New Perspectives in Science Education
International Conference
NEW PERSPECTIVES in SCIENCE EDUCATION

The Chronological Ascent to a Spatially Grounded World Model: Merging Geometric Architectures with Large Language Models to Meet Future Education Applied to Industry 6.0

Charlotte Sennersten[1], Kamilla Klonowska[2]

[1, 2] Computer Science Department, Kristianstad University, Sweden

Abstract

The development of a comprehensive computational World Model, capable of understanding and interacting with the three-dimensional physical world, embodies a truly global endeavour, now further propelled by the emergence of Industry 6.0. This new industrial paradigm, characterized by hyper-connectivity, intelligent systems, and sustainable innovation, unites research communities, industries, and educational institutions worldwide in the pursuit of transforming traditional science education. By bridging geometric data processing with advanced linguistic reasoning and leveraging technologies born out of Industry 6.0, the collaborative landscape of science education is rapidly evolving into a dynamic, interactive experience. This article is a conceptual, descriptive walk-through tracking the chronological milestones fuelling this transformation: from volumetric spatial indexing in 2016, to the rise of end-to-end deep learning for 3D feature extraction in 2017, followed by perceptual optimization with VoxelNeXt in 2022, and the parallel evolution of specialized knowledge representation through Galactica LLM. Industry 6.0 has also catalyzed international discourse around Digital Twins (DTs) in 2023/2024, establishing robust mathematical frameworks for modeling personalized, complex systems. The synthesis of these breakthroughs, realized in the 3D-LLMs (2023/2026), harnesses localization tokens to facilitate natural language querying and logical reasoning in (x, y, z)-space, signifying a pivotal advance for global science education and industrial practice. This highlights an educational paradigm which indirectly points towards a disconnect between general generative AI models 'appearing' to understand the 'reality' that they do not understand expressed in UNESCO's 'Guidance for Generative AI in Education and Research'. This calls for a slight urgency since generative AI are not yet informed by observations of the real world while in science real-world observations constitute scientific ground truth. A key aspect is to reach 3D structure comprehension including shape, orientation and location into practice into our current educations to meet this reality of ours. The worldwide shift from static, textbook-based instruction to spatially grounded World Models and Industry 6.0 innovations is creating an educational environment that nurtures curiosity, deep understanding, and scientific literacy across the globe, while preparing learners for an interconnected, intelligent future.

Keywords: *3D-LLM, Voxelization, Digital Twin, Spatial Grounding, World Model, Galactica, VoxelNeXt.*

1. Introduction

This is not the first time that education, particularly engineering education, has been required to closely align with industrial transformation to prepare students for rapid technological evolution. Previous industrial shifts, including Industry 4.0, demonstrated the necessity of integrating digitalization, automation, and intelligent systems into curricula. Today, the pace of change is even more accelerated. The extremely rapid growth of generative AI, 3D perception systems, and digital twin technologies, combined with intensified globalisation, demands a renewed collaboration between academia and industry.

This article shows that the emergence of spatially grounded large language models (3D-LLMs) marks the beginning of a new educational era shaped by Industry 6.0. As global technological ecosystems evolve at unprecedented speed, education must adapt proactively to prepare students not only to use advanced systems, but to critically understand, evaluate, and collaborate with them within increasingly complex, interconnected environments.

2. A Computational World Model

A World Model (WM) is a comprehensive computational framework designed to understand, reason about, and interact with the three-dimensional physical world. Unlike conventional artificial intelligence, which typically relies on episodic, partial observations, a genuine WM maintains long-term memories of entire environments, grasping intricate concepts such as spatial relationships, affordances, physics, and layout. A decade ago, the intersection of three-dimensional geometric modelling and natural language processing was an emerging research area, with very few publications exploring the integration of authentic volumetric data and linguistic analysis. At that time, the notion of a "WM" remained largely theoretical, limited by a predominantly "2D-centric" internet infrastructure that could process text and images but lacked the capacity to ground three-dimensional spatial datasets. In contrast, the current landscape has changed significantly; there is now a wealth of published work detailing the convergence of 3D perception, volumetric indexing, and textual data. This advancement signifies an important development in artificial intelligence: effective understanding of affordances, physical principles, and spatial arrangement requiring grounding in 3D environments, moving beyond reliance on 2D representations. A WM employing a comprehensive Earth-centric approach aims to unify diverse and traditionally independent methodologies, which are often maintained and referenced separately. Within the context of artificial intelligence, maintaining such separation is not a viable long-term strategy as pointed out in *'Guidance for Generative AI in Education and Research'* [1]. To comprehend three-dimensional concepts, it is useful to compare them with two-dimensional representations on a screen, where images are depicted using x- and y-coordinates with pixels. In a three-dimensional context, this extends to an x, y, and z coordinate framework, utilizing voxels rather than pixels as the fundamental units for representation.

