Queryable Spatio-Temporal 4D Representation for Construction Progress Monitoring

Sustainability Challenge and Background: The construction sector remains one of the most resource-intensive and wasteful global industries, frequently overestimating material, labor, and time needs [1]. Despite digitization efforts, on-site monitoring in construction largely relies on anecdotal estimates by construction managers. This leads to inconsistent documentation of key metrics such as progress, structural changes, installed assets, construction errors causing rework, and accidents. These gaps contribute to concerning statistics: 90% of infrastructure projects are delayed or over budget [2], and 52% of cost increases are due to rework from poor quality control [3]. In addition, 95% of non-hazardous construction and demolition waste is recyclable, yet it frequently ends up in landfills due to inadequate information and tracking of building materials [4]. As the world anticipates the significant need to house an estimated three billion people by 2030–equivalent to roughly 40% of the global population [5] –and confronting the urgent realities of climate change- there is a critical need for better management and monitoring practices to foster efficiency, safety, and environmental performance in construction.

The construction industry aims to create digital twins– dynamic, data-driven representations of real-world objects, processes, or systems that evolve- from design to end-of-life [6]. Current techniques employ reality capture methods at various project stages to generate a set of temporally sparse 3D measurements (4D). While this data enables off-site progress tracking, project managers must sift through extensive data to identify issues visually. With high costs, data limitations, and the industry's low error tolerance, automation efforts thus far have fallen short [7], highlighting the research opportunity for an automated and reliable construction progress tracking system.

Recent vision-linguistic models (VLMs) [8, 9] have made breakthroughs due to their zero-shot capability – generalizing to unseen tasks. Visual question answering (VQA) [10] combines computer vision and natural language using large language model (LLM) backbones, enabling 3D spatial [11–13] and video-based temporal reasoning [14, 15] and offering user-friendly natural language queries. VLM-based VQA, with its powerful reasoning for complex queries and causal understanding, is an ideal candidate for extending into 4D spatio-temporal reasoning. However, 3D VQA methods are already limited by scarce 3D data and narrow question diversity—shortcomings that would worsen in 4D—yielding weaker generalization compared to 2D or video domains. They also suffer from hallucinations (over-reliance on priors) and struggle to provide visually grounded outputs. Meanwhile, 3D scene understanding [16, 17] and open-vocabulary 3D scene understanding–which builds on foundation models trained extensively in 2D, transferring semantic features to point clouds [18]– inherently supports grounded segmentation (semantic, instance, and panoptic). Existing 4D research [19, 20] has not included data like the intermittent captures common in construction settings with significant scene changes. While scene understanding also shows promise, it relies on feature similarity, limiting the complexity of user queries.

There is a pronounced mismatch between the construction industry's need for automated 4D progress monitoring and the state of computer vision research. As discussed, VLMs and scene understanding show promise in 3D contexts but have not been adapted for 4D issues involving significant structural changes over time with sparse data capture intervals. Construction environments also pose unique challenges—low-texture, repetitive geometry, and specialized semantics—not found in standard vision benchmarks. Lastly, reliability is a critical barrier: without high accuracy and reliable outputs, trust for industrial adoption is hard to establish.

Proposed Solution: In Figure 1, we illustrate our proposed multi-agent system for creating a user-queryable 4D spatio-temporal scene representation. Our approach starts with a user inputting sparse 3D measurements (e.g., LiDAR or photogrammetry) captured at pivotal moments during construction. First, a novel spatiotemporal 4D visual-language model (VLM) extracts open-vocabulary visual and temporal features to interpret the scene. An off-the-shelf large language model (LLM) then parses a user's complex query and, through an API, iteratively interacts with the 4D representation to collect the necessary low-level details. Finally, the system gives users a synthesized, text-based conclusion paired with grounded indexes or visual outputs from the 4D data. By revisiting the same scene at different moments (4D), we can uncover relationships among snapshots, such as object displacements, occlusions, trends, and event identification (e.g., rework), providing feedback previously unavailable. Users can interact through natural language queries to identify materials, localize rework, spot safety issues, and monitor progress by comparing measured realizations with planned BIM across time points.

