COMPUTER AND CONTROL ENGINEERING

DAUIN - VideoLLMs for Multimodal Content Understanding, Retrieval, and Generation

Funded By	Dipartimento DAUIN
Supervisor	CAGLIERO LUCA - luca.cagliero@polito.it
Contact	GARZA PAOLO - paolo.garza@polito.it BARALIS ELENA MARIA - elena.baralis@polito.it
Context of the research activity	VideoLLMs extend the capabilities of Large Language Models to perform advanced video understanding and generation. Cost-effective approaches incorporate visually aligned auxiliary texts (e.g., OCR, ASR) to improve the VideoLLMs' performance on complex tasks, such as summarization, action recognition, machine unlearning, question answering, and anomaly detection, without the need of retraining the original model. The PhD scholarship aims to study new cost-effective Agentic AI solutions that combine VideoLLMs with Retrieval-Augmented Generation, Lightweight Fine-Tuning, and Adversarial Learning techniques as well as their application to real-world industrial scenarios.
Objectives	Objectives: Existing video understanding and generation approaches based on VideoLLMs incorporate visually aligned auxiliary texts (e.g., OCR, ASR, object detection). Auxiliary information is extracted from the videos by a static preprocessing step that ignores the objective of LLM prompting, the characteristics of the prompted video and text, and the time-evolving nature of video content. The research aims to (1) Design, implement, and test cost-effective autonomous agents extracting, indexing, and summarizing task-dependent multimodal auxiliary information from videos; (2) Make the retrieval, elaboration, and processing of video excerpts context- and time-aware; (3) Define the most appropriate strategies to prompt VideoLLMs with the retrieved content. Tentative work plan: In the first year the PhD student will explore the state-of-the-art of VideoLLMs, define benchmarking frameworks, and analyze the impact of auxiliary information on VideoLLMs' performance. In the second year, the PhD student will address task-specific challenges, with the aim to improve understanding, retrieval and generation of video content. The addressed tasks will include (but are not limited to) video segmentation,

	video summarization, action recognition, and anomaly detection from videos, video question answering, bias detection and mitigation, video classification, video clustering and co-clustering. The researcher develops and tests ad hoc autonomous agents and studies cost-effective prompting strategies for VideoLLMs. In the last year, the research focuses on developing context- and time-aware solutions based on VideoLLMs and investigating the synergical use of VideoLLMs with other LLM types such as SpeechLLMs and VisualLLMs. List of possible publication venues: - Conferences: ACL, EMNLP, ACM Multimedia, KDD, ACL, COLING, IEEE ICDM, ECML PKDD, ACM CIKM - Journals: IEEE TKDE, ACM TKDD, IEEE TAI, ACM TIST, IEEE/ACM TASLP, ACL TACL

Skills and competencies for the development of the activity	The PhD candidate is expected to - Have the ability to critically analyze complex systems, model them and identify weaknesses; - be proficient in Python programming; - know data science fundamentals; - have a solid background on machine learning and deep learning; - have natural inclination for teamwork; - be proficient in English speaking, reading, and writing.