LLAVIDAL : Benchmarking Large LAnguage VIsion Models for Daily Activities of Living

Abstract

With the increasing pervasiveness of video content throughout society, the demand for robust video-language models is increasingly urgent. In this work we introduce LLAVIDAL, a Large Language Vision Model tailored for Activities of Daily Living (ADL). Unlike existing models primarily trained on curated web videos, LLAVIDAL leverages a novel multiview RGB-D dataset, ADL-X, which includes 100K untrimmed video-instruction pairs, enriched with 3D skeletons and object trajectories to mimic real-world complexities. The model integrates these features to effectively understand intricate human behaviors and spatiotemporal dynamics typical of daily activities. We also introduce ADLMCQ, a new benchmark designed to evaluate the proficiency of video-language models in interpreting ADL content. Our evaluations demonstrate that LLAVIDAL significantly outperforms existing models, showcasing superior ability to process and reason about real-life video scenarios. The insights gained underscore the necessity for advanced processing techniques to handle the scale and multimodality of video data, alongside a need for comprehensive benchmarks that reflect real-world use cases more accurately. The instruction tuning data is available at Link.

1 Introduction

Large Language Vision Models (LLVMs) have made significant strides in processing and understanding internet videos [1, 2, 3, 4, 5], showcasing impressive capabilities that seem to challenge human intelligence. However, these models face substantial challenges when confronted with the complex, nuanced dynamics present in Activities of Daily Living (ADL) [6, 7, 8, 9, 10, 11, 12]. This limitation exposes a critical gap between the apparent sophistication of these AI systems and true general intelligence, particularly in real-world scenarios. The struggle of LLVMs with ADL stems from multiple factors: the lack of suitable datasets, the absence of models tailored to capture relevant cues, and most importantly, the inability to perform multimodal algorithmic reasoning required for understanding everyday human activities. ADL videos present unique challenges including multiple exocentric viewpoints, fine-grained activities with subtle motions, complex human-object interactions, and long-term temporal relationships. These aspects demand a level of perception and reasoning that goes beyond simple pattern recognition, touching on the core aspects of general intelligence. To address these challenges, we propose LLAVIDAL, a novel LLVM specifically designed for ADL understanding. LLAVIDAL integrates multiple modalities - video, 3D poses, and object cues - into a unified framework, demonstrating an approach to multimodal reasoning that more closely mimics human cognitive processes. This integration allows for a more nuanced understanding of spatial-temporal relationships and human-object interactions, key components in decoding the complexities of daily activities. Furthermore,we introduce ADLMCQ, a new benchmark for assessing LLVM performance in ADL scenarios. These tools not only facilitate the development of more capable AI systems but also provide a means to rigorously evaluate their performance, helping to illuminate the gap between current AI capabilities and human-like understanding. Through this work, we aim to contribute to the ongoing discussion about the foundations of general intelligence in AI systems. By focusing on the challenging domain of ADL, we hope to highlight both the progress made in multimodal reasoning and the significant hurdles that remain in achieving true artificial general intelligence that can match human cognitive capabilities in real-world scenarios.

Workshop on Video-Language Models at 38th Conference on Neural Information Processing Systems (NeurIPS 2024).

2 Related work

Recent advancements in Large Language Vision Models (LLVMs) have significantly improved video understanding [13, 14, 15, 16, 17, 18, 19] and dialogue capabilities. Datasets like VideoChat[13], Valley[3], VideoChatGPT[16], and TimeChat[17] have been instrumental in this progress. However, these datasets often lack the complexity and extended temporal nature required for understanding Activities of Daily Living (ADL). Existing models like VideoChatGPT[16], VideoLLaVA[14], and TimeChat[17] typically employ various strategies to integrate video information with language models. For instance, VideoChatGPT uses both temporal and spatial features of a video, while VideoLLaVA pre-aligns visual modalities to language using LanguageBind[20] encoders. TimeChat introduces a timestamp-aware frame encoder for temporal information. Despite these advancements, current LLVMs struggle with the intricate object interactions, fine-grained actions, and long-term temporal dependencies characteristic of ADL. This limitation stems from insufficient task coverage in training datasets and the lack of real-world complexity in existing video understanding frameworks, highlighting the need for specialized approaches in ADL comprehension.