2.1 VoxelNet -The Conceptualisation of Volumetric Indexing

In 2016, when VoxelNet was introduced as a WM implementing an earth indexation approach [2], as illustrated in figure 1, the field was at an early stage of development, with minimal research devoted to volumetric indexation and the integration of linguistic elements. A CMTE Development Ltd. and CSIRO patent was filed [3]. Today, the 'VoxelNet' search-term has exploded to approximately 7,390 references (Google Scholar, accessed 13-02-26), reflecting a shift from 2D-centric web models to spatially grounded intelligence. The development of these models follows a rigorous chronological path to manage increasing computational needs.

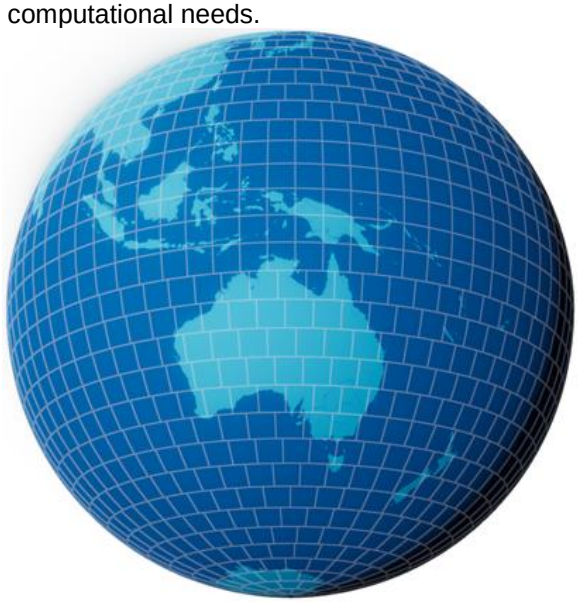


Fig. 1. A World Model: VoxelNet Earth Indexation Infrastructure by CSIRO Australia/Sennersten, source [4].

The VoxelNet Platform, developed by Sennersten et al. in Australia [2], [3], represented a notable advancement in volumetric spatial data management. Departing from traditional two-dimensional presentation models prevalent during the early web era, VoxelNet introduced Earth indexation alongside an octree indexing structure, thereby enabling hierarchical organization and efficient nesting of three-dimensional data. At that time, a major challenge was formulating an efficient computational representation for Earth on a large scale, particularly at the resolution of unit-sized voxels. To solve this issue, Sennersten proposed utilising IPv6 addresses as unique identifiers for each voxel, thereby

treating every cubic metre of space as an integrated component within the Internet of Things (IoT) to support computational needs. This methodology established the foundation for the VoxelNet framework, wherein specialised agents oversee discrete volumes of information, ultimately facilitating virtual presence and enabling Digital Twin (DT) [5] applications for physical environments. The VoxelNet indexation infrastructure established a row of furtherments. The CSIRO VoxelNET research team advanced UAV and Softbot approaches [6] and applied and incorporated Machine Learning and analytics to fuse labelled and unlabelled data including point clouds [7].

2.2. VoxelNet -The Birth of End-to-End Learning

By 2017, the focus shifted from data management to perceptual learning. Apple researchers in USA introduced a deep learning version of VoxelNet, the first end-to-end trainable network to unify feature extraction and bounding box prediction from raw LiDAR point clouds [8]. The key innovation was the Voxel Feature Encoding (VFE) layer. This allowed the model to transform disordered points within a 3D grid (voxels) into descriptive volumetric representations. By removing the need for manual feature engineering, such as bird's-eye-view projections, VoxelNet proved that a neural network could learn discriminative representations of objects with varying geometries (e.g., pedestrians and cyclists) directly from 3D data. This established the "Voxel" as the standard unit for 3D deep learning, much like the pixel is for 2D.

