Are Pretrained VideoLLMs Enough for Action Recognition? A Zero-Shot Evaluation on Charades

Andrea Lagorio[1], Giuseppe A. Tunfio[1], Matteo Poddighe[1], Massimo Tistarelli[1], and Pietro Ruiu[1]

Dipartimento di Ingegneria, Università degli Studi di Sassari, via Vienna 2, 07100 Sassari, Italy
{lagorio,trunfio,tista,pruiu}@uniss.it; m.poddighe27@studenti.uniss.it

Abstract. Human action recognition (HAR) in realistic video understanding scenarios remains challenging due to the presence of multiple concurrent activities, fine-grained action categories, and complex human-object interactions. Recent Video Large Language Models (VideoLLMs) have demonstrated strong multimodal reasoning capabilities, suggesting their potential as general-purpose solutions for video understanding. However, their effectiveness for zero-shot action recognition in complex video datasets is not yet well understood.

This work investigates the ability of pre-trained VideoLLMs to perform action recognition without task-specific architectures or supervised training on the target task. We evaluate representative VideoLLMs under different zero-shot inference strategies, including direct action prediction and description-based reasoning pipelines, on a challenging benchmark characterized by dense and diverse action annotations. The study analyzes how inference design choices affect recognition performance and examines the strengths and limitations of current VideoLLMs for fine-grained action understanding.

Our results show that VideoLLMs can achieve meaningful zero-shot action recognition performance, demonstrating the feasibility of using general-purpose multimodal models for action understanding. Among the evaluated strategies, direct video-to-label prediction consistently provides the strongest results, while description-based approaches can improve interpretability but may lose discriminative action information. Furthermore, fine-tuning significantly enhances recognition accuracy, with the best-performing model achieving competitive results on the benchmark. These findings highlight both the promise of VideoLLMs as flexible action recognition systems and the need for improved temporal reasoning to support robust fine-grained action understanding.

Keywords: Action recognition · Video Large Language Models · zero-shot learning · Charades dataset · multimodal video understanding

1 Introduction

Human action recognition (HAR) is a central problem in video understanding, with applications in surveillance, human-computer interaction, smart environments, assistive technologies, and video retrieval. The task requires a model not only to identify visual entities appearing in a scene, but also to interpret their temporal evolution, the interactions between people and objects, and the semantic meaning of motion patterns. For this reason, action recognition has traditionally been addressed through models specifically designed or fine-tuned for video classification, often relying on architectures able to encode temporal dynamics, such as 3D convolutional networks, two-stream networks, recurrent models, or transformer-based video encoders [8].

In recent years, large-scale Vision-Language Models and Video Large Language Models (VideoLLMs) have shown strong progress in multimodal understanding [13]. These models are trained on large and heterogeneous collections of image, video, and text data, and are able to perform a broad range of tasks through natural-language prompts. Their flexibility makes them attractive for zero-shot action recognition, where the model is asked to classify an observed action without being explicitly trained on the target dataset or on the specific action categories [11, 3, 1]. This setting is particularly relevant because it tests whether the semantic and temporal knowledge acquired during pre-training can generalize to standard action recognition benchmarks.

However, the actual effectiveness of pre-trained VideoLLMs for zero-shot action recognition remains unclear. Many works in the literature report strong results on video understanding tasks, but these results often involve models specifically adapted to the target task, task-specific architectures, supervised training, prompt engineering, or auxiliary modules [6, 5]. As a consequence, it is not always evident whether general-purpose VideoLLMs can directly replace specialized action recognition systems, especially when evaluated under strict zero-shot conditions on datasets with fine-grained and semantically overlapping action labels.

The main objective of this work is to assess whether pre-trained general-purpose VideoLLMs can perform action recognition on Charades dataset [10] in a zero-shot setting. Instead of designing a task-specific model, we evaluate how well these models transfer their multimodal pre-training capabilities to a structured action taxonomy. We also compare direct video-to-label prediction with a description-based strategy, in order to understand whether textual descriptions can improve the mapping between video content and action labels.

2 Background and Related Work

This section reviews prior work that uses Charades [10] as an evaluation benchmark, in order to position our study with respect to existing approaches for multi-label, zero-shot, and fine-grained action recognition.

Recent work in video understanding has increasingly moved from purely visual classification toward multimodal formulations that connect video content

with natural language. This direction is particularly relevant for Charades, where actions are often fine-grained, object-dependent, and temporally overlapping. ActionCLIP [11] is a representative example of this trend. It reformulates action recognition as a video-text matching problem, exploiting the semantic information contained in action labels instead of using only as numeric class identifiers. Based on a CLIP-derived *pre-train, prompt, and fine-tune* paradigm, ActionCLIP reports 44.3 mAP on Charades using a ViT-B/16 backbone with 32 input frames, providing a task-oriented reference for video-text action recognition.