3 LLAVIDAL: A Comprehensive Multimodal LLVM for ADL Understanding

LLAVIDAL adapts large language vision models (LLVMs) for understanding activities of daily living (ADL). Our proposed architecture transcends traditional RGB video inputs through the integration of 3D human pose information and object interaction cues through a multi-faceted approach.

At its core, LLAVIDAL builds upon the established LLVM paradigm [16]. The input video $V_i \in \mathbb{R}^{T\times H\times W\times C}$ is encoded using a pretrained CLIP ViT-L/14[21], yielding frame-level embeddings $x_i \in \mathbb{R}^{T\times h\times w\times D}$. These embeddings are aggregated temporally and spatially to derive video-level features $V_i \in \mathbb{R}^{F_v\times D_v}$, which are then projected into the LLM embedding space via the transformation T_v, resulting in $Q_v = T_v(V_i) \in \mathbb{R}^{F_v\times K}$, where $F_v = 356$ and $K = 4096$.

The integration of pose information in LLAVIDAL is achieved through three complementary methods. First, as QA pairs, where 3D joint coordinates and associated human actions are processed by GPT-3.5 Turbo [22] to generate descriptive QA pairs for LLM instruction tuning. Second, as context, where motion descriptions of five key peripheral joints are appended to the text query, forming an enriched query $Q_t^{new} = [Q_t^p Q_t]$. Third, as features, where 3D pose sequences $P_i \in \mathbb{R}^{T_p\times 3\times J}$ are encoded using a PoseCLIP model. Initially, the pose backbone [23] is pretrained on the NTU RGB+D dataset [24] for action classification.The pose embeddings are obtained as $z_i^p = \frac{1}{T_p}\sum f_p(P_i)$ and $z_i^t = f_t(t_i)$, where f_p is a Hyperformer pose encoder [23] and f_t is a frozen CLIP text encoder. These embeddings are aligned using a contrastive loss $L_{CE}(z_i^p, z_i^t) = -\sum_i \log \frac{\exp(\text{sim}(z_i^p, z_i^t)/\tau)}{\sum_j \exp(\text{sim}(z_j^p, z_j^t)/\tau)}$. The resulting pose features $P_i \in \mathbb{R}^{F_p\times D_p}$, where $F_p = 256$ and $D_p = 216$, are projected using a Random Projection as $Q_p = T_p(P_i)$.

Object information is integrated through parallel methods. QA pairs based on object trajectory coordinates are generated for instruction tuning. As context, relevant object labels Q_t^o are appended to each text query token. As features, objects detected by BLIP-2[25] and tracked by OWLv2[26] yield features $O_i \in \mathbb{R}^{8n\times D_o}$ for n objects across 8 sampled frames, where $D_o = 512$, projected into the LLM space as $Q_o = T_o(O_i)$.

The final input to the LLM concatenates the text query tokens with the projected video, pose, and object tokens: [USER: $\langle Q_t\rangle\langle Q_v\rangle\langle Q_o\rangle\langle Q_p\rangle$ Assistant:]. LLAVIDAL uses Vicuna v1.1 (7B)[27] as the base LLM, with its parameters frozen during training. Since the pose and object features are extracted from language grounded models, the projectors T_p, and T_o are also frozen during training. Only the projection layer T_v is optimized, allowing for efficient adaptation of the video features to the LLM space without altering the pretrained language knowledge.

The model is trained for 3 epochs with a batch size of 32 and a learning rate of $2e^{-5}$ using the Adam optimizer on 8 A6000 48GB GPUs, taking approximately 40 GPU hours. During inference, LLAVIDAL processes the input video through the vision encoder and eliminates the need for additional pose or object features. The resulting features are projected into the LLM space and concatenated with the text query, enabling the LLM to generate responses based on the input.

This comprehensive integration enables LLAVIDAL to leverage multiple modalities and representation methods, facilitating a more profound understanding of ADL scenarios and pushing the boundaries of multimodal AI in real-world applications. The model's ability to process and integrate diverse information sources allows it to capture subtle nuances in human activities, object interactions, and contextual details, making it particularly well-suited for understanding complex ADL scenarios.