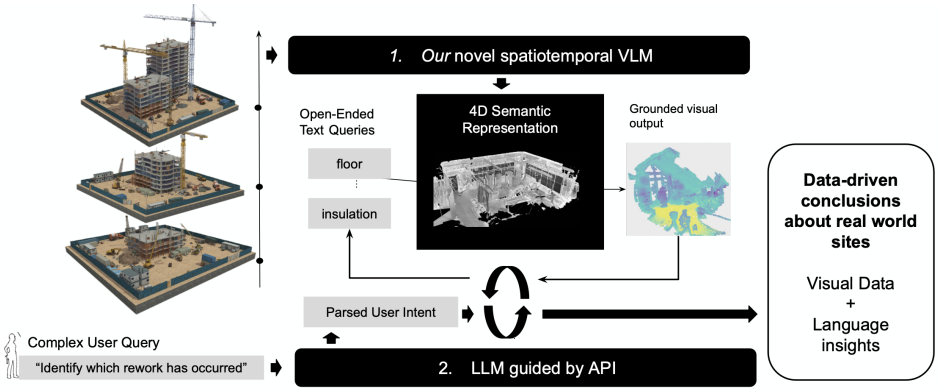
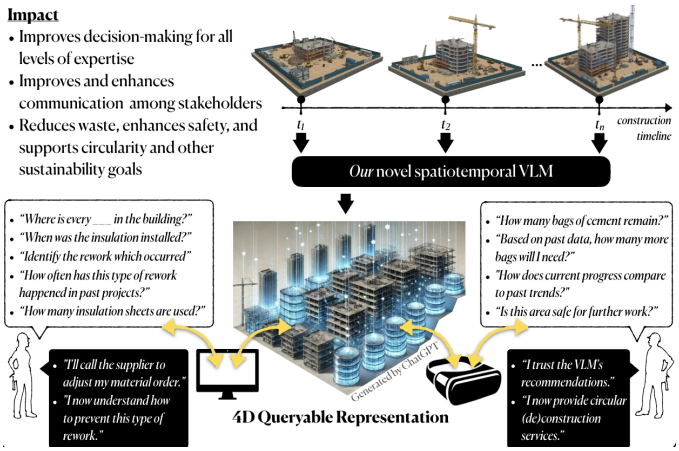
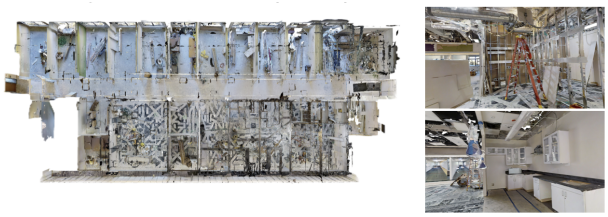


Figure 1: Overall Method

Potential Impacts: Rework often arises from poor planning, nonconforming work, misalignments with design changes, and quality control issues such as construction errors and material defects [21]. Implementing a unified tool for progress monitoring and analysis will streamline digital documentation and allow for early detection of discrepancies by integrating with existing tools like building information modeling (BIM), facilitating proactive decision-making. Additionally, automated analysis across sites will reduce friction in communication, enhance collaboration, and standardize the evaluation of trends and effectiveness, closing the feedback loop throughout the construction lifecycle for more efficient and sustainable projects. Through user case studies, the project will also assess the tool's feasibility, integration into current processes, and sustainability benefits. Figure 2a illustrates user interactions and impacts, creating a feedback loop for rapid learning and automation while minimizing issues related to operability, buildability, and waste.



(a) Impact and Applications

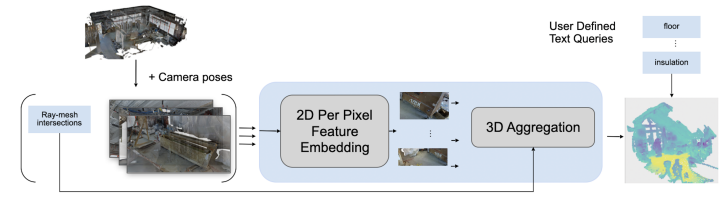


(b) Nothing Stands Still Dataset Example

(c) 3RScan Dataset Example

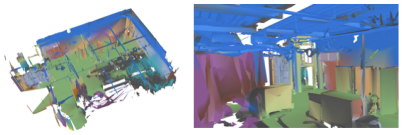
Figure 2: Left: Impact and applications. Right: Dataset Examples

Current Status: Development has begun with two existing datasets to develop and validate our approach, shown in Figure 2b and 2c. The NSS dataset [22] documents structural changes in buildings during construction but lacks semantic details, while the 3RScan dataset [23] captures indoor environments with semantic annotations during a building's usage phase. Research to date has (i) validated the performance of 2D foundation models with construction-specific concepts, integral to the overall success of our proposed design (Figure 3b) and (ii) developed a 3D open vocabulary pipeline that lifts 2D foundation models to 3D (without time) to examine the limitations of geometric non-trained feature lifting (Figure 3a 3c). We are currently (iii) designing and training a novel transformer and CNN-based architecture to merge temporal data for 4D closed-vocabulary instance segmentation (Figure 3d). Results thus far have shown improved ability to track instances of objects over time compared to the baseline.