To advance the understanding of voxels within the VFE layer (see figure 2), several key components must be established. A voxel is defined as a volumetric, three-dimensional pixel; utilizing this concept involves the following steps:

The three-dimensional space is partitioned into a uniform grid.
Each cell within this grid is designated as a voxel.
All data-points located within the same voxel are grouped together.

For effective use of the VFE layer, it is necessary to:

Group points according to their respective voxels.
Encode features at the point level within each voxel.
Aggregate these features into a fixed-length feature vector for each voxel.

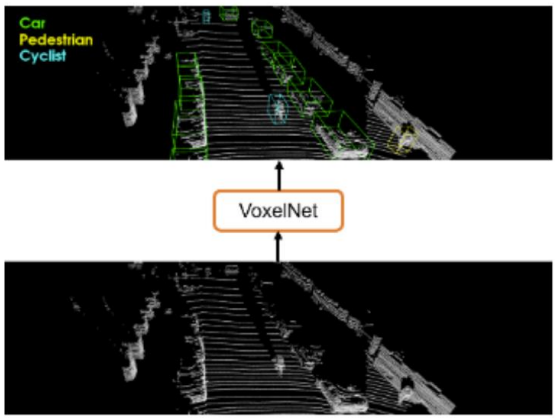


Fig. 2. "VoxelNet directly operates on the raw point cloud and produces the 3D detection results using a single end-to-end learning network", source [8].

3. Integrating 3D Spatial Understanding and Natural Language: Insights from ScanQA and ScanRefer

To illustrate the integration of 3D spatial knowledge with natural language processing, two significant research initiatives, ScanQA and ScanRefer, serve as prominent examples. Both projects, conducted in 2020, advanced the field of 3D vision-and-language understanding. ScanQA focused on 3D question answering within spatial contexts, while ScanRefer addressed 3D object localization in RGB-D scans using natural language descriptions. These endeavors pursued distinct objectives in examining the relationship between language and 3D environments and demonstrated that 3D large language models (3D-LLMs) significantly outperformed existing baseline approaches in both 3D question answering and spatial grounding tasks. Furthermore, these models exhibited the capability for complex task decomposition, including providing comprehensive instructions for identifying

ingredients within specific 3D spaces and facilitating navigation to target objects through conversation-based waypoints.

3.1 ScanQA

The objective of this project carried out in Japan was to accurately identify and localize the referenced object by predicting its 3D bounding box within a given scan [9]. For instance, when presented with a point cloud of a room and a description such as *"the red chair next to the window"*, the model generated the corresponding 3D bounding box for that specific chair (see figure 3). To effectively integrate language with 3D object data, both 'language embedding' and '3D geometry fusion' techniques were employed, enabling the model to interpret the provided description as well as the spatial context of the scene. ScanQA data is available at GitHub [10].

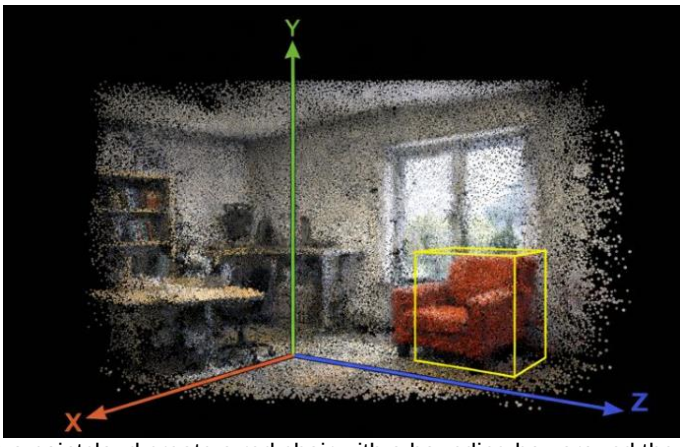


Fig. 3. Prompt "From a pointcloud create a red chair with a bounding box around the chair standing by the window. show x, y, and z for the room and the bounding box.", ChatGPT 5.2 (accessed 180126).

3.2 ScanRefer

The objective of the German Canadian project was to move beyond localization and address inquiries related to 3D scenes. Each input consisted of a complete 3D scan, obtained from an RGB-D setup such as ScanNet, accompanied by a question formulated in natural language [11]. To facilitate a clearer understanding, the following table (table 1) illustrates the functioning of both approaches. ScanRefer data is available at GitHub [12].

