Bootstrapping Indoor Semantic Digital Twins from 2D Video

Benjamin Alt[1], Luca Krohm[1], Patrick Mania[1], Maciej Stefańczyk[2], Artur Wilkowski[2] and Michael Beetz[1]

***Abstract*— Semantic digital twins (SemDTs) enable robots to reason about world semantics for robust real-world task planning and execution. Most existing frameworks require manual modeling of geometry and ontologies. We present an end-to-end pipeline that bootstraps SemDTs from monocular video using vision-language models. Our system reconstructs 3D geometry, segments objects, classifies them against an extensible taxonomy, and persists the result to a queryable database. The VLM dynamically proposes novel semantic classes when existing categories are insufficient. We demonstrate promising first results on a real-world kitchen environment.**

I. INTRODUCTION

Digital twins, i.e. explicit models of the environment, its dynamics and the agents within it, enable robots to solve long-horizon manipulation tasks through deliberative planning. *Semantic* digital twins (SemDTs) extend geometry and kinodynamics with explicit object semantics such as taxonomies, affordances and social functions [1], allowing robots to reason about complex world dynamics and enables human inspection and modification of generated plans.

SemDTs have traditionally required significant modeling, such as CAD models and manual semantic annotation. We envision a world in which creating semantically rich digital twins is as easy as pointing a phone camera at the scene. To that end, we propose a system that constructs complete semantic digital twins from short monocular videos in a zero-shot, end-to-end manner, reconstructing geometry, segmenting objects and inferring semantics without human labeling.

A. Related Work

3D Reconstruction. Recent advances in Structure from Motion (SfM) and multi-view stereo enable dense 3D reconstruction from monocular video [2]. Neural Radiance Fields (NeRFs) [3] and 3D Gaussian Splatting (3DGS) [4] achieve photorealistic novel view synthesis but produce implicit representations unsuitable for object-level reasoning. Surface extraction methods [5], [6] can recover explicit meshes from neural fields.

Semantic Scene Understanding. Panoptic segmentation [7] unifies instance and semantic segmentation. Recent works lift 2D segmentation to 3D via multi-view fusion [8], [9] or feature distillation into radiance fields [10], [11]. Vision-language models (VLMs) [12], [13], [14] enable open-vocabulary 3D understanding [15], including zero-shot affordance detection [16], [17] and articulation estimation [18].

Semantic Digital Twins. Prior SemDT frameworks for robotics [19], [20] require predefined CAD models or manual ontology construction. Our work proposes bootstrapping complete semantic representations end-to-end from video using VLMs, without requiring manual modeling.

B. Contribution

We present an end-to-end pipeline for creating semantic digital twins from monocular video (Fig. 1). Our contributions are (1) a complete pipeline from raw video to SemDT, integrating 3D reconstruction, panoptic segmentation, semantic classification and persistence; (2) open-vocabulary semantic understanding via VLMs that dynamically extends the object taxonomy when novel classes are encountered; and (3) zero-shot operation requiring no task-specific training or manual annotation. Our code is available as open-source.[1]

II. METHODS

Our pipeline transforms 2D video into a queryable semantic digital twin through four stages (Fig. 1).

3D Reconstruction and Segmentation. We reconstruct scene geometry using semantically-augmented Gaussian Splatting. We use COLMAP [2] to infer camera poses and initialize sparse 3D Gaussians. Each Gaussian is augmented with semantic feature vectors combining class features and instance embeddings. These are supervised with SAM3 [21] outputs using binary cross-entropy loss for class features and contrastive loss for instance embeddings, such that embedding distances reflect instance membership probabilities. We cluster the optimized per-point features into object instances with class probabilities using hierarchical DBSCAN.

In parallel, we reconstruct object meshes via a classical SfM pipeline [22] using COLMAP to obtain camera positions and a sparse point cloud, followed by multi-view stereo reconstruction based on semi-global matching to obtain dense depth images. 3D Delaunay tetrahedralization and Graph Cut Max-Flow are applied to obtain meshes. Semantic labels are transferred to the meshes by rendering the Gaussian semantics from training viewpoints and re-texturing the mesh surface.