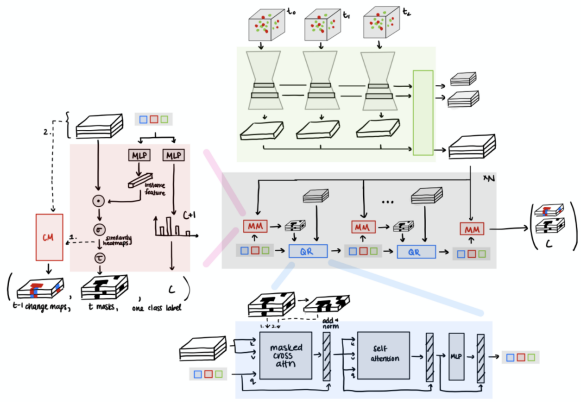


(a) Base Open-Vocabulary 3D Pipeline

(b) Example Output from a 2D VLM



(c) Resulting Semantic Segmentation for an NSS Scene



(d) 4D Scene Understanding Architecture

Figure 3: Current Status of the Project

Research Plan: The research timeline in Table 1 is dived into three phases and an ongoing data effort.

Date	Stage	Metric	Justification
Feb 2026	Phase 1	CV metrics (mIoU, mAP)	Ensure quantifiable accuracy
April 2026	Phase 2	User studies and task completion rate	Ensure user ease and complex task success
July 2026	Data	CV metrics (mIoU, mAP)	Ensure effectiveness for domain specific data
July 2027	Phase 3	Lifecycle carbon footprint, rework reduction	Quantify sustainability impact

Table 1: Timeline

Phase 1 will extend our prior work by uniting our previous methods (ii) and (iii) to enable open-vocabulary 4D instance segmentation. To leverage open-vocabulary embeddings we will replace the feature backbone in (iii) and build on initial progress with (i) and (ii) (as shown in Figure 3) by lifting 2D CLIP features [8] into 3D—using techniques such as Gaussian Splats [24]. The existing architecture will incorporate temporal information to label instances across multiple scans consistently.

Phase 2 will integrate a large language model (LLM), such as GPT-4o [9], which is trained on internet-scale data and excels at parsing user intent and generating sophisticated language queries. We will interface the LLM with our 4D semantic representation through an API, enabling the model to produce visio-linguistic answers rooted in real-world scene understanding. This open-domain capability of the LLM, combined with the domain-specific 3D scene knowledge, removes the need for a specialized VLM trained on limited 4D data. The integrated pipeline for Phase 1 and Phase 2 is illustrated in Figure 1.

Phase 3 will focus on user and case studies to ensure the technology aligns with real-world needs and delivers meaningful impact. We will iteratively refine VR and desktop interfaces so that diverse stakeholders—regardless of their construction expertise—can reliably extract necessary information. Early engagement will identify the most critical stakeholders and sustainability questions and guide evaluation metrics to reflect these. The construction industry is known for its resistance to novel technologies; by prioritizing user interaction, we aim to create a reliable and relevant system that is deployable in practice.

Data: To diversify the training and validation, there will be an ongoing project effort to enrich the data sources by (i) providing human semantic annotation for the NSS construction dataset and (ii) collecting additional data (3D + time) at construction sites and in entirely constructed buildings which change. By curating high-quality, domain-specific 3D scans with rich annotations, we can perform model fine-tuning to improve performance in construction tasks.

Future Trajectory: Beyond this fellowship, site-based studies will assess how this new progress monitoring approach reduces material waste, energy use, and costs to align with sustainability goals.

References

[1] Jan Mischke, Kevin Stokvis, and Koen Vermeltfoort, "Delivering on construction productivity is no longer optional.," *McKinsey & Company*, 2024.

[2] "Efficiency eludes the construction industry,," *Economist*, 2017.

[3] N. Forcada, M. Gangolells, M. Casals, and M. Macarulla, "Factors Affecting Rework Costs in Construction," *Journal of Construction Engineering and Management*, vol. 143, p. 04017032, Aug. 2017.

[4] I. Armeni, D. Raghu, and C. De Wolf, "Artificial Intelligence for Predicting Reuse Patterns," in *A Circular Built Environment in the Digital Age* (C. De Wolf, S. Çetin, and N. M. P. Bocken, eds.), pp. 57–78, Cham: Springer International Publishing, 2024. Series Title: Circular Economy and Sustainability.