The role of label semantics is also investigated by Calabrese et al. [3], who study zero-shot multi-label action recognition for object-based actions. Their work shows that many Charades labels have a compositional structure, typically combining an action verb with an object, and that this structure directly affects zero-shot generalization. In the same direction, Dual-VCLIP [3] combines VCLIP with DualCoOp to address zero-shot multi-label action recognition, aiming to improve the recognition of unseen action categories while reducing the dependence on task-specific training data.

Other works using Charades focus on temporal localization and complex action structures in untrimmed videos. Gao et al. [6] propose the Human Action Aware Network, which combines RGB frames and estimated multi-person skeletons through a cross-modality fusion module to improve multi-label temporal action detection. Nawhal et al. [9] introduce the Activity Graph Transformer, an end-to-end model that represents video content through graph-based structures and predicts sets of action instances, allowing the model to reason over overlapping and recurring actions. Dehghani et al. [4] present Scenic, an open-source JAX framework for computer vision research, which includes transformer-based video models and provides reference results used in the literature for Charades.

The difficulty of fine-grained action recognition with vision-language foundation models has been further highlighted by Ali et al. [1]. Their study shows that, although these models transfer well to several video understanding tasks, they still struggle when actions require precise temporal reasoning and accurate alignment between visual evidence and textual labels. Their evaluation includes Charades and confirms that object-dependent and fine-grained human action recognition remains challenging for pre-trained video-language models.

Overall, previous studies show that language supervision, visual-semantic embeddings, prompt design, temporal modeling, skeleton-RGB fusion, and task-specific adaptation can improve action recognition on Charades. However, these methods generally rely on architectures or training procedures specifically designed for action recognition. In contrast, our work evaluates recent general-purpose generative VideoLLMs, including InternVL and Qwen-based models [12, 2], without introducing a new task-specific architecture. Our goal is to assess whether action recognition capabilities already emerge from large-scale multimodal pre-training and can be exploited through zero-shot prompting.

We evaluate two prompt-based strategies. In the direct video-to-label setting, the VideoLLM receives the video and predicts the Charades action labels directly. In the description-based setting, the VideoLLM first generates a textual

description of the video, and a separate language model maps this description to the target action classes. This allows us to compare direct visual action prediction with an intermediate language-based representation, and to assess their effectiveness for zero-shot action recognition on Charades.

3 Experimental Setup and Methodology

Fig. 1. Example of a Charades video with multiple temporally localized action annotations. Each action is associated with a start and end time, enabling the extraction of action-level clips from the original video.

This study evaluates several pre-trained VideoLLMs in a zero-shot setting using the Charades dataset [10], a large-scale benchmark for human activity recognition in everyday indoor environments. The dataset contains 9,848 videos recorded in home settings by 267 different subjects, and each video is annotated with temporal action intervals, object labels, and textual descriptions (see Fig. 1). It covers 157 action classes and 46 object categories, which makes it particularly suitable for studying complex activities involving multiple actions and object interactions within the same sequence. This dataset is used because it reflects realistic daily behavior rather than simplified laboratory actions. In many videos, actions overlap in time and occur in cluttered domestic scenes, with noticeable variations in viewpoint, illumination, and background. These characteristics make Charades a challenging and representative benchmark for evaluating methods for practical video understanding.

In our experiments, we use two evaluation settings based on Charades. The first one corresponds to the original Charades dataset, using the full label set of 157 action classes. This setting is used to assess the scalability and robustness of the selected models under the standard, more complex action recognition scenario, where the label space is large and several actions are visually or semantically close. The second setting is based on a newly derived dataset, named Charades-24 Balanced. This dataset was constructed from the original Charades annotations by converting video-level annotations into action-level clips. Since

Charades provides the start and end timestamps for each annotated action (Fig. 1), each original video was temporally segmented into shorter clips, with each clip associated with a single action label. Table 1 reports the action classes included in this subset.