Table 1. Explanation and comparison of ScanQA and ScanRefer.

Feature	ScanQA	ScanRefer
Primary Task	Answer questions about a 3D scene	Localize a referred object in 3D
Input	3D scan + natural language question	3D scan + natural language reference
Output	Answer text + (often) object bounding boxes	3D bounding box of referenced object
Focus	Scene reasoning and understanding	Object grounding via language
Dataset Size	~41 k QA pairs	~51 k object descriptions
Complexity	Reasoning about relations + answering questions	Finding one referenced object

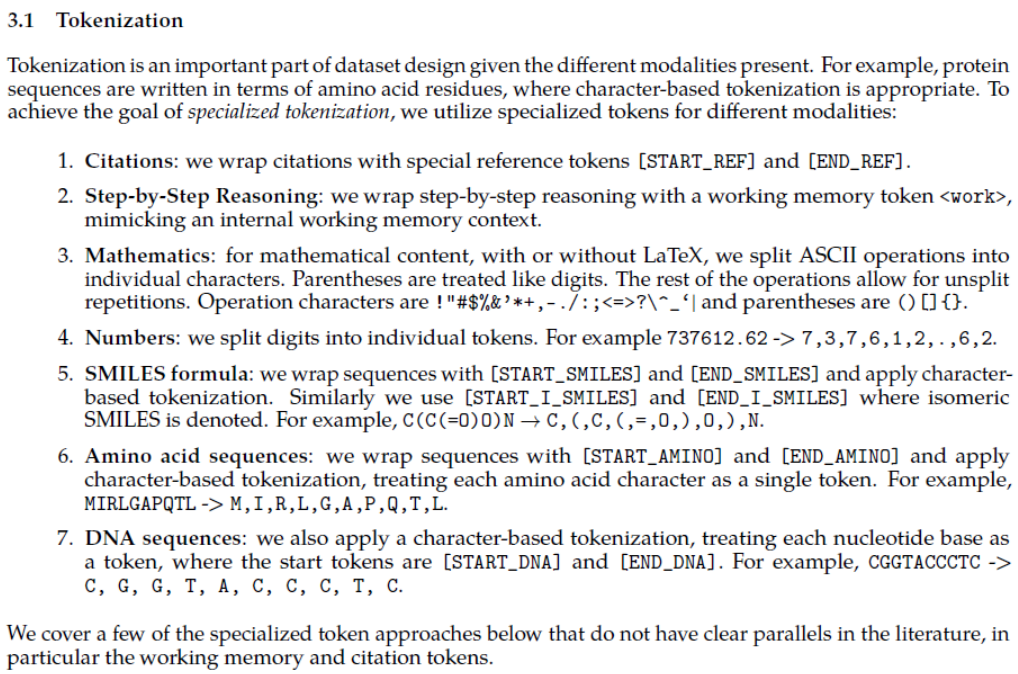
In summary, the two methodologies are outlined as follows:
ScanQA: *"Respond to a question pertaining to the entirety of this 3D scene."*
ScanRefer: *"Identify the object referenced within this 3D scan."*

4. The Rapid Rise of Large Language Models and Specialised Scientific AI

In 2022, two parallel tracks of innovation emerged: the optimization of 3D perception (mentioned briefly above) and the specialized tokenization of scientific knowledge. With the focus on natural language releases of OpenAI's Large Language Model (LLM) ChatGPT in November 2022, Meta's Galactica research language model released November 2022 and the year after Meta's Large Language Model Meta AI (LlamA) in February 2023. Half a year later, xAI's Grok Model was released in November 2023 followed by Google Deep Mind's AI Model Gemini in December 2023. The commercial LLMs really started to take off. Who would have imagined such a breakthrough or rather explosion of linguistic language models 10 years before.

