

Reading status. The detailed review below preserves the earlier literature synthesis and its proposed ARIA-NBV adaptations. Recommendations about model inputs, training stages or implementation are research proposals, not a description of the current system or evidence of measured benefit. The active thesis and implementation specify the current choices.

SceneScript: Structured Scene Language As Semantic Bridge

Primary source. SceneScript: Reconstructing Scenes With An Autoregressive Structured Language Model [1].

Local source. `main.tex`, `sections/dataset.tex`, and `sections/structured_scene_language.tex`.

Related ARIA-NBV pages. Project Aria, EFM3D/EVL, and target-aware research questions.

Core contribution

SceneScript represents indoor scenes as a structured language of geometric commands. Instead of producing only a dense mesh or point cloud, the model predicts an editable sequence of entities such as walls, doors, windows, and object bounding boxes [1].

The ARIA-NBV relevance is semantic structure, not immediate planner replacement. SceneScript shows that Aria Synthetic Environments (ASE) scale, Machine Perception Services (MPS)-style point clouds, and typed scene primitives can support a compact global representation that might later guide target selection or semantic subgoals.

Figure 1 separates predicting a structured description from interpreting it as geometry.

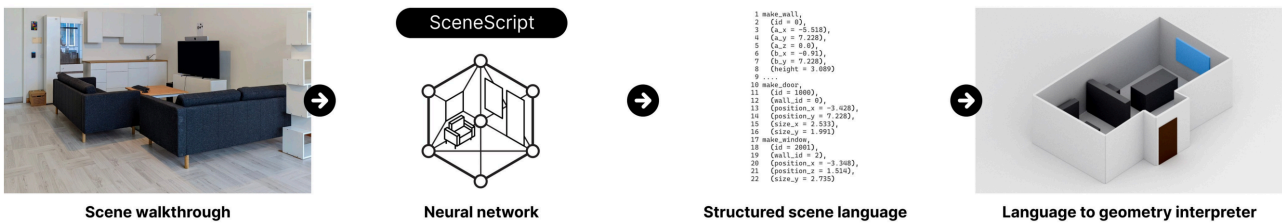


Figure 1: SceneScript maps an egocentric walkthrough to structured scene language, which a geometry interpreter turns into a scene model. Reproduced from [1, Fig. 1, top panel].

Verified paper signals

signal	source-backed detail	ARIA-NBV relevance
Structured language	Scenes are encoded as commands such as <code>make_wall</code> , <code>make_door</code> , <code>make_window</code> , and <code>make_bbox</code> .	Useful stretch representation for named targets, layout regions, and semantic planning.
Input evidence	The model consumes point clouds from Project Aria / MPS-style reconstruction, discretized to a fixed spatial resolution.	Aligns with ARIA-NBV’s semi-dense observed-state premise.
Architecture	A sparse 3D encoder feeds an autoregressive Transformer decoder with grammar/type constraints.	Suggests a future sequence representation, not a required thesis component.
ASE scale	SceneScript uses large-scale ASE scenes and annotations.	Supports ASE as the semantic/global planning ecosystem around ARIA-NBV.
Editability	Structured commands can be inspected and edited more directly than opaque dense features.	Useful for human-facing target definitions and future semantic diagnostics.

Example command types:

command	role
make_wall	architectural wall primitive
make_door / make_window	openings attached to walls
make_bbox	object-level bounding box/entity
make_prim	extensible primitive command in later variants

ARIA-NBV adoption

- **Stretch/bridge:** SceneScript is a semantic/global representation layer after observed target selection, target RRI labels, finite-candidate rollouts, and Finite-Horizon Q Function (Q_H) training are stable.
- **Proposal/diagnostic:** typed scene entities can help organize reports by walls, objects, doors, or rooms.
- **Future target interface:** structured commands could become a human-readable target source, but must be matched to observed/predicted Oriented Bounding Box (OBB) or support evidence before being actor-visible.

Do not adopt

- Do not make SceneScript a thesis-core dependency.
- Do not use GT semantic commands or GT entity layouts as actor-visible inputs in the main ARIA-NBV protocol.
- Do not assume high-level layout correctness is enough for fine-detail target-RRI supervision.
- Do not replace geometric validity, candidate masks, or mesh-supervised RRI with language-level scene completeness.

Open risks / caveats

- Structured language can miss fine geometry that matters for Chamfer-style reconstruction quality.
- Command vocabularies and discretization choices constrain what can be represented.
- SceneScript is valuable as a future semantic/global planning bridge, but the current thesis must first prove target-conditioned quality-driven NBV on ASE/EFM evidence.

Terminology

Dataset

ASE – Aria Synthetic Environments: Synthetic indoor dataset with simulated Aria observations, trajectories, and scene annotations. The thesis uses ASE reference meshes and rendered depth to measure target reconstruction improvement. Semidense actor evidence is a separate input, not the default oracle reconstruction source.

MPS – Machine Perception Services: Project Aria services that derive poses, mapping, and other perception outputs from recordings. MPS-style poses and semidense points describe available sensing products. They should not be confused with the ASE-depth reconstruction used by the current oracle evaluation.

Geometry

OBB – Oriented Bounding Box: Box described by a center, orientation, and three extents. It approximates an object extent and defines a local target frame. Semantic class and confidence can accompany the detection but are not geometric properties of the box itself.

Model

Q_H – Finite-Horizon Q Function: Expected finite-horizon return after taking a specified candidate action first. The continuation policy determines later choices. The ideal optimal value maximizes over policies using the admitted actor information; a learned scorer approximates it from a finite representation. A sampled return, the fitted training target and the predicted value are distinct. The Q function specifies neither encoder architecture nor training algorithm.

Bibliography

- [1] A. Avetisyan *et al.*, “SceneScript: Reconstructing Scenes With An Autoregressive Structured Language Model.” [Online]. Available: <https://arxiv.org/abs/2403.13064>