Class	**Description**	**Train + Val.**	**Test**	**TOT**
c000	Holding some clothes	480	200	680
c001	Putting clothes somewhere	480	200	680
c006	Closing a door	480	200	680
c008	Opening a door	480	200	680
c009	Putting something on a table	480	200	680
c015	Holding a phone/camera	480	200	680
c026	Holding a book	480	200	680
c033	Holding a towel/s	480	200	680
c059	Sitting in a chair	480	200	680
c061	Holding some food	480	200	680
c062	Putting some food somewhere	480	200	680
c063	Taking food from somewhere	480	200	680
c097	Walking through a doorway	480	200	680
c106	Drinking from a cup/glass/bottle	480	200	680
c107	Holding a cup/glass/bottle of something	480	200	680
c109	Putting a cup/glass/bottle somewhere	480	200	680
c110	Taking a cup/glass/bottle from somewhere	480	200	680
c118	Holding a dish	480	200	680
c141	Grasping onto a doorknob	480	200	680
c149	Someone is laughing	480	200	680
c151	Someone is going from standing to sitting	480	200	680
c152	Someone is smiling	480	200	680
c154	Someone is standing up from somewhere	480	200	680
c156	Someone is eating something	480	200	680
Total		11520	4800	16320

Table 1. Composition of the Charades-24 Balanced dataset, derived from Charades by temporally segmenting the original videos according to the annotated start and end timestamps.

After generating the action-level clips, we computed the number of available samples for each class and selected the 24 most frequent action classes with at least 200 clips in the test partition. To obtain a balanced training set, the number of training clips was fixed to 480 per class, corresponding to the smallest class cardinality among the selected classes. Each class therefore contains 480 training clips and 200 test clips, for a total of 680 clips per class. The resulting dataset includes 11,520 video clips overall. Charades-24 Balanced provides a controlled and balanced protocol for evaluating the ability of the models to distinguish common actions under a reduced label space. In contrast, the full Charades setting preserves the complete 157-class taxonomy and is used to test model behaviour under a broader and more challenging classification scenario. In both

settings, the models are prompted to classify the observed action directly from the video.

The experimental evaluation considers recent general-purpose VideoLLMs belonging to the Qwen-VL and InternVL families. These models were selected because they provide strong multimodal understanding capabilities, support instruction-based prompting, and are not specifically designed for action recognition. They are also publicly available in multiple scales, which makes them suitable for assessing the zero-shot recognition abilities of pre-trained VideoLLMs under a realistic prompt-based evaluation setting.

Qwen3-VL [2] combines a vision encoder, an MLP-based vision–language merger, and a large language model (LLM). The Qwen-VL family includes Qwen2.5-VL-72B, Qwen3-VL-32B, and Qwen3-VL-8B. These models allow us to analyze the impact of model scale, ranging from large high-capacity architectures to a more compact 8B variant. The Qwen3-VL-8B model was also evaluated after fine-tuning, in order to compare the strict zero-shot setting with a supervised adaptation scenario. This model was selected for fine-tuning because it is the smallest among the evaluated Qwen3-VL variants and therefore represents the most feasible option under the available computational resources. Despite its reduced size, fine-tuning still required a high-memory GPU server equipped with two NVIDIA H100 GPUs, each providing 95 GB of VRAM. We fine-tuned the model using supervised fine-tuning with LoRA. The vision encoder and language model were kept frozen, while the multimodal merger and LoRA adapters were optimized. LoRA was applied with rank 32, alpha 64, and dropout 0.05. Training was conducted for 100 epochs on two GPUs with a global batch size of 32, using bfloat16 precision and gradient checkpointing. The optimizer used a learning rate of 2×10^{-5}, a weight decay of 0.01, and a cosine learning-rate schedule with a warmup ratio of 0.03. Videos were sampled at 1 FPS, and validation was performed at the end of each epoch. The fine-tuning procedure was performed using the training split of the Charades dataset. This split was further divided into training and evaluation subsets according to an 80/20 partition. The original test split was kept separate and used only for the final performance assessment.

InternVL3.5 [12] introduces three key innovations, namely Cascade Reinforcement Learning (Cascade RL), the Visual Resolution Router (ViR), and Decoupled Vision–Language Deployment (DvD). The InternVL family includes InternVL3-78B (INT8) and InternVL3.5-38B. The INT8 variant was considered to assess whether quantized models can retain useful recognition capabilities while reducing memory requirements and computational cost.

3.1 Proposed approaches

Two different pipelines are investigated to evaluate the capabilities of VideoLLMs for action recognition, as summarized in Fig. 2. The first approach, referred to as the *Direct video-to-label pipeline*, directly performs video-to-label classification by predicting the relevant action label(s) from the input video. The second approach, denoted as the *Description-based pipeline*, prompts the VideoLLM

