CReLeRI: Explainable, Concept-centric, Representation, Learning, Reasoning, and Interaction Video Analysis System

Michael Pérez
michaelperez012@ufl.edu
UF
Gainesville, USA

Yichi Yang
yiy067@ufl.edu
UCSD
San Diego, USA

Yuheng Zha
UCSD
San Diego, USA

Enze Ma
UCSD
San Diego, USA

Danish Tamboli
UF
Gainesville, USA

Haodi Ma
UF
Gainesville, USA

Reza Shahriari
UF
Gainesville, USA

Vyom Pathak
UF
Gainesville, USA

Dzmitry Kasinets
UF
Gainesville, USA

Rohith Venkatakrishnan
UF
Gainesville, USA

Daisy Zhe Wang
daisyw@cise.ufl.edu
UF
Gainesville, USA

Jaime Ruiz
jaime.ruiz@ufl.edu
UF
Gainesville, USA

Eric Ragan
UF
Gainesville, USA

Zhiting Hu
zhh019@ucsd.edu
UCSD
San Diego, California, USA

Eric Xing
CMU
Pittsburgh, USA

Jun-Yan Zhu
CMU
Pittsburgh, USA

Abstract

Existing video analysis models often lack explainability, struggle with long videos, and hallucinate. Commercial solutions are closed-source and costly. We introduce *CReLeRI*, an open-source[1] system for action detection in untrimmed videos. CReLeRI integrates segmentation, action detection, argument detection, and grounding to improve interpretability and reduce hallucinations, enhancing transparency and trust in AI-driven video analysis. This paper is accompanied by a demonstration video[2].

CCS Concepts

• **Computing methodologies** → **Activity recognition and understanding**; **Video segmentation**; **Visual inspection**.

[1]https://github.com/michaelperez023/creleri-video
[2]https://youtu.be/XDCue9EYNTU

MM '25, Dublin, Ireland

ACM ISBN 978-1-4503-XXXX-X/2025/10
https://doi.org/XXXXXXX.XXXXXXX

Keywords

Multimedia Interaction, Video Action Detection, Object Detection, Interpretability, Grounding, Vision-Language Models, Large Language Models

ACM Reference Format:
Michael Pérez, Yichi Yang, Yuheng Zha, Enze Ma, Danish Tamboli, Haodi Ma, Reza Shahriari, Vyom Pathak, Dzmitry Kasinets, Rohith Venkatakrishnan, Daisy Zhe Wang, Jaime Ruiz, Eric Ragan, Zhiting Hu, Eric Xing, and Jun-Yan Zhu. 2025. CReLeRI: Explainable, Concept-centric, Representation, Learning, Reasoning, and Interaction Video Analysis System. In *Proceedings of (MM '25).* ACM, New York, NY, USA, 4 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

The rapid growth of video content in recent years has made manual analysis unfeasible. Vast amounts of data including satellite imagery [9], medical imaging [7], and social media [18] must be processed efficiently. To address this challenge, AI-powered video analysis systems have been developed to assist human analysts.

However, many existing video analysis models have critical limitations. Many systems [2, 10, 13, 15] function as black boxes, offering minimal transparency in their decision-making process [16]. Users need tools that not only classify actions but also justify predictions to improve trust and usability. We define explainability as a

system's ability to provide human-interpretable evidence behind its predictions, enabling users to understand why outputs are generated and assess their validity. Many state-of-the-art vision language models [2, 10, 13, 15] are optimized for trimmed video clips and cannot effectively process unstructured, continuous video streams, limiting their applicability to real-world scenarios. Such models often hallucinate [8], generating outputs that appear plausible but are fabricated and are not grounded in the input video.

To address the limitations of existing systems, we developed a novel image and video analysis tool called Concept-centric Representation, Learning, Reasoning, and Interaction (CReLeRI). While the tool supports both images and videos, this paper focuses on the video component. The system was designed with explainability at its core, aiming to provide users with visual justifications for predicted actions. To achieve this, we overlay segmentation outlines on the video, highlighting detected actions, associated arguments, and their physical points of contact. A key innovation of CReLeRI is its video segmentation pipeline, which partitions untrimmed videos into manageable clips based on both scene changes and action shifts, allowing efficient downstream processing. Additionally, to reduce hallucinations commonly observed in vision-language models, we integrated a grounding mechanism that verifies spatial consistency between predicted arguments and their locations in 3D space. This ensures that recognized actions are not only semantically plausible but also physically grounded in the visual input.

2 System Design

Before diving into the system architecture, we briefly outline how users interact with CReLeRI. Users upload videos through the user interface, receive segmented and labeled outputs, and can explore detected actions and arguments through interactive visualizations. This design prioritizes modularity and scalability to handle untrimmed videos of varying lengths and sources, ensuring usability in real-world contexts where videos are long and complex.

2.0.1 Architecture. The system consists of modular back-end APIs for image analysis, video segmentation, action recognition, and grounding, enabling rapid iteration and integration of new models. FastAPI handles API requests, while Celery and Redis manage long-running tasks, queuing multiple submissions for scalable processing. The front-end provides interactive visualizations and query controls. CReLeRI's deployment on 10 NVIDIA A100 GPUs enables fast and accurate processing of untrimmed videos: 15-second videos take 10 minutes to process, and 1-minute videos take 30 minutes.

