The American Journal of Interdisciplinary Innovations and Research
ISSN 2642-7478 | Open Access

TYPE Original Research
PAGE NO. 64-70
DOI- 10.37547/tajiir/Volume07Issue11-08

Check for updates

Ontology of Quantum Information for Efficient Visual and Language Control

Yevhen Petrov
CEO <Guardnova>
Bothell, Washington, USA

OPEN ACCESS
SUBMITTED 26 October 2025
ACCEPTED 11 November 2025
PUBLISHED 26 November 2025
VOLUME Vol.07 Issue 11 2025

CITATION
Yevhen Petrov. (2025). Ontology of Quantum Information for Efficient Visual and Language Control. The American Journal of Interdisciplinary Innovations and Research, 7(11), 64–70. https://doi.org/10.37547/tajiir/Volume07Issue11-08

Abstract: The article analyzes the architecture of intelligent video surveillance based on an ontology of information quanta. The relevance of the study is determined by tightening requirements for IoT video analytics, which must operate under limited network bandwidth, strict privacy requirements, and tight budgets. The novelty of the approach lies in tokenizing video streams at the edge: low-level visual descriptors are transformed into semantically stable information quanta (IQ), after which their cloud processing is performed by vision-language models. The paper formalizes the principles of a two-tier edge-cloud architecture and analyzes data-centric methods that increase the robustness of models to noisy data. Special attention within the work is paid to the ontology of actions as a connecting link between pose detection and subsequent semantic interpretation. The aim of the study is to demonstrate that the proposed architecture ensures ultra-low latency, compliance with privacy requirements, and high interpretability of decisions. To achieve this aim, methods of systems analysis and comparative performance analysis are employed. In conclusion, it is shown that the architecture enables efficient solving of complex event analysis tasks in real time. The material is addressed to specialists in the fields of computer vision, the Internet of Things, and security systems.

Keywords: Internet of Things, edge-cloud architecture, data-centric AI, ontology of actions, information quantum, vision-language control, low latency, pose recognition, video surveillance.

Introduction

The Internet of Things ecosystem increasingly relies on video analytics, yet it operates under strict bandwidth constraints, stringent personal data protection

requirements, and energy budget limitations. The classical paradigm that presupposes transmitting the full video stream to the cloud creates excessive network load, complicates compliance with privacy regulations, and leads to unstable latencies. At the same time, fully local processing at the edge is often unattainable due to limited computational resources. These contradictions focus attention on hybrid architectures in which perception and decision-making tasks are distributed between the edge and the cloud so that critically important services are guaranteed predictable quality-of-service (QoS) metrics.

The aim of the study is to develop and rigorously substantiate an ontological approach to video data processing in hybrid edge-cloud systems, based on the concept of quanta of information, to ensure efficient vision-language control in real time while preserving confidentiality.

Research objectives

- Analyze existing edge-cloud architectures of video analytics and identify their bottlenecks in terms of latency, privacy, and scalability.

- Formulate the concept of an ontology of actions based on quanta of information, where human behavioral primitives (fall, shooter pose, raised hands) are represented as discrete semantic tokens.

- Propose an integrated system architecture that combines data-centric principles for training models at the edge and zero-shot vision-language modeling in the cloud for composing complex events.

The scientific novelty lies in a different interpretation of video stream tokenization: instead of transmitting images to the cloud, only semantically rich, compact, and depersonalized IQ events are sent. This enables the use of powerful language models for situational analysis without access to the original frames, radically reducing network traffic while simultaneously addressing the privacy problem.

The author's hypothesis is based on the premise that discretizing a continuous video stream at the edge into a sequence of semantic quanta of information makes it possible, on the one hand, to reduce network bandwidth requirements and, on the other, to enable the application of large language models to complex event analysis with minimal latency, which is fundamentally unattainable for traditional video streaming architectures.

Materials and Methods

A comprehensive review of the scientific corpus on hybrid computing, computer vision, and natural language processing was conducted for the preparation of this article, covering foundational works and contemporary studies relevant to the stated problem.

Gan et al. (2022). establish the methodological framework of vision-language pre-training (VLP), providing a detailed systematization of contrastive image–text alignment, cross-modal transformers, modality masking, and generative tasks; their survey sets an upper-level TBox basis for an ontology of visual–linguistic entities and relations suitable for controlled scenarios (event detection, explainability, instructional prompting).