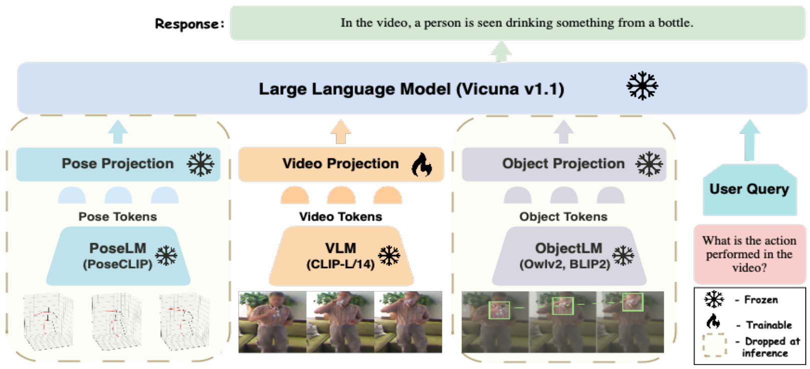

Figure 1: Overview of **LLAVIDAL**, which utilizes an LLM to integrate multiple modalities, including video, pose, and object features. Videos are represented by embeddings obtained from a **VLM**, poses are processed through **(PoseLM)**, and object embeddings are obtained through **(ObjectLM)**. These embeddings are projected into the LLM space, where they are concatenated with tokenized text queries for instruction tuning.

4 Experimental Evaluation

We evaluate LLAVIDAL on ADL tasks using metrics for video description generation, Mementos evaluation [28], and our novel ADLMCQ benchmarks. Datasets include Charades [29], Toyota Smarthome [30], LEMMA [31], and TSU [32]. Table 1 presents the impact of integrating pose and object cues into LLAVIDAL. Pose features (PF) outperform pose context (PC) and QA approaches, suggesting effective language contextualization. Object features (OF) derived from ObjectLM yield superior performance across most tasks, highlighting their significance in ADL understanding.

Notably, LLAVIDAL with OF surpasses the PF model on all tasks. However, combining PF and OF results in performance convergence towards the PF-only model, possibly due to challenges in optimizing the projection layer T_v to align with both T_p and T_o effectively. Multi-cue integration remains an open challenge for future work. Given its superior performance, we employ LLAVIDAL with OF for subsequent experiments. Detailed Experiments are in Appendix **??**

Table 1: Performance of LLAVIDAL with Pose and Object Cues

Method	ADLMCQ-AR		ADLMCQ-AF		AD (Charades)		AD (TSU)	
	Charades	Smarthome	LEMMA	TSU	Object	Action	Object	Action
Pose QA	48.5	49.0	42.0	21.2	31.8	14.0	16.5	15.9
Pose Context (PC)	50.8	54.0	45.0	22.3	30.5	**14.8**	18.6	15.4
Pose Features (PF)	56.7	**57.0**	**51.3**	26.0	**32.7**	13.5	18.2	13.0
PC + PF	52.5	53.1	44.6	24.9	32.1	13.6	17.5	15.6
Object QA	51.1	50.1	40.3	23.0	32.1	13.7	17.0	16.0
Object Context	44.6	46.2	41.8	21.0	31.2	**14.7**	17.2	16.5
Object Features (OF)	**59.0**	**58.8**	**52.6**	**27.0**	**33.1**	14.3	18.0	**17.7**
PF + OF	56.2	56.1	51.0	26.6	30.4	14.1	**20.0**	14.1

5 Conclusion & Future Work

LLAVIDAL, a novel LLVM, integrates 3D poses and human-object interaction cues to enhance ADL understanding. Evaluated using our proposed ADLMCQ benchmark, LLAVIDAL outperforms existing baselines, demonstrating superior temporal reasoning in ADL scenarios. Future work will explore innovative training strategies to more effectively combine pose and object cues within the LLAVIDAL framework.

References

[1] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *IEEE International Conference on Computer Vision*, 2021.

[2] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–970, 2015.

[3] Ruipu Luo, Ziwang Zhao, Min Yang, Junwei Dong, Minghui Qiu, Pengcheng Lu, Tao Wang, and Zhongyu Wei. Valley: Video assistant with large language model enhanced ability. *arXiv preprint arXiv:2306.07207*, 2023.