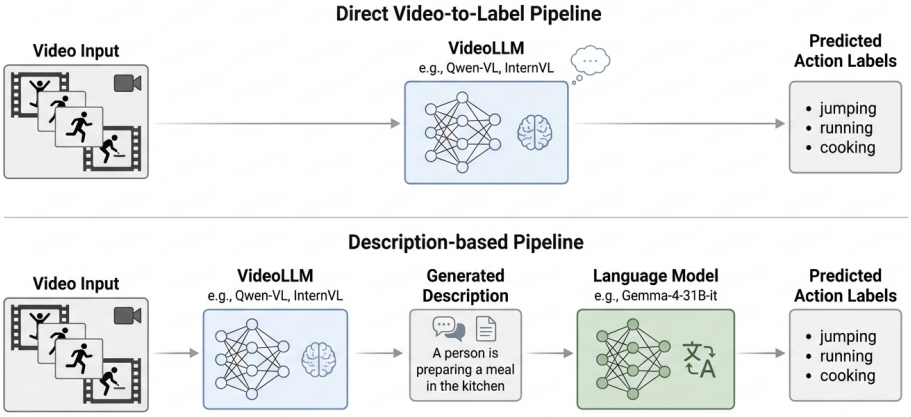

Fig. 2. Overview of the two evaluation pipelines considered in this work. In the direct video-to-label pipeline, the VideoLLM receives the input video and directly predicts the relevant Charades action labels. In the description-based pipeline, the VideoLLM first generates a textual description of the video, which is then processed by Gemma-4-31B-it to produce the final action labels.

to generate a textual description of the video content. The generated description is subsequently provided to Gemma-4-31B-it [7], which performs the final classification by identifying the relevant action label(s). This second protocol enables the separate evaluation of two complementary capabilities: the ability of the VideoLLM to produce semantically meaningful video descriptions and the ability of the LLM to infer the corresponding action classes from the generated text. Comparing the two pipelines provides insight into the effectiveness of intermediate textual representations for zero-shot action recognition.

3.2 Evaluation metrics

The mAP is computed from continuous class-wise confidence scores, rather than from the final discrete set of predicted labels. For each video, we estimate a real-valued score for each of the 157 action classes in the Charades taxonomy. In the direct video-to-label setting, the VideoLLM is prompted separately for each candidate action to assess whether that action is present in the video. In the description-based setting, the video is first converted into a textual description, and a LLM is then prompted with the same binary question for each action class to determine whether the action is present in the generated description.

For each binary query, the logits associated with the tokens `Yes` and `No` are extracted, and the softmax-normalized probability of `Yes` is used as the confidence score for the corresponding class. Repeating this procedure over all action classes produces a 157-dimensional score vector for each video. Average Precision is then computed independently for each class by ranking all videos according to their confidence scores and comparing the resulting ranking with the

binary Charades ground-truth labels. The final mAP is obtained by averaging the class-wise AP values over all classes with at least one positive instance.

Precision, Recall, and F1 are computed differently: they are based on the discrete set of action labels explicitly generated by the classifier and parsed from its textual output. For the full multi-label Charades evaluation, these metrics are computed in a micro-averaged manner by accumulating true positives, false positives, and false negatives over all videos and all action classes before applying the standard Precision, Recall, and F1 formulas. Therefore, the mAP evaluation relies on continuous confidence scores derived from token logits, whereas Precision/Recall/F1 rely on discrete predicted labels.

4 Results

This section reports the experimental results obtained on the Charades dataset under three complementary evaluation settings. First, we evaluate the selected VideoLLMs on Charades-24 Balanced, with the goal to measure the ability of pre-trained VideoLLMs to distinguish common actions when class imbalance and the size of the label taxonomy are controlled.

Second, we extend the evaluation to the original Charades label set, composed of 157 action classes. This setting is more challenging because it preserves the full taxonomy of the dataset, including fine-grained, object-dependent, and partially overlapping actions. For this evaluation, we consider the two prompt-based strategies: *video-to-label* and *description-based* approaches. Comparing these two strategies allows us to assess whether an intermediate textual representation helps bridge the gap between visual content and fine-grained action labels.

Finally, we compare our results with representative literature baselines evaluated on Charades. Since most prior works report results using the standard multi-label Charades protocol, the comparison is mainly based on mean Average Precision (mAP). This analysis is used to position the proposed zero-shot VideoLLM evaluation with respect to task-oriented action recognition methods from the literature.

4.1 Charades-24 Balanced

Table 2 reports the overall performance obtained by the tested VideoLLMs on the Charades-24 Balanced dataset. In this setting, each model performs classification directly from the input video, predicting the action labels associated with the observed activity. For the zero-shot evaluation, only the 200 test clips available for each selected class are used. No training samples from Charades-24 Balanced are provided to the models. This makes the task particularly challenging, since the selected classes include complex human actions that are not always visually distinct. In several cases, different actions may share similar objects, scenes, or motion patterns, and some labels can partially overlap from a semantic or visual point of view. The results show that zero-shot action recognition remains

difficult for general-purpose VideoLLMs. Although the models can identify high-level visual and semantic cues in the videos, their predictions are not always aligned with the fine-grained action labels defined in Charades. This is reflected in the gap between precision and recall, suggesting that the models may either miss relevant actions or produce labels that are only partially consistent with the ground truth.