Abu Tami et al. (2024). operationalize this framework in the applied domain of road safety, demonstrating how multimodal LLMs implement automatic detection of critically dangerous events in traffic flow.

Jebur et al. (2023). classify methods for video anomaly detection (reconstructive autoencoders, predictive models, probabilistic densities/flows, graph-based and transformer variants), in essence redefining anomaly as an ontologically composite object (violations of regularities at the levels of trajectories, interactions, scene).

Diraco et al. (2023) conduct a review of action recognition in a smart environment and introduce key parameters of context awareness, personalization, and privacy: it is precisely the contextual modifiers (illumination, layout, user profile) that become meaningful attributes of the ontology of events and significantly influence the robustness of visual–linguistic control.

Huang et al. (2023) propose semantically private video surveillance services on edge infrastructure: the authors demonstrate how preselected relevant categories/events and masking of identifying attributes allow only safe subpredicates to be transferred to subsequent pipeline stages.

Silva et al. (2021) formalize energy-aware adaptive offloading of soft real-time tasks in mobile edge clouds: their policies for distributing loads between the device and the edge cloud create prerequisites for stratifying ontological inference into local fast predicates and heavy semantic deductions with consideration of SLA and energy budget.

Cho et al. (2025) introduce a lightweight predictive layer; in ontological terms this makes it possible to annotate control predicates with operational metadata (inference cost, evaluation latency) and automatically select admissible inference strategies.

Md Nur Hasan Mamun (2024) systematizes the integration of artificial intelligence and DevOps in scalable product development, outlining a systematic review of frameworks: the practice of data and model versioning, CI/CD, drift monitoring, and artifact traceability becomes an ontological process layer (relations among datasets, versions, environments, and policies).

Hamid (2023) formulates the dual engine of Industry 4.0 as a balance of model-centric and data-centric approaches (compactness and robustness are achieved through synchronous improvement of architectures and data quality)

Zha et al. (2025) analyze programmatic labeling, thereby turning data operations (construction, auditing, validation) into first-class ontological actions at the ABox level.

Despite the diversity of approaches, the studies also exhibit limitations: first, some investigations focus on the scale and universality of representations, whereas others demonstrate strict energy budgets, giving rise to tension between the completeness of semantics and the permissible cost of inference. Second, the following remain underexplored: formal languages of cost-aware inference logic linking energy/latency to the choice of visual–linguistic predicates; causal ontologies of events over multimodal latents for explainable control in critical scenarios; provable privacy guarantees at the predicate level during MLLM inference; ontologization of MLOps metadata (dataset versions, labeling policies, drift metrics) as first-class entities for audit, reproducibility, and automated compliance.

Results

The working hypothesis is that complex behavioral patterns can be decomposed into a finite alphabet of discrete, semantically indivisible tokens — quanta of information. Therefore, it is necessary to introduce an Ontology of actions, that is, an extensible vocabulary in which each observed behavior is represented by such a quantum. Classical examples include a fall, prolonged squatting, a raised-hands stance, and a shooter pose — all of them are interpreted as distinct IQ. A lightweight detector based on skeletal tracking (MediaPipe, MoveNet) recognizes these IQ on an edge device and transmits to the cloud not a video stream but only a sequence of symbols (for example, JSON). Owing to this, cloud components, including large language models (LLM), can process the sequence as text, ensuring millisecond-scale latencies at the edge and deterministic response characteristics (Gan et al., 2022; Cho et al., 2025) .

For Internet of Things video analytics tasks, a two-tier edge–cloud architecture should be used, comprising edge and cloud levels.

As for the edge level, it is characterized by a video stream captured by a camera or a low-power node located near the camera. Using OpenCV and MediaPipe, a marker-based pose estimation function extracts 33 body keypoints. A custom algorithm analyzes the angles between joints and their dynamics, matching observations with predefined IQ parameters from the ontology. When a condition is met (for example, a person remains in a bent posture for more than 10 seconds), the node forms a message and sends it to the cloud via MQTT or HTTP. The original frames remain on the device to ensure privacy.