2.1 Video Action Detection Pipeline

The CReLeRI video processing pipeline is designed to efficiently analyze untrimmed videos, detect action transitions, and provide explainable visualizations. It consists of three main components: segmentation, action recognition, and grounding, all integrated into an interactive web interface.

2.1.1 Video Segmentation. To analyze untrimmed videos, CReLeRI first identifies action boundaries by detecting two types of transitions. Scene changes, abrupt transitions between different camera shots, are detected using a simple frame difference thresholding method. Action shifts, transitions within a single camera shot, are detected using a pre-trained temporal action detection model, AdaTAD [12]. We fine-tuned this model on an untrimmed dataset that was created by stitching together trimmed videos from 100 classes. After detecting boundaries, the system refines the results to ensure that every moment in the video falls within one temporal segment, so that no portion is left unclassified.

2.1.2 Action Recognition. The D-Fine object detector [14] performs per-frame object detection in trimmed clips. StrongSort [5] tracks detected objects across frames, generating actor-specific trajectories. For each tracked object, the system creates a new video with bounding boxes overlaid on the actor's trajectory. These per-actor videos are then passed to the 72-billion-parameter Instruct VLM Qwen2.5-VL [15], which generates descriptive captions summarizing the actions of each tracked actor. Next, the 70-billion-parameter LLM Instruct LLaMA3.1 [1] parses these captions to extract action-arguments tuples. These tuples are then encoded using the MPNet [17] language model, and DBSCAN [6] clustering is applied to group semantically similar pairs based on their relative distances in a computed similarity matrix. Finally, the clustering results determine the most probable action-argument tuples.

2.1.3 Grounding. The grounding component enhances explainability by linking detected actions and objects with their physical locations in the video, ensuring 3D spatial consistency. To locate arguments in a video, this component first detects candidate objects using GroundingDino [11], which produces bounding boxes for textual labels. In parallel, Molmo [4] generates point-based annotations for arguments guided by natural language prompts created by Qwen2.5-VL [15]. Results from GroundingDino and Molmo are merged and filtered. SegmentAnything2 tracks the grounded objects throughout the video, producing segmentation masks. DepthPro [3] estimates 3D depth on sampled frames, and contact between arguments is verified by checking for proximity and physical interaction in the 3D space. Only grounded arguments that are in physical contact are retained, reducing hallucinations by ensuring spatial plausibility of the predicted actions.

2.1.4 Web User Interface. The user interface is a web application built with Flask for the server side and HTML, CSS, and Javascript for the front-end. The top-left section of the interface features a conversation area, where users can upload one or more videos and receive real-time feedback on processing status, either running or completed. The bottom section displays a timeline of the untrimmed video, with color-coded segments representing different detected actions. Clicking on a segment reveals predicted actions and their arguments below the timeline. Figure 1 shows a close-up of this. Additionally, clicking a segment plays the corresponding video segment in the top-right section with superimposed action/arguments names, segmentation outlines, and contact between arguments highlighted in yellow for visual reference and explainability. Figure 2 shows a close-up of the media viewer.

3 Live Interactive Demo

During the live demonstration, users will interact with CReLeRI by uploading videos for real-time analysis. For example, a user might upload an instructional video, like a cooking tutorial, to automatically segment the video into meaningful steps. They would examine

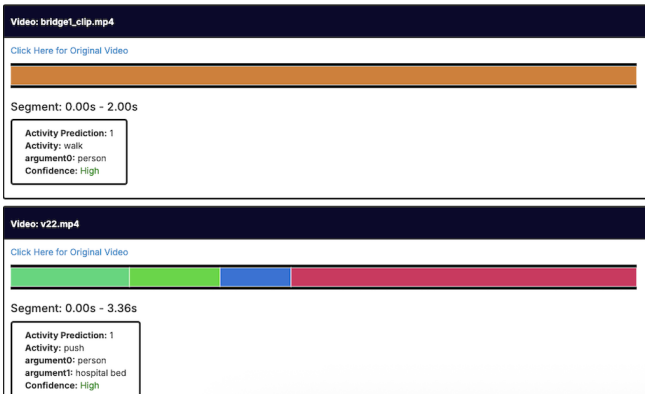

Figure 1: Concept Visualization. CReLeRI partitions untrimmed videos into segments using scene and action shift detection. This timeline color-codes segments, each labeled with predicted actions and arguments, to help users interpret long video streams.