Open-Vocabulary Semantic Classification. Each segmented mesh is classified using a VLM (Qwen3 VL 30B A3B Instruct). We render the scene from three viewpoints

*Funded by German Academic Exchange Service Grant 57754532 and Polish National Agency for Academic Exchange grant BPN/BDE/2024/1/00044

[1]B. Alt, L. Krohm, P. Mania and M. Beetz are with the AICOR Institute for Artificial Intelligence, University of Bremen, Germany {balt,krohm,pmania,mbeetz}@uni-bremen.de

[2]M. Stefanczyk and A. Wilkowski are with the Institute of Control and Computation Engineering, Warsaw University of Technology, Poland {artur.wilkowski,maciej.stefanczyk}@pw.edu.pl

[1]https://github.com/Sanic/cognitive_robot_abstract_machine/tree/semdt-creation-from-video

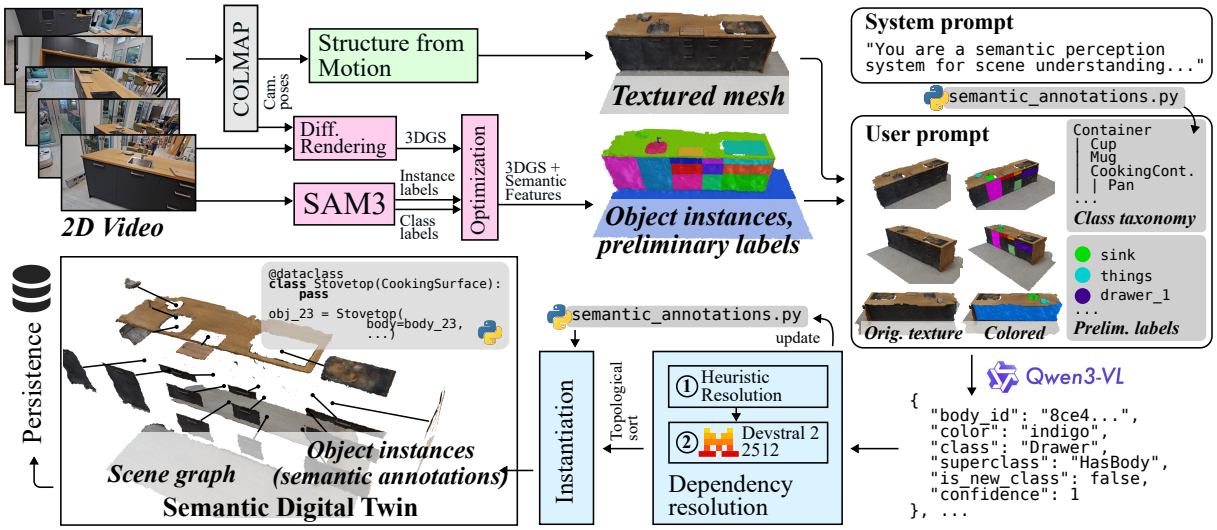

Fig. 1. Overview: Zero-shot, end-to-end construction of semantic digital twins from monocular video.

(back, diagonal front-left, diagonal front-right), highlighting target objects in distinct colors while preserving surrounding textures for context. Objects are processed in batches; for each batch, the VLM receives six images (three original, three highlighted), a hierarchical object taxonomy (class hierarchy from the SemDT codebase), and prior semantic labels from the segmentation stage. The VLM outputs structured JSON with class assignments and confidence scores. If no suitable class exists in the taxonomy, the VLM proposes new subclasses under the appropriate parents, enabling open-vocabulary understanding without a fixed label set.

Semantic Annotation Instantiation. Classification results are instantiated as semantic annotations in a three-phase process. In *class inference*, novel classes proposed by the VLM are dynamically generated from Jinja2 templates inheriting from appropriate superclasses and imported for instantiation. In *dependency resolution*, we iteratively resolve typed constructor fields (e.g., a `Cabinet` requires a `Container` reference) using domain heuristics (e.g. type hints) with optional LLM fallback (Devstral 2 2512) for ambiguous cases. In *instantiation*, annotations are topologically sorted by their dependencies and instantiated in order, ensuring all references are valid. The result is a fully instantiated SemDT that can be used for simulation, task planning and robot control by downstream applications.

Persistence. The world model (bodies, kinematic structure, semantic annotations) is persisted to PostgreSQL via an auto-generated object-relational mapping. A class diagram analyzer generates Data Access Objects (DAOs) for all annotation classes, including dynamically created ones; the database schema is updated at runtime when new classes are introduced.

III. PRELIMINARY RESULTS

We evaluated our pipeline on a real-world kitchen environment (see Fig. 1). The reconstruction and segmentation stage produced 20 mesh bodies from monocular video. The VLM classified all 20 objects and assigned them to 10 distinct semantic classes. Four novel classes not present in the base taxonomy were proposed by the VLM (`CookingSurface`, `Stovetop`, `CuttingBoard`, `Tap`) and dynamically generated at runtime.

During dependency resolution, the system inferred 12 additional `Container` annotations to satisfy constructor field constraints for `Drawer` and `Cabinet` instances. All dependencies were resolved in a single iteration using heuristics, without falling back on the LLM. The final SemDT contains 32 semantic annotations across 11 classes.

This preliminary experiment demonstrates the feasibility of SemDT construction from raw video of a real-world kitchen to a queryable, persistable semantic digital twin with automatically extended taxonomy. Limitations include the lack of articulation (e.g. doors, tap handle, ...) and the reliance on accurate upstream segmentation (drawer handles were not segmented). Future work will evaluate on additional environments and integrate physics and affordances.