The cloud level is a service that aggregates IQ event streams from multiple nodes. A vision-language model (for example, LLaVA) is used, which accepts a sequence of tokens plus the operator's text query, enabling complex zero-shot queries without fine-tuning. For example, the operator can set a rule of the form Notify if a person has left the facility (IQ5: Abandoned object), and then quickly left the area (IQ3: sudden acceleration). The LLM interprets the IQ sequence and generates an alert with a textual explanation.

A practical implementation based on the authors' developments demonstrated high efficiency. On a test bench with an Intel Core i5 processor (without a GPU), the edge node stably processed two 1080p streams at a rate of ~25 frames/s. The median latency from the moment of the event to receipt of the cloud notification did not exceed 3 seconds.

The ontology was empirically tested on six key IQ indicators, which are shown in Figure 1 for greater clarity.

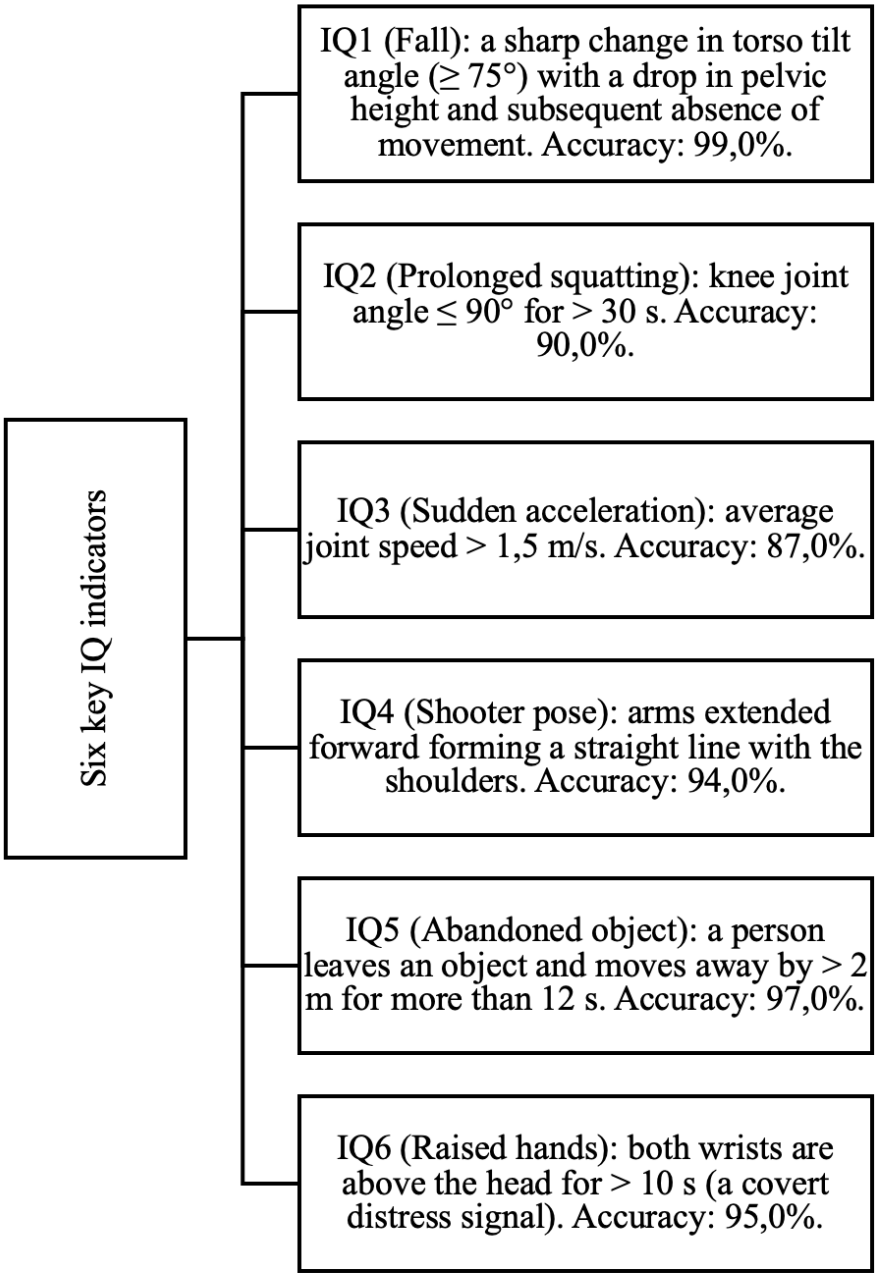

Fig.1. Six key IQ indicators

Experiments with LLaVA-1.6-13B processing the tokenized IQ stream showed logical accuracy of 92,3% on 15 complex natural-language queries. This exceeded the baseline variant with individual frames as input (86,7%), while the average inference latency decreased from 780 ms to 110 ms.