4.1 Galactica: Bridging Science and Language

Computing has indeed revolutionized how research is conducted, but information overload remains an overwhelming problem. A huge increase in published articles can be seen over a 20-year period at arXiv (arXiv, 2022) [13]. Beyond papers, scientific data is also growing much more quickly than our ability to process it. Given the volume of information, it is impossible for a single person to read all the papers in a given field; and it is likewise challenging to organize data on the underlying scientific phenomena. Meta AI in USA introduced Galactica (GAL), a large language model that can store, combine and reason about scientific knowledge [14]. Galactica pioneered the concept of "tokenizing nature" (see figure 4) using specialized tokens to represent non-linguistic data like DNA sequences, protein chains, and chemical formulas (SMILES). Crucially, Galactica introduced the Working Memory Token (<work>), which wrapped step-by-step reasoning paths. This mimicked an internal "scratchpad," allowing the LLM to perform complex scientific reasoning that a single forward pass could not achieve. Galactica's success in predicting citations and chemical properties showed that LLMs could become context-associative interfaces for the physical and scientific worlds.



3.1 Tokenization

Tokenization is an important part of dataset design given the different modalities present. For example, protein sequences are written in terms of amino acid residues, where character-based tokenization is appropriate. To achieve the goal of *specialized tokenization*, we utilize specialized tokens for different modalities:

1. **Citations**: we wrap citations with special reference tokens `[START_REF]` and `[END_REF]`.
2. **Step-by-Step Reasoning**: we wrap step-by-step reasoning with a working memory token `<work>`, mimicking an internal working memory context.
3. **Mathematics**: for mathematical content, with or without LaTeX, we split ASCII operations into individual characters. Parentheses are treated like digits. The rest of the operations allow for unsplit repetitions. Operation characters are `!"#$%&'*+,-./:;<=>?\^_‘|` and parentheses are `()[]{}`.
4. **Numbers**: we split digits into individual tokens. For example `737612.62 -> 7,3,7,6,1,2,.,6,2`.
5. **SMILES formula**: we wrap sequences with `[START_SMILES]` and `[END_SMILES]` and apply character-based tokenization. Similarly we use `[START_I_SMILES]` and `[END_I_SMILES]` where isomeric SMILES is denoted. For example, `C(C(=O)O)N` → `C,(,C,(,=,O,),O,),N`.
6. **Amino acid sequences**: we wrap sequences with `[START_AMINO]` and `[END_AMINO]` and apply character-based tokenization, treating each amino acid character as a single token. For example, `MIRLGAPQTL -> M,I,R,L,G,A,P,Q,T,L`.
7. **DNA sequences**: we also apply a character-based tokenization, treating each nucleotide base as a token, where the start tokens are `[START_DNA]` and `[END_DNA]`. For example, `CGGTACCCTC -> C, G, G, T, A, C, C, C, T, C`.

We cover a few of the specialized token approaches below that do not have clear parallels in the literature, in particular the working memory and citation tokens.

Fig. 4. Excerpt from article [14] describing the team's tokenisation structure.

Galactica was developed using an extensive and carefully curated collection of scientific resources, encompassing 48 million papers, textbooks, lecture notes, millions of compounds and proteins, as well as scientific websites, encyclopedias, and additional materials.

5. VoxelNeXt: Enhancing 3D Object Detection through Sparse Computational Efficiency

Observing that dense prediction heads are often computationally inefficient, given that typically less than 1% of the 3D space contains relevant objects, researchers from Hong Kong introduced VoxelNeXt [15] in 2023. Instead of using dense grids of data or manually designed reference points known as "anchors", this network used a sparse approach meaning it only processed the parts of the

3D space where actual data points (voxels contained points) exist. By doing so, it could efficiently identify and predict the presence of 3D objects directly from the information contained within these voxels, making the process both faster and more resource efficient. VoxelNeXt showed that a simpler approach such as adding extra down-sampling steps to increase the area each voxel could 'see' could then perform better than complicated, multi-stage detection systems.

But how does VoxelNeXt differ from VoxelNet? In VoxelNet, the process involves heavy 3D convolution operations and dense grids full of voxels, many of which are empty, and the Voxel Feature Encoding (VFE) layers are computationally demanding. VoxelNeXt changes this by redesigning how voxels are processed from the ground up (see table 2). Instead of wasting resources on empty space, VoxelNeXt focuses on computation only where there are actual data points, maintaining a much simpler and more efficient architecture.

Table 2. Comparison between VoxelNet [8] and VoxelNeXt [15].