Model	Precision	Recall	F1
Zero-Shot Models			
Qwen2.5-VL-72B	0.3121	0.2888	0.2659
Qwen3-VL-32B	0.3490	0.3025	0.2683
InternVL3-78B (INT8)	0.3749	0.3102	0.2815
InternVL3.5-38B	**0.3625**	**0.3344**	**0.3068**
Qwen3-VL-8B	0.3376	0.3067	0.2909
Fine-Tuned Model			
Qwen3-VL-8B (FT)	**0.4429**	**0.4198**	**0.4153**

Table 2. Overall performance of the Direct Video-to-label pipeline on the Charades-24 balanced dataset.

4.2 Full Charades dataset

Model	Accuracy	Precision	Recall	F1
Zero-Shot Models				
Qwen3-VL-32B	-	0.4142	**0.3808**	0.3968
InternVL3.5-38B	-	**0.6076**	0.3043	**0.4055**
Qwen3-VL-8B	-	0.4369	0.2975	0.3540
Fine-Tuned Model				
Qwen3-VL-8B (FT)	-	**0.7325**	**0.4866**	**0.5847**

Table 3. Overall performance on the Charades dataset across all 157 classes. In this approach, models perform classification directly from the video by predicting the class labels of the observed actions.

Table 3 reports the results on the full Charades label space using the direct video-to-label protocol. Among the zero-shot models, InternVL3.5-38B achieves the highest precision, showing a more selective behaviour, but with lower recall. In contrast, Qwen3-VL-32B provides a more balanced trade-off between precision and recall, while Qwen3-VL-8B obtains the weakest zero-shot performance among the evaluated models.

The comparison between zero-shot and fine-tuned Qwen3-VL-8B clearly shows the impact of supervised adaptation. Fine-tuning substantially improves precision, recall, and F1-score, allowing the smaller model to outperform all zero-shot

configurations, including larger models. These results indicate that pre-trained VideoLLMs provide useful visual-semantic priors, but task-specific adaptation remains important for aligning predictions with the full Charades action taxonomy.

Model	Accuracy	Precision	Recall	F1
Zero-Shot Models				
Qwen3-VL-32B	-	0.6728	**0.4973**	**0.5719**
InternVL3.5-38B	-	0.7096	0.4195	0.5273
Qwen3-VL-8B	-	0.7219	0.4619	0.5633
Fine-Tuned Model				
Qwen3-VL-8B (FT)	-	**0.7897**	0.3056	0.4407

Table 4. Overall performance on the Charades dataset across all 157 classes. In this approach, models first generate a textual description of the video; the final classification is then performed by Gemma-4-31B-it using the generated description as input.

Table 4 reports the results obtained on the full Charades label space using the description-based pipeline. In this setting, Qwen3-VL-32B achieves the best zero-shot balance between precision and recall, resulting in the highest F1-score among the zero-shot configurations. InternVL3.5-38B reaches higher precision, but its lower recall indicates a more selective behaviour, with fewer but more reliable predictions.

Qwen3-VL-8B also performs competitively in the zero-shot setting, despite its smaller size, with results close to those of Qwen3-VL-32B. This suggests that, when the generated description contains explicit action-related information, the description-based pipeline can partially reduce the gap due to model scale.

The comparison between zero-shot and fine-tuned Qwen3-VL-8B shows a different trend from the direct video-to-label setting. Fine-tuning increases precision but substantially reduces recall, leading to a lower F1-score. This indicates that the fine-tuned model becomes more conservative when classification is based on generated descriptions: its predictions are more reliable, but less complete.

Overall, textual descriptions provide an effective intermediate representation for zero-shot action recognition on Charades, but their usefulness depends on the amount and specificity of action-related information preserved in the generated text. The fact that the best zero-shot description-based configuration outperforms the fine-tuned Qwen3-VL-8B in terms of F1-score shows that fine-tuning does not necessarily improve performance when the final decision relies on textual descriptions rather than direct video evidence.