IV. CONCLUSION

We presented an end-to-end pipeline for bootstrapping semantic digital twins from monocular video that integrates open-vocabulary scene understanding, integration with existing taxonomies, and compatibility with the CRAM cognitive architecture [23] for virtual reality (VR) simulation and robot planning.

Limitations. Our evaluation is limited to a single real-world kitchen environment. Furthermore, the appropriate level of semantic abstraction is application-dependent and requires further experimentation.

Outlook. Future work will include benchmarking on large-scale scene understanding datasets and qualitative evaluation on diverse real-world environments. We will assess the utility of the generated SemDTs for robotic mobile manipulation, where the required semantic granularity can be empirically determined by task performance.

REFERENCES

[1] A. Melnik, B. Alt, G. Nguyen, A. Wilkowski, Q. Wu, S. Harms, H. Rhodin, M. Savva, M. Beetz, *et al.*, "Digital twin generation from visual data: A survey," *arXiv preprint arXiv:2504.13159*, 2025.

[2] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4104–4113.

[3] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[4] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering." *ACM Trans. Graph.*, vol. 42, no. 4, pp. 139–1, 2023.

[5] P. Wang, L. Liu, Y. Liu, C. Theobalt, T. Komura, and W. Wang, "Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction," *arXiv preprint arXiv:2106.10689*, 2021.

[6] H. Chen, C. Li, Y. Wang, and G. H. Lee, "Neusg: Neural implicit surface reconstruction with 3d gaussian splatting guidance," *arXiv preprint arXiv:2312.00846*, 2023.

[7] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollár, "Panoptic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 9404–9413.

[8] J. Schult, F. Engelmann, A. Hermans, O. Litany, S. Tang, and B. Leibe, "Mask3d: Mask transformer for 3d semantic instance segmentation," *arXiv preprint arXiv:2210.03105*, 2022.

[9] A. Takmaz, E. Fedele, R. W. Sumner, M. Pollefeys, F. Tombari, and F. Engelmann, "Openmask3d: Open-vocabulary 3d instance segmentation," *arXiv preprint arXiv:2306.13631*, 2023.

[10] S. Kobayashi, E. Matsumoto, and V. Sitzmann, "Decomposing nerf for editing via feature field distillation," *Advances in neural information processing systems*, vol. 35, pp. 23 311–23 330, 2022.

[11] M. Qin, W. Li, J. Zhou, H. Wang, and H. Pfister, "Langsplat: 3d language gaussian splatting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 20 051–20 060.

[12] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[13] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, pp. 34 892–34 916, 2023.

[14] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[15] S. Peng, K. Genova, C. Jiang, A. Tagliasacchi, M. Pollefeys, T. Funkhouser, *et al.*, "Openscene: 3d scene understanding with open vocabularies," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 815–824.

[16] T. Nguyen, M. N. Vu, A. Vuong, D. Nguyen, T. Vo, N. Le, and A. Nguyen, "Open-vocabulary affordance detection in 3d point clouds," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 5692–5698.

[17] G. Li, D. Sun, L. Sevilla-Lara, and V. Jampani, "One-shot open affordance learning with foundation models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 3086–3096.

[18] L. Le, J. Xie, W. Liang, H.-J. Wang, Y. Yang, Y. J. Ma, K. Vedder, A. Krishna, D. Jayaraman, and E. Eaton, "Articulate-anything: Automatic modeling of articulated objects via a vision-language foundation model," *arXiv preprint arXiv:2410.13882*, 2024.

[19] P. Mania, S. Stelter, G. Kazhoyan, and M. Beetz, "An open and flexible robot perception framework for mobile manipulation tasks," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 17 445–17 451.

[20] M. Kümpel, C. A. Mueller, and M. Beetz, "Semantic digital twins for retail logistics," in *Dynamics in logistics: Twenty-five years of interdisciplinary logistics research in Bremen, Germany*. Springer International Publishing Cham, 2021, pp. 129–153.

[21] N. Carion, L. Gustafson, Y.-T. Hu, S. Debnath, R. Hu, D. Suris, C. Ryali, K. V. Alwala, H. Khedr, A. Huang, *et al.*, "Sam 3: Segment anything with concepts," *arXiv preprint arXiv:2511.16719*, 2025.

[22] AliceVision, "AliceVision Photogrammetric Computer Vision Framework," https://alicevision.org/#photogrammetry, [Accessed 11-01-2026].

[23] M. Beetz, G. Kazhoyan, and D. Vernon, "Robot manipulation in everyday activities with the cram 2.0 cognitive architecture and generalized action plans," *Cognitive Systems Research*, p. 101375, 2025.