Discussion

The proposed architecture, grounded in the ontology of quanta of information, demonstrates substantial superiority over classical video analytics pipelines: it not only addresses latency and bandwidth bottlenecks, but also establishes a different logic of human–security-system interaction, shifting it from passive monitoring to a meaningful dialogue in natural language. At the same time, the obtained results delineate several avenues for development.

In the current implementation, a static, predefined ontology (IQ1–IQ6) is used. This scheme is effective but, by definition, constrained by the repertoire of already known patterns. Therefore, it is necessary to introduce the concept of an adaptive ontology with feedback, which removes this limitation. The improvement is achieved by incorporating an anomaly-detection module at the edge level. Such a module identifies behavioral trajectories that do not match any of the existing IQs yet statistically fall outside the norm;

instead of discarding them, the system vectorizes the corresponding motion sequences and transmits them to the cloud for expert interpretation. The operator assigns a semantic label to the detected behavior (for example, vandalism or fight), thereby forming a new quantum of information (say, IQ7). After verification, the definition is automatically propagated to all edge nodes of the network, enriching their ontology without the need for a full software update. The process is illustrated in Fig. 2.

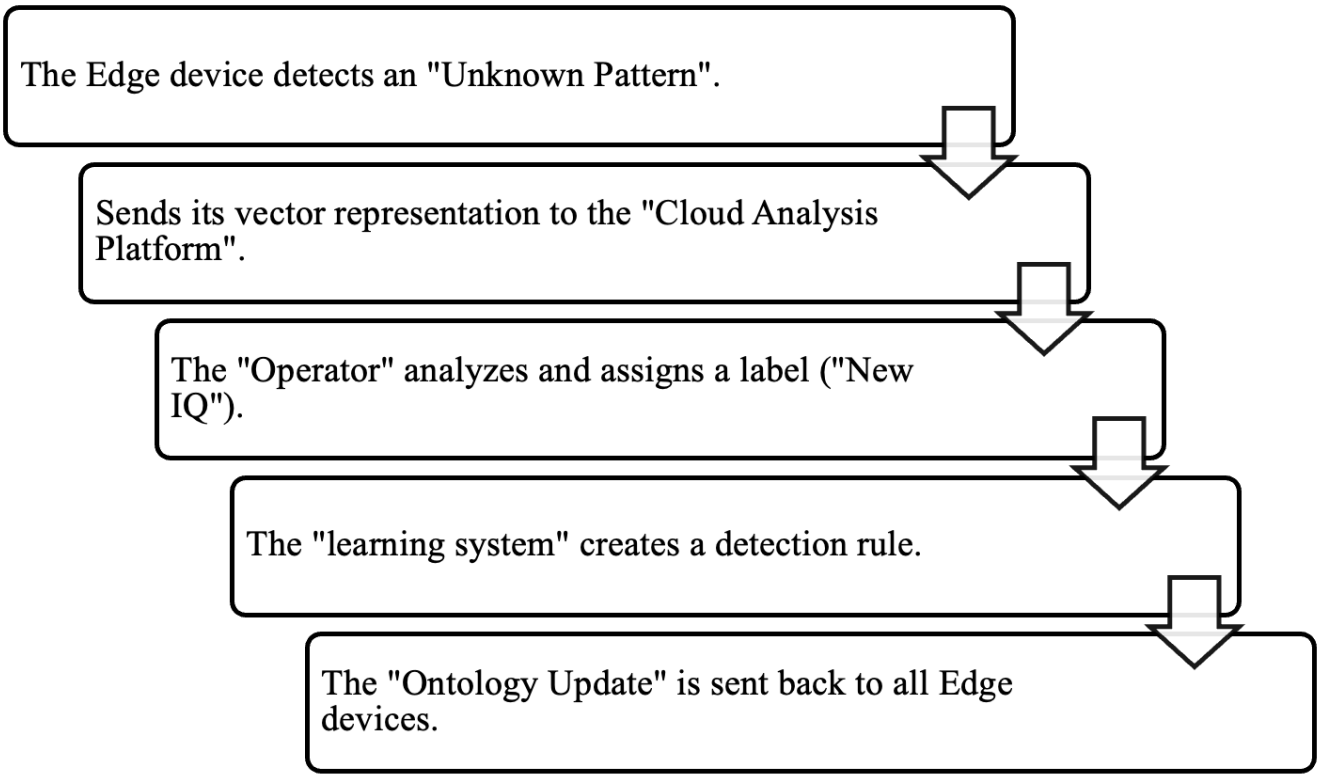

Fig. 2. Adaptive ontology cycle diagram (Abu Tami et al., 2024; Huang et al., 2023 ; Zha et al., 2025)

Such a scheme transforms the system from a static detector into a dynamic, self-learning ecosystem capable of adapting to evolving threats and the context of the specific object under observation. Owing to this adaptivity, not only operational effectiveness increases, but also the strategic resilience of the solution, which directly enhances its long-term value and efficiency. At the same time, the current model treats IQ as a linear, unweighted sequence. It is proposed to introduce a hierarchical organization of events, where quanta form levels of abstraction, and to further enrich interpretation with cross-source data, which will deepen causal analysis and improve the accuracy of inferences without altering the original semantics.Below in Table 1 a comparison of approaches to video analytics is demonstrated.

Table 1. Comparison of approaches to video analytics

Parameter	Traditional (Cloud-Only)	Edge (Edge-Only)	Authors approach (IQ-based Edge-Cloud)	Proposed evolution (Hierarchical)
Latency	High, unpredictable	Low	Low (<3 s)	Very low (for simple events)
Privacy	Low (video in the cloud)	High	Very high (tokens only)	Very high
Bandwidth	Very high	Minimal	Low	Low

Analysis complexity	High	Limited	High (Zero-shot)	Very high (contextual)
Adaptability	Low (requires retraining)	Low	Medium	High (self-learning)