Figure 3 compares the direct video-to-label protocol and the description-based pipeline on the full 157-class Charades setting. Overall, the description-based strategy improves zero-shot performance, especially in terms of precision and F1-score, suggesting that textual descriptions can provide a useful semantic bridge between video content and action labels. The two approaches show different precision-recall trade-offs. In the direct setting, InternVL3.5-38B is more

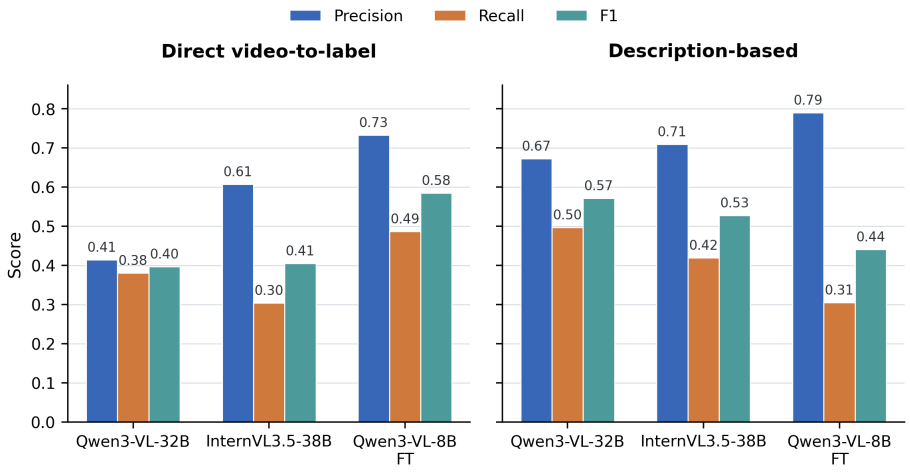

Fig. 3. Micro-averaged Precision, Recall, and F1-score on the full 157-class Charades evaluation. The two panels compare the direct video-to-label pipeline and the description-based pipeline, highlighting the different precision-recall trade-offs produced by the evaluated VideoLLM configurations.

selective, with high precision but lower recall, while Qwen3-VL-32B provides a more balanced behaviour. In the description-based setting, precision generally increases, and Qwen3-VL-32B achieves the best zero-shot balance, reaching the highest recall and F1-score. The comparison with Qwen3-VL-8B highlights the role of fine-tuning. In the direct video-to-label setting, fine-tuning produces the best overall result. In contrast, in the description-based setting, the fine-tuned model achieves the highest precision but a much lower recall, reducing its F1-score. This indicates that fine-tuning makes predictions more conservative when classification relies on generated descriptions. Overall, the figure shows that the description-based pipeline is effective for zero-shot recognition, while the best overall performance is obtained by the fine-tuned Qwen3-VL-8B in the direct video-to-label setting.

Figures 4 and 5 provide a class-level comparison between the two best Qwen-based configurations: the fine-tuned Qwen3-VL-8B model in the direct video-to-label setting and the Qwen3-VL-32B model in the description-based pipeline.Figure 4 shows the five best-performing classes for the two approaches. In both cases, the strongest results are obtained for actions with clear object-action associations and distinctive visual cues. For example, classes such as *drinking from a cup/glass/bottle* are easier to recognize because the relevant object and the associated gesture provide strong evidence for the correct label. In the direct video-to-label setting, fine-tuning helps the model align these visual patterns with the Charades taxonomy. In the description-based setting, the same type of action is correctly recognized when the generated description explicitly contains discriminative semantic cues. Figure 5 reports the five worst-performing

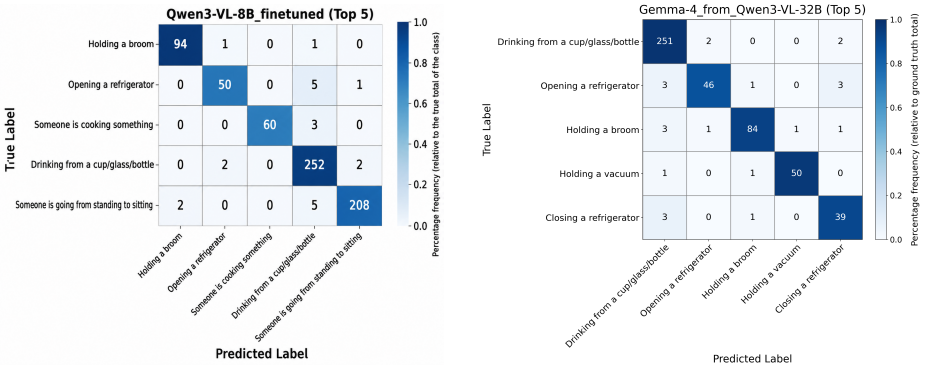

Fig. 4. Top-5 class-level results for the Qwen-based evaluations. The two matrices report the best-performing classes obtained by the top-performing model in each evaluation setting: the direct video-to-label approach on the left and the description-based pipeline on the right.

