

Hestia: Direct Pose Prediction for Object Reconstruction

Hestia learns an NBV policy for object-centric capture that predicts a continuous viewpoint instead of ranking a sampled candidate table. It first predicts a look-at point, then a camera position conditioned on that point; pitch and yaw follow from their geometric relation. The predicted position is moved to a nearby collision-free location before acquisition [1, Secs. 1 and 3.2].

Reviewed method: arXiv:2508.01014v4, Sections 3.1–3.2. The locally cached PDF is v3; the version-four method claims below were checked against the official v4 text.

Figure 1 shows the hierarchy in the cached version-three architecture; the method discussion below was checked against version four.

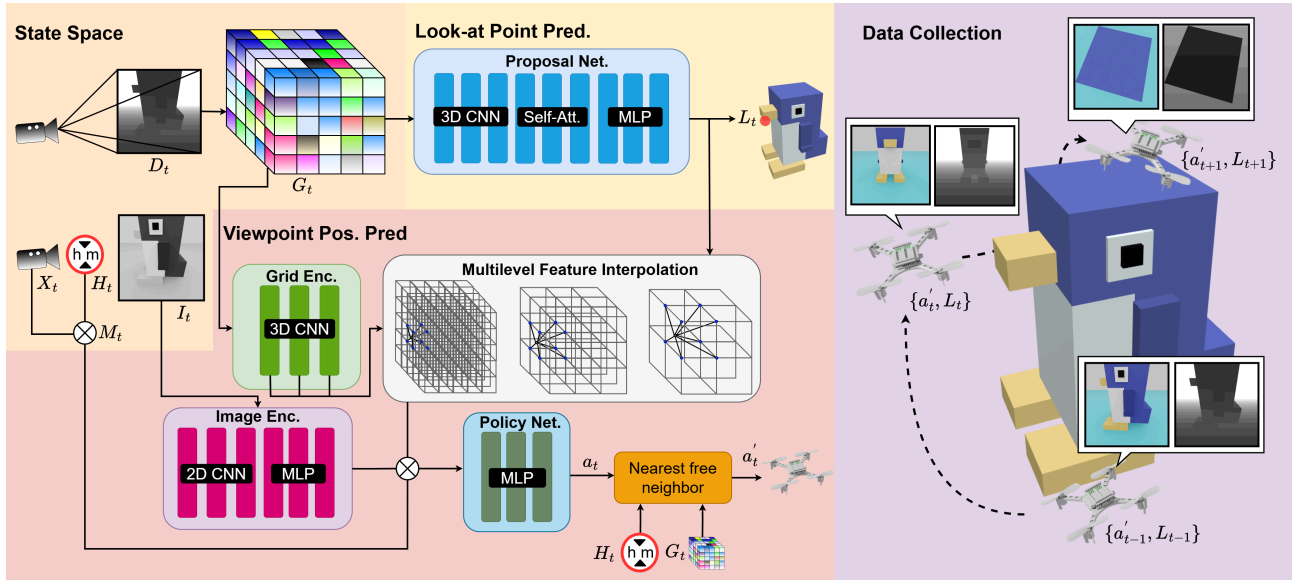


Figure 1: Hestia predicts an internal look-at point before a viewpoint position, then relocates that position to free space. This is the version-three source figure. Reproduced from [1, Fig. 2; arXiv v3].

What the Method Values

Its state accumulates occupancy and six directional face-visibility indicators per voxel. Rewards favour newly observed voxel faces. The paper sets the discount to $\gamma = 0.1$ to emphasize immediate coverage; the look-at proposal is supervised by uncaptured ground-truth surface. Thus hierarchical pose prediction and repeated acquisitions do not by themselves establish planning for delayed reconstruction benefit [1, Secs. 3.1–3.2].

Relevance and Boundary

Hestia provides a concrete example beyond GenNBV for explaining direct pose prediction. Its look-at point is an internally proposed attention location, whereas the thesis conditions scoring on a requested object embedded in a larger scene. A proposed look-at location therefore does not provide object identity or establish that a particular object’s reconstruction improved.

Hestia shows a different action representation from finite candidate ranking. Changing the action representation also changes which views can be selected, so its benefits cannot be isolated by comparing unmatched policies. Hestia’s object-centric capture and drone demonstration do not establish performance for requested-object reconstruction in egocentric indoor scenes.

Detailed review and retained proposals

What the target means

Hestia’s internal look-at point is a predicted attention location, not an externally requested semantic instance. Its main benchmarks scan individual scaled objects, while real-drone demonstrations and a qualitative complex-scene example broaden the tested settings. Neither establishes requested-object selection among cluttered distractors. The selected camera position is relocated to collision-free space; the reward counts newly observed voxel faces with discount 0.1. Evaluating geometric reconstruction afterward does not change that coverage reward into reconstruction-error supervision. See version 4, Sections 3.1–3.2, 4.1, 4.8 and supplementary S4/S9.