	Aspect	VoxelNet	VoxelNeXt
Voxel Representation	Voxel grid	Dense 3D grid	Sparse voxelset
	Empty voxels	Explicitly processed	Ignored
	Efficiency	Low	High
Feature Encoding	Encoding	VFE layers (PointNet-style)	Simple voxel aggregation
	Point-level MLPs	Yes	Minimal or removed
	Complexity	High	Much lower
Backbone Architecture	Backbone	Dense 3D CNN	Sparse convolution backbone
	Computation	3D everywhere	Only on occupied voxels
	Memory Use	Very high	Much lower
Detection Head	Detection style	Anchor-based RPN	Anchor-free, centered-base
	Output	Region proposals	Object centers + regression
	Simplicity	Complex	Simpler & Faster

6. Digital Twins and Mathematical Rigor

Advancements in 3D modelling and LLMs have increased focus on DTs, which virtually replicate physical assets in real time. In digital twins, geolocation data, longitude and latitude, obtained for example using GPS for location. Mapping and surveying often use grid coordinates representing a flat plane: northings indicate distance along the y-axis while eastings show distance along the x-axis from a set origin. The three-dimensional nature of Earth is characterized by its curvature, distinguishing it from a flat, two-dimensional surface.

The 2023 MATH-DT workshop in the USA [16] underscored the need for foundational mathematical advances to transition from generic physical laws to personalized applications. Researchers emphasized that DTs rely on solving inverse problems, such as using sensor data to determine real-world properties. This has brought attention to what is called Uncertainty Quantification (UQ), which is a foundational mathematical discipline required for the development and operation of DTs, ensuring that virtual models accurately reflect reality and support reliable decision-making. It involves systematically identifying and managing unknowns at every stage of the industrial pipeline, including initial modeling, forward problems, optimization, and engineering design. The sources of this uncertainty are diverse, ranging from physical parameters, such as material variations or unknown physics, to external factors like environmental loads, boundary conditions, and inaccuracies in sensor data. The DTs serve as a central system, leveraging 3D data to support decision-making. Mathematics is essential across sciences, including AI and machine learning (AI/ML), with computational engineers increasingly applying mathematical concepts like optimization, statistics, and numerical analysis. New mathematical developments are crucial for advancing AI/ML, proving

convergence of algorithms, and addressing limitations in current data-driven methods that often struggle with basic physics models. Key challenges remain unresolved, such as optimal network design, training data requirements, and accuracy expectations. The MATH-DT report [17] found that there is no universal solution; for instance, while forecasting tools such as Kalman filters are effective in engineering, they prove less successful in biomedical applications. Likewise, AI and machine learning techniques demonstrate variable outcomes across different disciplines, indicating a need for more systematic research efforts. In summary, research in digital twins presents substantial mathematical challenges and opportunities related to the development of Euclidean space and trigonometric concepts for DTs.

7. The Synthesis of 3D-LLMs

In 2023 the culmination of these pillars is the 3D-LLM, an architecture designed to "inject" the 3D world into large language models [18]. By synthesizing 3D point clouds, voxel-based perception (VoxelNet [8]/ VoxelNeXt [15]), and specialized reasoning (Galactica [14]), 3D-LLMs perform tasks far beyond the scope of 2D Vision-Language Models. The core architectural pillars of 3D-LLMs are:

1) 3D Localization Mechanism: To bridge the gap between language and space, researchers introduced localization tokens (e.g., $\langle x_{min}, y_{min}, z_{min}, x_{max}, y_{max}, z_{max} \rangle$) into the LLM vocabulary. This allows the model to "speak" in coordinates, grounding descriptions in space.

2) Cross-Modal Alignment: Using 3D feature extractors, 3D representations are mapped into the same feature space as 2D pretrained features (like Contrastive Language-Image Pretraining (CLIP)), allowing standard 2D VLMs to be used as efficient backbones for 3D reasoning. CLIP is a multimodal AI model developed by OpenAI that learns the associations between images and text.

3) Holistic Scene Understanding: Unlike episodic partial-view observations, 3D-LLMs store long-term memories of entire scenes in holistic 3D representations.

Hong et al. [18] in an American Chinese collaboration were fusing LLMs and Vision-Language Models (VLMs) to show and prove how these two disparate models can excel multiple tasks, such as commonsense reasoning. The authors highlighted that these models could be very powerful, but they were not grounded in the 3D physical world, which involves richer concepts such as spatial relationships, affordances, physics, layout, and so forth. In their work, they proposed to inject the 3D world into large language models and introduce a whole new family of 3D-LLMs. They demonstrated how 3D-LLMs could take 3D point clouds and their features as input and perform a diverse set of 3D-related tasks, including answering 3D questions, task decomposition, 3D-assisted dialog and navigation. Kanazawa of UC Berkeley [19] notes that LLMs can act as interfaces, providing common sense communication, while detailed world models enable spatial-temporal memory. These advancements are likely to reshape knowledge sharing in the next decade, bridging the gap between language and real-time spatial-temporal understanding.