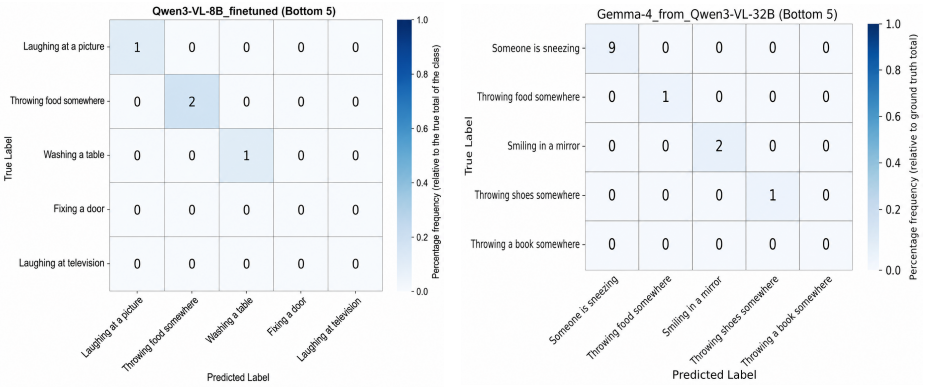

Fig. 5. Bottom-5 class-level results for the Qwen-based evaluations. The two matrices report the worst-performing classes obtained by the top-performing model in each evaluation setting: the direct video-to-label approach on the left and the description-based pipeline on the right.

classes. The weakest results are associated with actions that are subtle, short, or strongly dependent on context. For example, *throwing food somewhere* is difficult because the action may occur briefly and can be missed if the model focuses on static objects or scene content. This limitation affects both approaches, but for different reasons: the direct method may fail to capture fine temporal evidence, while the description-based method may omit the relevant action from the generated text. Overall, the comparison confirms that both strategies work better for visually explicit object-centred actions and struggle with fine-grained or temporally localized events. Overall, the two figures highlight a consistent trend. Both approaches perform best on actions with strong object-action associations and distinctive visual evidence, while they fail on actions requiring fine temporal localization, subtle human behaviour interpretation, or context-dependent reasoning. The main difference is that the direct video-to-label approach benefits from fine-tuning and can learn dataset-specific visual-label associations, whereas the description-based approach benefits from explicit textual descriptions but is limited by what the VideoLLM decides to mention. Therefore, the description-based pipeline can improve recognition for semantically explicit actions, but it may miss actions that are visually present yet not verbalized in the generated description.

4.3 Comparison with Literature Baselines

To contextualize the obtained results, we compare our findings with different representative literature baselines evaluated on the Charades dataset, reported in Fig. 6. Since these methods differ in architecture, supervision, input processing, and evaluation protocol, the comparison should be interpreted as a high-level positioning rather than as a strictly controlled benchmark.

The results show that the direct video-to-label strategy achieves the best overall performance among the reported methods in this comparison. In particular, InternVL3.5-38B with direct video-to-label method obtains the highest mAP, exceeding both the selected literature baselines and the other tested configurations under the reported evaluation setup. The comparison also highlights a clear difference between the two proposed strategies. The description-based pipeline improves over several earlier methods, but remains below ActionCLIP and below the direct video-to-label configurations. This suggests that the intermediate textual description is useful, but may introduce an information bottleneck: if the generated description omits subtle motion cues or does not match the Charades label vocabulary, the final classification stage cannot recover the missing information. By contrast, the direct video-to-label approach appears more effective for this benchmark. Both Qwen3-VL-8B and Qwen3-VL-32B in the direct setting obtain higher mAP than the description-based variants, indicating that predicting labels directly from the video better preserves the visual and temporal evidence needed for action recognition. Overall, these results indicate that general-purpose VideoLLMs can be competitive with established action recognition methods on Charades when they are prompted to perform direct label

