69.5	(+0.7)	**69.7**	(+0.9)
Qwen2.5-VL (7B) †	69.3	(+0.5)	**69.9**	(+1.1)	**70.2**	(+1.4)

5. Conclusion

We introduce FineBench, a densely annotated benchmark of 199 420 QA pairs probing fine-grained, human-centric video understanding. Our evaluation exposes two systematic weaknesses in current open-source VLMs: poor spatial reasoning in multi-person scenes, and limited sensitivity to subtle human movements and interactions. FineAgent directly targets these bottlenecks via spatial grounding and contextual captioning, yielding consistent gains across diverse architectures without retraining. We hope FineBench serves as a rigorous testbed to drive future progress in this underexplored yet practically critical domain.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 3, 5, 6, 7, 8

[2] Mu Cai, Reuben Tan, Jianrui Zhang, Bocheng Zou, Kai Zhang, Feng Yao, Fangrui Zhu, Jing Gu, Yiwu Zhong, Yuzhang Shang, et al. Temporalbench: Benchmarking fine-grained temporal understanding for multimodal video models. *arXiv preprint arXiv:2410.10818*, 2024. 2, 3

[3] Yuxuan Cai, Jiangning Zhang, Zhenye Gan, Qingdong He, Xiaobin Hu, Junwei Zhu, Yabiao Wang, Chengjie Wang, Zhucun Xue, Xinwei He, et al. Hv-mmbench: Benchmarking mllms for human-centric video understanding. *arXiv preprint arXiv:2507.04909*, 2025. 3

[4] 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, 6, 8

[5] Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM international conference on multimedia*, pages 11198–11201, 2024. 5

[6] Gueter Josmy Faure, Jia-Fong Yeh, Min-Hung Chen, Hung-Ting Su, Winston H. Hsu, and Shang-Hong Lai. Hermes: temporal-coherent long-form understanding with episodes and semantics, 2024. 3

[7] Gueter Josmy Faure, Min-Hung Chen, Jia-Fong Yeh, Ying Cheng, Hung-Ting Su, Yung-Hao Tang, Shang-Hong Lai, and Winston H. Hsu. Moviecore: Cognitive reasoning in movies, 2025. 2

[8] Chaoyou Fu, Yuhan Dai, Yondong 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. *arXiv preprint arXiv:2405.21075*, 2024. 3

[9] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6047–6056, 2018. 2, 4

[10] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 6

[11] 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, 3

[12] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 3

[13] 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, 3

[14] Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, et al. Smolvlm: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299*, 2025. 6

[15] OpenAI. Hello gpt-4o. https://openai.com/index/hello-gpt-4o/, 2024. [Accessed 01-11-2024]. 6

[16] OpenAI. Introducing gpt-5. https://openai.com/index/introducing-gpt-5/, 2025. [Accessed 31-08-2025]. 5, 6

[17] Shaina Raza, Aravind Narayanan, Vahid Reza Khazaie, Ashmal Vayani, Mukund S Chettiar, Amandeep Singh, Mubarak Shah, and Deval Pandya. Humanibench: A human-centric framework for large multimodal models evaluation. *arXiv preprint arXiv:2505.11454*, 2025. 3

[18] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18221–18232, 2024. 2, 3

[19] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 5, 6

[20] Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. Star: A benchmark for situated reasoning in real-world videos. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2024. 2, 3

[21] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857, 2024. 3

[22] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9777–9786, 2021. 2, 3

[23] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016. 2, 3

[24] Le Xue, Manli Shu, Anas Awadalla, Jun Wang, An Yan, Senthil Purushwalkam, Honglu Zhou, Viraj Prabhu, Yutong Dai, Michael S Ryoo, et al. xgen-mm (blip-3): A family of open large multimodal models. *arXiv preprint arXiv:2408.08872*, 2024. 6

[25] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024. 3, 6, 8

[26] Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. *arXiv preprint arXiv:2408.04840*, 2024. 3, 6, 8

[27] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 13040–13051, 2024. 6

[28] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9127–9134, 2019. 2, 3

[29] Yuxuan Zhang, Tianheng Cheng, Lianghui Zhu, Rui Hu, Lei Liu, Heng Liu, Longjin Ran, Xiaoxin Chen, Wenyu Liu, and Xinggang Wang. Evf-sam: Early vision-language fusion for text-prompted segment anything model, 2024. 7

[30] Ting Zhou, Daoyuan Chen, Qirui Jiao, Bolin Ding, Yaliang Li, and Ying Shen. Humanvbench: Exploring human-centric video understanding capabilities of mllms with synthetic benchmark data. *arXiv preprint arXiv:2412.17574*, 2024. 3