8. Discussion and Conclusion

The importance of 3D-LLMs in education and Industry 6.0 can be understood most clearly through the chronological trajectory presented in this article. Each technological milestone reflects a structural step toward spatially grounded intelligence, and together they form a logical progression with direct educational implications.

The introduction of volumetric indexing through VoxelNet (2016) marked a conceptual departure from a 2D-centric web toward three-dimensional spatial representation. By proposing Earth-indexed voxel structures and hierarchical octree systems, early research established the idea that physical space could be computationally structured. This was not yet educationally transformative, but it laid the infrastructural foundation for representing reality in machine-readable 3D form. The next milestone, end-to-end voxel-based learning (2017), shifted the focus from indexing to perception. VoxelNet demonstrated that neural networks could learn directly from raw point clouds without manual feature engineering. This transition signaled that 3D space was no longer merely stored; it could be interpreted. In educational terms, this corresponds to moving from static diagrams toward interactive spatial data analysis, enabling learners to explore geometric reasoning dynamically. The subsequent

integration of vision and language (ScanRefer, ScanQA) expanded capabilities from perception to linguistic grounding. These systems connected spatial datasets to natural language queries, bridging geometry and communication. This development is particularly relevant for education, as it transforms complex spatial modeling into accessible conversational interaction while preserving coordinate-based structure. Parallel to these advances, specialized scientific LLMs such as Galactica (2022) introduced structured tokenization for scientific data, including domain-specific symbolic representations. This demonstrated that large language models could serve as interfaces to structured scientific knowledge rather than merely generating fluent text. Importantly, it revealed both the potential and limitations of generative AI when disconnected from physical grounding. The introduction of VoxelNeXt (2023) addressed computational efficiency, emphasizing sparse spatial reasoning aligned with real-world constraints. This refinement reflects Industry 6.0 priorities: efficient, real-time processing of cyber-physical systems and digital twins.
Finally, the synthesis into 3D-LLMs (2023–2026) integrates these threads: volumetric indexing, end-to-end 3D perception, language grounding, and scientific tokenization, into models capable of coordinate-aware reasoning. This culmination aligns structurally with Industry 6.0, which relies on digital twins, robotics, intelligent infrastructure, and spatially integrated systems.

Historical progression shows that spatial grounding has moved from conceptual possibility to operational reality. Therefore, integrating 3D-LLMs into education is not merely technological adoption, it is alignment with the evolving cognitive infrastructure of industry and society. However, this chronological argument also reinforces the need for critical integration. At every stage, from voxel indexing to generative reasoning, computational models' approximate reality rather than embody it. Education must therefore emphasize verification, empirical grounding, and ethical accountability alongside technical competence. In doing so, 3D-LLMs can become structured learning tools that support, rather than substitute, scientific reasoning within the emerging paradigm of Industry 6.0.

9. Summary and Future Work

Constructing a chronological overview of global developments highlights the swift evolution of scientific tools and methodologies in response to the demand for contextual knowledge identified by the UNESCO guidance. As a result, there is an imperative for scientific education and industry to adapt by integrating these emerging tools and methods in order to effectively assess risks and opportunities. The conceptual analysis presented in this article will serve as the foundation for its authors to further incorporate these insights into existing Computer Science AI curricula.

In summary, the rapid advancement of scientific and technological innovation in digital twins and 3D large language models represents a significant transition toward integrated and context-aware systems that will challenge and enhance staff and student understanding.

Acknowledgement
The authors express their gratitude to the reviewers for their insightful feedback, which contributed significantly to enhancing the quality of the article.