As follows from the data of Table 1, the proposed method already demonstrates an advantage over classical solutions; however, transitioning to a hierarchical organization can qualitatively enhance its efficiency. At the lowest level reside base IQs. The middle level (Events) provides automatic composition of simple IQ sequences into more complex structural units: thus, the combination of IQ2 (Squat) and IQ5 (Abandoned item) gives rise to the Event Deposit of a suspicious object. The top level (Scenarios) interprets combinations of Events with the involvement of contextual features arriving from other sensory channels (audio, smoke detectors, access control systems), which makes it possible to identify complex behavioral patterns, including Unauthorized entry followed by preparation for a terrorist attack (Jebur et al., 2023; Silva et al., 2021).

The stream of IQ events forms a strictly ordered time series, optimal for training recurrent neural networks (RNN) and/or transformers. When trained on a representative archive of incidents, such a model identifies stable patterns in IQ sequences that statistically precede critical episodes. In particular, it can learn that a certain trajectory of actions (IQ3, IQ2, …) often anticipates an act of vandalism. This enables the system to generate not post factum notifications about an event that has already occurred, but proactive warnings of likely escalation, providing the operator with a critical time window for preventive action (Diraco et al., 2023; Hamid, 2023).

The proposed multimodal hierarchy shifts the task from simple detection to contextual interpretation of observed behavior, which fundamentally reduces the frequency of false alarms and increases the reliability of threat level assessment. Moreover, it is advisable to employ IQ sequences not only for the retrospective analysis of incidents but also in predictive mode. Table 2 demonstrates existing risks and methods for their mitigation in the models.

Table 2. Risks and methods of their mitigation in the proposed models (Diraco et al., 2023; Md Nur Hasan Mamun, 2024).

Risk	Mitigation method
Ontology poisoning (insertion of false IQ)	Introduction of a strict protocol for operator verification of new IQ; ontology version control.
Incorrect context interpretation (in a multimodal system)	Use of attention mechanisms to weight the importance of different modalities.
False predictive alerts	Setting a configurable confidence threshold for predictive alerts; use of explainable AI models (XAI) for interpreting predictions.

The proposed solutions — adaptive ontology, hierarchical event composition, and predictive analytics — represent a successive deepening of the baseline architecture. Their integration transforms a video surveillance system from a predominantly reactive loop into a proactive, context-informed, and self-learning security platform, thereby expanding the horizons of intelligent video analytics in mission-critical domains.

