

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

Michael Francis Perez
michaelperez012@ufl.edu
UF
Gainesville, USA

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

Yuheng Zha
Enze Ma
UCSD
San Diego, USA

Danish Tamboli
Haodi Ma
Reza Shahriari
Vyom Pathak
Dzmitry Kasinets
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 D. Ragan
UF
Gainesville, USA

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

Eric Xing
Jun-Yan Zhu
CMU
Pittsburgh, USA

Abstract

Existing video analysis models often lack explainability, perform poorly on long videos, and frequently hallucinate. Commercial solutions are closed-source and costly. We introduce *CReLeRI*, an open-source¹ system for action detection in untrimmed videos. *CReLeRI* segments videos using scene and action transitions, detects actions and their arguments and grounds them in 3D space to improve interpretability and reduce hallucinations. The system promotes transparency and trust in AI-driven analysis of complex, real-world videos. A demonstration video is also available².

CCS Concepts

• **Computing methodologies** → **Activity recognition and understanding; Video segmentation; • Human-centered computing** → *User interface design; Web-based interaction.*

Keywords

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

¹<https://github.com/michaelperez023/creleri-video>

²<https://youtu.be/XDCue9EYNTU>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

MM '25, Dublin, Ireland

© 2025 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2035-2/2025/10

<https://doi.org/10.1145/3746027.3754479>

ACM Reference Format:

Michael Francis Perez, Yichi Yang, Yuheng Zha, Enze Ma, Danish Tamboli, Haodi Ma, Reza Shahriari, Vyom Pathak, Dzmitry Kasinets, Rohith Venkatakrishnan, Daisy (Zhe) Wang, Jaime Ruiz, Eric D. 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 the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland*. ACM, New York, NY, USA, 3 pages. <https://doi.org/10.1145/3746027.3754479>

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 using scalable, automated tools. 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. These systems [2, 10, 13, 15] often function as black boxes, offering minimal transparency in their decision-making process [16]. To improve trust and usability, systems must not only classify actions but also explain their predictions through human-interpretable evidence. Many state-of-the-art vision language models [2, 10, 13, 15] are optimized for trimmed video clips and cannot process unstructured, continuous video streams, limiting their applicability to real-world scenarios. Such models often hallucinate [8], producing plausible but fabricated outputs not grounded in the input video.

To overcome these limitations, we present Concept-centric Representation, Learning, Reasoning, and Interaction (CReLeRI), a system designed for explainable video analysis. The tool highlights detected actions, their associated arguments, and points of physical

interaction using segmentation overlays. A key innovation is the video segmentation pipeline, which partitions untrimmed video into manageable clips based on both scene changes and action shifts. To mitigate hallucinations, CReLeRI includes a grounding mechanism that ensures spatial consistency between predicted arguments and their physical locations in 3D space, enhancing plausibility and supporting explainability.

2 System Design

Before describing the system architecture, we briefly outline how users interact with CReLeRI. Users upload videos and explore action segments and arguments through interactive visualizations. The system is scalable and modular, supporting diverse untrimmed videos and future research.

2.1 Architecture

The system consists of modular back-end APIs for 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 asynchronous processing. The front-end provides interactive visualizations and query controls. Deployed on 10 NVIDIA A100 GPUs, CReLeRI enables high-throughput processing of untrimmed videos: a 15-second video takes 10 minutes to analyze, and a 1-minute video takes 30 minutes.

2.2 Video Action Detection Pipeline

The CReLeRI video processing pipeline analyzes untrimmed videos by detecting transitions, recognizing actions, and grounding them in 3D space. It includes three main components: segmentation, recognition, and grounding, integrated into an interactive interface.

2.2.1 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. The system then integrates both transition types by partitioning the video based on scene changes, then further subdividing each scene using high-confidence action proposals that satisfy a minimum duration constraint. This refinement ensures complete coverage of the video timeline while preserving semantically meaningful boundaries.

2.2.2 Action Recognition. The D-Fine object detector [14] performs object detection on each frame of 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 passed to the 72-billion-parameter Instruct VLM Qwen2.5-VL [15], which generates captions summarizing the actions of each tracked actor. Next, the 70-billion-parameter LLM Instruct LLaMA3.1 [1] parses these captions to extract action-argument tuples using an instruction-tuned prompt designed to

produce structured outputs. The tuples are encoded using the MP-Net [17] language model, and DBSCAN [6] clustering is applied to group semantically similar pairs and select representative action-argument tuples.

2.2.3 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]. The results from GroundingDino and Molmo are merged and filtered. SegmentAnything2 tracks grounded objects throughout the video, producing segmentation masks. DepthPro [3] estimates 3D depth in sampled frames, and contact between arguments is verified by checking proximity and physical interaction in the 3D space. Only grounded arguments that are in physical contact are retained.



Figure 1: Media Viewer. CReLeRI overlays segmentation masks, actions, arguments, and contact highlights on videos. This visualization enhances interpretability by grounding predictions and highlighting physical interactions.

2.2.4 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. It features a conversation area for uploading one or more videos and tracking processing status, a color-coded timeline of action and scene segments, and a media viewer that visualizes actions and interactions. See the media viewer in Figure 1.

3 Live Interactive Demo

During the live demonstration, users will interact with CReLeRI by uploading videos for analysis. For example, a user might upload an instructional video, such as a cooking tutorial, to automatically segment it into steps. They would examine the timeline showing these segmented clips, each labeled with predicted actions and associated arguments. The media viewer will allow users to verify steps and interactions. By verifying action sequences and grounding them to 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 to help users understand the 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 Millican, 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 Wittliff, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. 2025. Molmo and PixMo: Open Weights and Open Data for State-of-the-Art Vision-Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 91–104.
- [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, Jiayi 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, Jiayi 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].