REFERENCES

[1] Miao., F., and Holmes., W., *"Guidance for Generative AI in Education and Research",* UNESCO, 2023, https://doi.org/10.54675/EWZM9535
[2] Sennersten, C., Davie, A., and Lindley, C., "VoxelNet - An Agent Based System for Spatial Data Analytics", COGNITIVE 2016, The Eight International Conference on Advanced Cognitive Technologies and Applications, Rome, Italy, 2016. download_full-libre.pdf
[3] Sennersten, C., Lindley, C., Evens, B., Grace, A., and Wise, J., "Spatial Data Processing System and Method", Commonwealth Scientific and Industrial Research Organisation (CSIRO), 2017-2024, US20200213426A1, https://patents.google.com/patent/CA3075119A1/en
[4] VoxelNET 4D data integration platform, CSIRO, https://www.csiro.au/en/work-with-us/industries/mining-resources/Mining/VoxelNET, accessed 130226.
[5] Semeraro, C., Lezoche, M., Panetto, H., Dassisti, M., "Digital Twin Paradigm: A Systematic Literature Review", Computers in Industry, Volume 130, 2021, https://doi.org/10.1016/j.compind.2021.103469

[6] Sennersten, C., Evans, B., and Lindley, C., “VoxelNET’s Geo-Located Spatio Temporal Softbots”, COGNITIVE 2019, The Eleventh International Conference on Advanced Technologies and Applications, Venice, Italy, 2019, https://personales.upv.es/thinkmind/dl/conferences/cognitive/cognitive_2019/cognitive_2019_2_20_40040.pdf
[7] Azhari, F., Sennersten, C., Milford, M., and Peynot, T., ”PointCrack3D: Crack Detection in Unstructured Environments using a 3D-Point-Cloud-Based Deep Neural Network”, arXiv, 2021, https://doi.org/10.48550/arXiv.2111.11615
[8] Zhou, Y., and Tuzel, O., “VoxelNet: End-to-End Learning for Point Cloud Based 3D Object Detection”, 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 4490-4499, 2017/2018, doi:10.1109/CVPR.2018.00472
[9] Azuma, Y., Miyazawa, K., Ito, Y., Hayashi, Y., Oda, Y., & Uhno, K., “ScanQA: 3D Question Answering for Spatial Scene Understanding”, In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, (pp. 19129–19139), CVPR 2022 Open Access Repository
[10] ScanQA, Github access, https://github.com/ATR-DBI/ScanQA, accessed 130226
[11] Chen, D. Z., Chang, A. X., & Nießner, M., “ScanRefer: 3D Object Localization in RGB-D Scans using Natural Language”, In European Conference on Computer Vision (ECCV), 2020, pp. 202-221, Cham: Springer International Publishing, https://doi.org/10.48550/arXiv.1912.08830
[12] ScanRefer, Github access, https://daveredrum.github.io/ScanRefer/, accessed 130226
[13] Cornell University, Submission Rate Statistics - Data for 1991 through 2021, updated 3 January 2022, ArXiv, https://info.arxiv.org/help/stats/2021_by_area/index.html, accessed 130226
[14] Taylor, R., Kardas, M., Cucurull, G., Scialom, T., Hartshorn, A., Saravia, E., Poulton, A., Kerkez, V., and Stojnic, R., “Galactica: A Large Language Model for Science”, 2022, https://doi.org/10.48550/arXiv.2211.09085
[15] Chen, Y., Liu, J., Zhang, X., Qi, X., & Jia, J., ”VoxelNeXt: Fully Sparse VoxelNet for 3D Object Detection and Tracking”, 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 21674-21683, DOI:10.1109/CVPR52729.2023.02076
[16] Mathematical Opportunities in Digital Twins (MATH-DT) Workshop, https://dcn.nat.fau.eu/math-dt-workshop-2023/, accessed 130226
[17] Antil, H., “Mathematical Opportunities in Digital Twins (MATH-DT)”, ArXiv abs/2402.10326 (2024), https://arxiv.org/pdf/ 2402.10326.pdf
[18] Hong, Y., Zhen, H., Chen, P., Zheng, S., Du, Y., Chen, Z., & Gan, C., ”3D-LLM: Injecting the 3D World into Large Language Models”, Advances in Neural Information Processing Systems, 36, 2023, 20482-20494, https://doi.org/10.48550/arXiv.2307.12981
[19] Bechard, D.E., “World Models Could Unlock the Next Revolution in Artificial Intelligence”, Scientific American, 2025, https://www.scientificamerican.com/article/world-models-could-unlock-the-next-revolution-in-artificial-intelligence/, accessed 170126