[5] Department of Economic and Social Affairs: Population Division, "World Population Prospects 2022," Tech. Rep. vol 145, Department of Economic and Social Affairs, UN, New York, USA, 2022.

[6] C. Liu, P. Zhang, and X. Xu, "Literature review of digital twin technologies for civil infrastructure," *Journal of Infrastructure Intelligence and Resilience*, vol. 2, p. 100050, Sept. 2023.

[7] A. B. Ersoz, "Demystifying the Potential of ChatGPT-4 Vision for Construction Progress Monitoring,"

[8] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning Transferable Visual Models From Natural Language Supervision," Feb. 2021. arXiv:2103.00020 [cs].

[9] OpenAI, "GPT-4 Technical Report," Mar. 2024. arXiv:2303.08774 [cs].

[10] J. Zhang, J. Huang, S. Jin, and S. Lu, "Vision-Language Models for Vision Tasks: A Survey," Feb. 2024. arXiv:2304.00685 [cs].

[11] B. Chen, Z. Xu, S. Kirmani, B. Ichter, D. Driess, P. Florence, D. Sadigh, L. Guibas, and F. Xia, "SpatialVLM: Endowing Vision-Language Models with Spatial Reasoning Capabilities," Jan. 2024. arXiv:2401.12168 [cs].

[12] Y. Hong, H. Zhen, P. Chen, S. Zheng, Y. Du, Z. Chen, and C. Gan, "3D-LLM: Injecting the 3D World into Large Language Models," July 2023. arXiv:2307.12981 [cs].

[13] C. Zhu, T. Wang, W. Zhang, J. Pang, and X. Liu, "LLaVA-3D: A Simple yet Effective Pathway to Empowering LMMs with 3D-awareness," Sept. 2024. arXiv:2409.18125 [cs].

[14] S. Ren, L. Yao, S. Li, X. Sun, and L. Hou, "TimeChat: A Time-sensitive Multimodal Large Language Model for Long Video Understanding," Mar. 2024. arXiv:2312.02051 [cs].

[15] H. Zhang, X. Li, and L. Bing, "Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding," Oct. 2023. arXiv:2306.02858 [cs, eess].

[16] J. Schult, F. Engelmann, A. Hermans, O. Litany, S. Tang, and B. Leibe, "Mask3D: Mask Transformer for 3D Semantic Instance Segmentation," Apr. 2023. arXiv:2210.03105 [cs].

[17] R. Huang, S. Peng, A. Takmaz, F. Tombari, M. Pollefeys, S. Song, G. Huang, and F. Engelmann, "Segment3D: Learning Fine-Grained Class-Agnostic 3D Segmentation without Manual Labels," Dec. 2023. arXiv:2312.17232 [cs].

[18] X. Wu, Z. Tian, X. Wen, B. Peng, X. Liu, K. Yu, and H. Zhao, "Towards Large-scale 3D Representation Learning with Multi-dataset Point Prompt Training," July 2024. arXiv:2308.09718 [cs].

[19] J. Seidenschwarz, A. Ošep, F. Ferroni, S. Lucey, and L. Leal-Taixé, "SeMoLi: What Moves Together Belongs Together," Mar. 2024. arXiv:2402.19463 [cs].

[20] K. Zeng, H. Shi, J. Lin, S. Li, J. Cheng, K. Wang, Z. Li, and K. Yang, "MambaMOS: LiDAR-based 3D Moving Object Segmentation with Motion-aware State Space Model," Aug. 2024. arXiv:2404.12794 [cs, eess].

[21] Andrew Roe, "Five Causes of Construction Rework and How to Avoid Them," Apr. 2022.

[22] T. Sun, Y. Hao, S. Huang, S. Savarese, K. Schindler, M. Pollefeys, and I. Armeni, "Nothing Stands Still: A Spatiotemporal Benchmark on 3D Point Cloud Registration Under Large Geometric and Temporal Change," Nov. 2023. arXiv:2311.09346 [cs].

[23] J. Wald, A. Avetisyan, N. Navab, F. Tombari, and M. Niessner, "RIO: 3D Object Instance Re-Localization in Changing Indoor Environments," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, (Seoul, Korea (South)), pp. 7657–7666, IEEE, Oct. 2019.

[24] B. Kerbl, G. Kopanas, T. Leimkuehler, and G. Drettakis, "3D Gaussian Splatting for Real-Time Radiance Field Rendering," *ACM Transactions on Graphics*, vol. 42, pp. 1–14, Aug. 2023.