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.

Hestia: Hierarchical Continuous NBV Bridge

Primary source. Hestia: Voxel-Face-Aware Hierarchical Next-Best-View Acquisition for Efficient 3D Reconstruction [1].

Local source. `main.tex` and section sources under `tex-src/arXiv-Hestia/sec/`.

Related ARIA-NBV pages. GenNBV, RL planning, and roadmap.

Core contribution

Hestia is the most relevant continuous-policy bridge in the current corpus. It remains coverage-driven, but it improves the continuous NBV control structure by decomposing action selection into a look-at proposal and a feasible camera-position policy, and by tracking directional voxel-face visibility [1].

For ARIA-NBV, the transferable idea is not the coverage reward. It is the factorization:

$$\text{target/look-at proposal} \rightarrow \text{feasible camera pose}.$$

Verified paper signals

signal	source-backed detail	ARIA-NBV relevance
Face-aware state	Hestia tracks directional visibility of voxel faces rather than only occupancy.	Inspires directional observation histograms for semi-dense points, EVL voxels, or target crops.
Hierarchical action	The controller predicts a look-at point and a 3-DoF position; yaw/pitch are derived from the look-at relation.	Future bridge from finite candidates to target-then-pose continuous control.
Feasibility projection	Predicted positions are adjusted to collision-free locations.	Mirrors ARIA-NBV’s need for hard validity masks and feasible candidate transitions.
Close-greedy reward	The paper uses immediate face-coverage gains with a small discount factor.	Supports short-horizon evidence gates before long-horizon RL claims.
Intermediate supervision	The look-at proposal is supervised by a target derived from uncaptured surface geometry.	Suggests supervised target/missing-surface proposal heads after target-RRI labels are trusted.

Directional face visibility can be summarized as:

$$F_t = f_t \vee F_{t-1}, \quad f_t(v_i, j) = \mathbb{1} \left(\frac{a'_t - p_{v_i}}{\|a'_t - p_{v_i}\|} \cdot n_{i,j} > 0 \right).$$

ARIA-NBV adoption

- **Stretch/bridge:** use Hestia to design a later continuous target-then-pose actor, not as the first quantitative thesis baseline.
- **Proposal/diagnostic:** add directional observation features or incidence histograms to candidate reports.
- **Validity:** keep feasibility projection and invalid-action handling as hard constraints rather than reward-only penalties.

Do not adopt

- Do not replace Relative Reconstruction Improvement (RRI) or Target-Specific RRI (target RRI) with face coverage.
- Do not make continuous actor-critic training mandatory for the thesis result.
- Do not transfer object-centric Objaverse coverage results as evidence for ASE/EFM target-conditioned NBV.

Open risks / caveats

- Directional face coverage is still a proxy and may disagree with Chamfer-style quality.
- Continuous policies need simulator fidelity, collision handling, and reward throughput that ARIA-NBV does not yet have.
- Hestia is most useful after finite-candidate oracle rollouts and Finite-Horizon Q Function (Q_H) learning establish the quality-driven baseline.

Terminology

Metrics

RRI – Relative Reconstruction Improvement: Reduction in reconstruction error relative to the error before one acquisition. Target-specific RRI measures this change on the selected object. Oracle RRI obtains the measurement from privileged reference geometry. Unlike the root-gain reward, its denominator changes with the current reconstruction.

target RRI – Target-Specific RRI: State-relative reconstruction improvement restricted to the selected target. It specializes RRI to the target geometry and differs from the root-normalized reward used to accumulate gain across a rollout.

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.

Planning

candidate table – Candidate Table: Finite collection of proposed camera poses and their aligned row data at one decision. A candidate shell need not lie on a sphere. The admissible action set contains only the row indices retained by the validity mask.

NBV – Next-Best-View: Selection of a sensor viewpoint to improve reconstruction under a limited acquisition budget. A myopic selector ranks immediate improvement; a finite-horizon planner also values views that enable useful later acquisitions.

Bibliography

- [1] C.-Y. Lu *et al.*, “Hestia: Voxel-Face-Aware Hierarchical Next-Best-View Acquisition for Efficient 3D Reconstruction,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2026. [Online]. Available: https://openaccess.thecvf.com/content/WACV2026/papers/Lu_Hestia_Voxel-Face-Aware_Hierarchical_Next-Best-View_Acquisition_for_Efficient_3D_Reconstruction_WACV_2026_paper.pdf