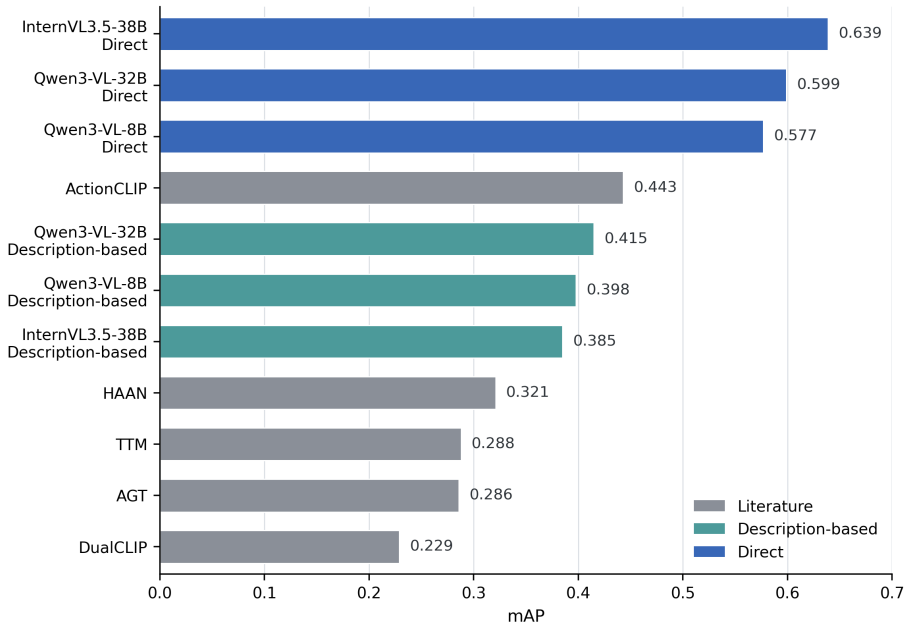

Fig. 6. mAP comparison on Charades across representative literature baselines and the proposed VideoLLM evaluation pipelines. The plot separates prior methods, description-based and direct video-to-label classification, showing that the best direct VideoLLM configuration achieves the highest mAP among the reported results.

prediction. The best result also suggests that, for this task, direct video understanding is more effective than relying on a textual intermediate representation.

5 Conclusion

This work assessed the ability of general-purpose VideoLLMs to perform prompt-based action recognition on Charades, without introducing a task-specific architecture. The experiments show that pre-trained VideoLLMs can produce meaningful action predictions, but their performance depends strongly on the evaluation setting and on the adopted prediction strategy.

On the full Charades taxonomy, the task remains challenging because of fine-grained, object-dependent, and overlapping actions. The direct video-to-label approach achieved the best overall results, especially after fine-tuning Qwen3-VL-8B, confirming the benefit of task-specific adaptation. Among zero-shot models, InternVL3.5-38B achieved high precision, while Qwen3-VL-32B showed a more balanced precision-recall behaviour.

The description-based pipeline proved useful in the zero-shot setting, but its effectiveness depends on the quality of the generated descriptions. When

relevant motion cues or object-action relations are omitted, the final classifier cannot recover the missing information. Class-level results confirm that both approaches work better on visually explicit object-centred actions, while they struggle with subtle, short, or context-dependent actions.

Overall, the results indicate that VideoLLMs are promising tools for flexible action recognition on Charades, but robust fine-grained recognition still benefits from task-specific adaptation and stronger temporal reasoning.

Bibliography

[1] Ali, M., Yang, D., Brémond, F.: Current challenges on fine-grained human action recognition. arXiv preprint arXiv:2410.17149 (2024), https://arxiv.org/abs/2410.17149

[2] Bai, S., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025). https://doi.org/10.48550/arXiv.2511.21631, https://arxiv.org/abs/2511.21631

[3] Calabrese, C., Berti, S., Pasquale, G., Natale, L.: The impact of compositionality in zero-shot multi-label action recognition for object-based tasks. In: Proceedings of the 33rd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN). pp. 857–863 (2024)

[4] Dehghani, M., Gritsenko, A., Arnab, A., Minderer, M., Tay, Y.: Scenic: A jax library for computer vision research and beyond. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 21393–21398 (2022)

[5] Dutta, S.J., Boongoen, T., Zwiggelaar, R.: Human activity recognition: A review of deep learning-based methods. IET Computer Vision **19**(1), e70003 (2025)

[6] Gao, Z., Qiao, P., Dou, Y.: Haan: Human action aware network for multi-label temporal action detection. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 5059–5069 (2023)

[7] Google DeepMind: Gemma 4 model overview. Google AI for Developers documentation (2026), https://ai.google.dev/gemma/docs/core, documentation for the Gemma 4 open model family

[8] Karim, M., Khalid, S., Lee, S., Almutairi, S., Namoun, A., Abohashrh, M.: Next generation human action recognition: A comprehensive review of state-of-the-art signal processing techniques. IEEE Access (2025)

[9] Nawhal, M., Mori, G.: Activity graph transformer for temporal action localization. arXiv preprint arXiv:2101.08540 (2021)

[10] Sigurdsson, G.A., Varol, G., Wang, X., Farhadi, A., Laptev, I., Gupta, A.: Hollywood in homes: Crowdsourcing data collection for activity understanding. In: European conference on computer vision. pp. 510–526. Springer (2016)

[11] Wang, M., Xing, J., Liu, Y.: Actionclip: A new paradigm for video action recognition. CoRR **abs/2109.08472** (2021), https://arxiv.org/abs/2109.08472

[12] Wang, W., et al.: Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265 (2025). https://doi.org/10.48550/arXiv.2508.18265, https://arxiv.org/abs/2508.18265

[13] Wu, J., Liu, W., Liu, Y., Liu, M., Nie, L., Lin, Z., Chen, C.W.: A survey on video temporal grounding with multimodal large language model. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)