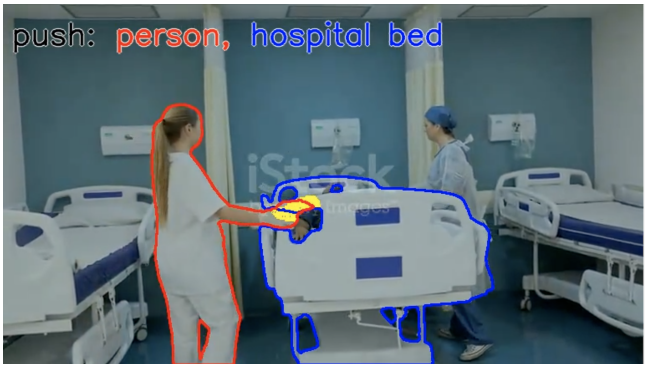

Figure 2: Media Viewer. CReLeRI overlays segmentation masks, argument labels, and contact highlights on detected actions. This visualization enhances interpretability by grounding predictions spatially and highlighting physical interactions.

the timeline showing these segmented clips, each labeled with predicted actions and associated arguments, and explore the media viewer to confirm the detected steps and physical interactions. By verifying action sequences and grounding them to specific points in the video, the user can extract a step-by-step summary of the instructional content, making it easier to follow complex procedures. A pre-recorded demonstration is also available for users to understand the working system.

Acknowledgments

This research is based upon work supported by U.S. DARPA ECOLE Program No. HR00112390063. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of DARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

[1] Meta AI. 2024. The Llama 3 Herd of Models. arXiv:2407.21783 [cs.AI] https://arxiv.org/abs/2407.21783

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millicah, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022. Flamingo: a visual language model for few-shot learning. In *Proceedings of the 36th International Conference on Neural Information Processing Systems* (New Orleans, LA, USA) *(NIPS '22)*. Curran Associates Inc., Red Hook, NY, USA, Article 1723, 21 pages.

[3] Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R. Richter, and Vladlen Koltun. 2024. Depth Pro: Sharp Monocular Metric Depth in Less Than a Second. arXiv:2410.02073 [cs.CV] https://arxiv.org/abs/2410.02073

[4] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. 2024. Molmo and PixMo: Open Weights and Open Data for State-of-the-Art Vision-Language Models. arXiv:2409.17146 [cs.CV] https://arxiv.org/abs/2409.17146

[5] Yunhao Du, Zhicheng Zhao, Yang Song, Yanyun Zhao, Fei Su, Tao Gong, and Hongying Meng. 2023. StrongSORT: Make DeepSORT Great Again. *Trans. Multi.* 25 (Jan. 2023), 8725–8737. doi:10.1109/TMM.2023.3240881

[6] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* (Portland, Oregon) *(KDD'96)*. AAAI Press, 226–231.

[7] Moomal Farhad, Mohammad Mehedy Masud, Azam Beg, Amir Ahmad, and Luai Ahmed. 2023. A Review of Medical Diagnostic Video Analysis Using Deep Learning Techniques. *Applied Sciences* 13, 11 (2023). doi:10.3390/app13116582

[8] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Trans. Inf. Syst.* 43, 2, Article 42 (Jan. 2025), 55 pages. doi:10.1145/3703155

[9] Shengyang Li, Xian Sun, Yanfeng Gu, Yixuan Lv, Manqi Zhao, Zhuang Zhou, Weilong Guo, Yuhan Sun, Han Wang, and Jian Yang. 2023. Recent Advances in Intelligent Processing of Satellite Video: Challenges, Methods, and Applications. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 16 (2023), 6776–6798. doi:10.1109/JSTARS.2023.3296451

[10] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2024. Video-LLaVA: Learning United Visual Representation by Alignment Before Projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 5971–5984. doi:10.18653/v1/2024.emnlp-main.342

[11] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. 2024. Grounding DINO: Marrying DINO with Grounded Pre-training for Open-Set Object Detection. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XLVII* (Milan, Italy). Springer-Verlag, Berlin, Heidelberg, 38–55. doi:10.1007/978-3-031-72970-6_3

[12] Shuming Liu, Chen-Lin Zhang, Chen Zhao, and Bernard Ghanem. 2024. End-to-End Temporal Action Detection with 1B Parameters Across 1000 Frames. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 18591–18601. doi:10.1109/CVPR52733.2024.01759

[13] OpenAI. 2023. Introducing GPT-4V: Multimodal Capabilities in GPT-4. OpenAI Blog. https://openai.com/research/gpt-4v Accessed: March 13, 2025.

[14] Yansong Peng, Hebei Li, Peixi Wu, Yueyi Zhang, Xiaoyan Sun, and Feng Wu. 2024. D-FINE: Redefine Regression Task in DETRs as Fine-grained Distribution Refinement. arXiv:2410.13842 [cs.CV] https://arxiv.org/abs/2410.13842

[15] Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue,

Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 Technical Report. arXiv:2412.15115 [cs.CL] https://arxiv.org/abs/2412.15115

[16] Avinab Saha, Shashank Gupta, Sravan Kumar Ankireddy, Karl Chahine, and Joydeep Ghosh. 2024. Exploring Explainability in Video Action Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 8176–8181.

[17] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. MPNet: masked and permuted pre-training for language understanding. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) *(NIPS '20)*. Curran Associates Inc., Red Hook, NY, USA, Article 1414, 11 pages.

[18] YouTube and Google. 2022. *Hours of video uploaded to YouTube every minute as of February 2022*. https://www.statista.com/statistics/259477/hours-of-video-uploaded-to-youtube-every-minute/ [Online; accessed 29-May-2025].

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009