[4] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina González, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jáchym Kolář, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbeláez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18995–19012, June 2022.

[5] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In *ICCV*, 2019.

[6] Fabien Baradel, Christian Wolf, Julien Mille, and Graham W. Taylor. Glimpse clouds: Human activity recognition from unstructured feature points. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[7] Fabien Baradel, Christian Wolf, and Julien Mille. Human activity recognition with pose-driven attention to rgb. In *The British Machine Vision Conference (BMVC)*, September 2018.

[8] Srijan Das, Saurav Sharma, Rui Dai, Francois Bremond, and Monique Thonnat. Vpn: Learning video-pose embedding for activities of daily living. In *European Conference on Computer Vision*, pages 72–90. Springer, 2020.

[9] Srijan Das, Rui Dai, Di Yang, and Francois Bremond. Vpn++: Rethinking video-pose embeddings for understanding activities of daily living. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021.

[10] Dominick Reilly and Srijan Das. Just add π! pose induced video transformers for understanding activities of daily living. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024.

[11] Saurabh Gupta, Judy Hoffman, and Jitendra Malik. Cross modal distillation for supervision transfer. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2827–2836, 2015.

[12] Xiaolong Wang and Abhinav Gupta. Videos as space-time region graphs. In *Proceedings of the European conference on computer vision (ECCV)*, pages 399–417, 2018.

[13] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023.

[14] Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023.

[15] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *ArXiv*, abs/2306.02858, 2023.

[16] Muhammad Maaz, Hanoona Abdul Rasheed, Salman H. Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *ArXiv*, abs/2306.05424, 2023.

[17] Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. *arXiv preprint arXiv:2312.02051*, 2023.

[18] Kevin Qinghong Lin, Alex Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Zhongcong Xu, Difei Gao, Rongcheng Tu, Wenzhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. *arXiv preprint arXiv:2206.01670*, 2022.

[19] Antoine Yang, Arsha Nagrani, Paul Hongsuck Seo, Antoine Miech, Jordi Pont-Tuset, Ivan Laptev, Josef Sivic, and Cordelia Schmid. Vid2seq: Large-scale pretraining of a visual language model for dense video captioning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2023.

[20] Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, Hongfa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Wancai Zhang, Zhifeng Li, Wei Liu, and Liejie Yuan. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. *ArXiv*, abs/2310.01852, 2023.

[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

[22] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *ArXiv*, abs/2005.14165, 2020.

[23] Yuxuan Zhou, Zhi-Qi Cheng, Chao Li, Yifeng Geng, Xuansong Xie, and Margret Keuper. Hypergraph transformer for skeleton-based action recognition. *arXiv preprint arXiv:2211.09590*, 2022.

[24] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis. In *IEEE Conf. Comput. Vis. Pattern Recog.*, June 2016.

[25] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, 2023.

[26] Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 72983–73007. Curran Associates, Inc., 2023.

[27] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023.

[28] Xiyao Wang, Yuhang Zhou, Xiaoyu Liu, Hongjin Lu, Yuancheng Xu, Fuxiao Liu, Feihong He, Jaehong Yoon, Taixi Lu, Gedas Bertasius, Mohit Bansal, Huaxiu Yao, and Furong Huang. Mementos: A comprehensive benchmark for multimodal large language model reasoning over image sequences. *ArXiv*, abs/2401.10529, 2024.

[29] Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding. In *European Conference on Computer Vision(ECCV)*, 2016.

[30] Srijan Das, Rui Dai, Michal Koperski, Luca Minciullo, Lorenzo Garattoni, Francois Bremond, and Gianpiero Francesca. Toyota smarthome: Real-world activities of daily living. In *Int. Conf. Comput. Vis.*, 2019.

[31] Baoxiong Jia, Yixin Chen, Siyuan Huang, Yixin Zhu, and Song-Chun Zhu. Lemma: A multiview dataset for learning multi-agent multi-view activities. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[32] Rui Dai, Srijan Das, Saurav Sharma, Luca Minciullo, Lorenzo Garattoni, Francois Bremond, and Gianpiero Francesca. Toyota Smarthome Untrimmed: Real-World Untrimmed Videos for Activity Detection. *arXiv preprint arXiv:2010.14982*, 2020.