Conclusion

Within the conducted study, the central objective was achieved: an innovative edge–cloud architecture for video analytics was designed and theoretically substantiated, relying on an ontological model of quanta of information.

A comprehensive analysis of existing approaches confirmed their systemic limitations: overload of

network infrastructure, privacy vulnerabilities, and latency instability. The proposed architecture addresses these most vulnerable points by offloading the primary computational loop to the edge and transmitting only lightweight semantic tokens to the cloud.

The constructed ontology of actions and quanta of information (IQ) provided a rigorous formalization of complex behavioral patterns as machine-readable discrete entities. This created a coupling link between low-level computer vision procedures and high-level semantic inference implemented by language models.

The integrated system demonstrated compelling empirical characteristics: processing of two 1080p streams on a CPU-only device at a rate of approximately ~25 fps and a median event-to-notification latency of less than 3 seconds. These results support the hypothesis that tokenization of the video stream at the edge is a key mechanism for building highly efficient and scalable vision–language control systems.

In summary, the proposed approach not only surpasses existing solutions on technical metrics but also establishes a foundation for the next generation of intelligent security systems — proactive, adaptive, and capable of capturing complex contextual interdependencies. Further movement toward adaptive ontologies and predictive analytics, as shown in the discussion, appears to be a fruitful research direction.

References

1. Silva, J., Marques, E. R., Lopes, L. M., & Silva, F. (2021). Energy-aware adaptive offloading of soft real-time jobs in mobile edge clouds. Journal of Cloud Computing, 10(1), 38.
2. Gan, Z., Li, L., Li, C., Wang, L., Liu, Z., & Gao, J. (2022). Vision-language pre-training: Basics, recent advances, and future trends. Foundations and Trends® in Computer Graphics and Vision, 14(3–4), 163-352. http://dx.doi.org/10.1561/0600000105
3. Hamid, O. H. (2023). Data-centric and model-centric AI: Twin drivers of compact and robust industry 4.0 solutions. Applied Sciences, 13(5), 2753.
4. Abu Tami, M., Ashqar, H. I., Elhenawy, M., Glaser, S., & Rakotonirainy, A. (2024). Using Multimodal Large Language Models (MLLMs) for Automated Detection of Traffic Safety-Critical Events. Vehicles, 6(3), 1571-1590. https://doi.org/10.3390/vehicles6030074
5. Huang, A. Y., Chen, Y., Huang, D., & Zhao, M. (2023, December). Semantic Privacy-Preserving for Video Surveillance Services on the Edge. In Proceedings of the Eighth ACM/IEEE Symposium on Edge Computing (pp. 300-305). https://doi.org/10.1145/3583740.3626820
6. Diraco, G., Rescio, G., Caroppo, A., Manni, A., & Leone, A. (2023). Human Action Recognition in Smart Living Services and Applications: Context Awareness, Data Availability, Personalization, and Privacy. Sensors, 23(13), 6040. https://doi.org/10.3390/s23136040
7. Md Nur Hasan Mamun (2024). INTEGRATION OF ARTIFICIAL INTELLIGENCE AND DEVOPS IN SCALABLE AND AGILE PRODUCT DEVELOPMENT: A SYSTEMATIC LITERATURE REVIEW ON FRAMEWORKS. ASRC Procedia: Global Perspectives in Science and Scholarship, 4(1), 01–32. https://doi.org/10.63125/exyqj773
8. Zha, D., Bhat, Z. P., Lai, K. H., Yang, F., Jiang, Z., Zhong, S., & Hu, X. (2025). Data-centric artificial intelligence: A survey. ACM Computing Surveys, 57(5), 1-42. https://doi.org/10.1145/3711118
9. Cho, W., Kim, D., Lim, B., & Gu, J. (2025). PreEdgeDB: A Lightweight Platform for Energy Prediction on Low-Power Edge Devices. Electronics, 14(10), 1912. https://doi.org/10.3390/electronics14101912
10. Jebur, S. A., Hussein, K. A., Hoomod, H. K., Alzubaidi, L., & Santamaría, J. (2023). Review on Deep Learning Approaches for Anomaly Event Detection in Video Surveillance. Electronics, 12(1), 29. https://doi.org/10.3390/electronics12010029