AEC Interpreter: Cross-Modal Alignment and Ontology-Driven Retrieval for Site-to-BIM Element Grounding

Chia Hui Yen
Carnegie Mellon University
huiyenc@andrew.cmu.edu

2026

Abstract

Architecture, Engineering, and Construction workflows still depend on people translating egocentric site evidence — photographs, marked plans, and short notes — into allocentric BIM records. Before an AI system can reason about status, schedule, or corrective action, it must answer a more basic spatial question: *which exact model element is this?* We introduce a neuro-symbolic image-to-BIM grounding method that maps a construction-site photograph and natural-language note to a *unique* IFC element GUID using only the BIM model itself for supervision, with no real on-site labels. The task is hard because many candidate elements are *visually identical*, the target identity lives in *graph* space while the evidence lives in *image* space, and a cold-start deployment has no paired site-image training set. Our method recovers a **type-conditional spatial address** — a class-specific relational key, such as an opening's ordinal position-slot or a wall's connectivity fingerprint, computable from the raw IFC model with no labels. The opening slot is also image-recoverable and we realize it; the wall fingerprint is oracle-discriminative but *not* image-realizable (its fields collapse in the floorplan), so its numbers are reported at oracle level only. A vocabulary-constrained neural layer reads coarse semantics; deterministic visual specialists recover discriminating address fields; and a symbolic executor consumes a per-field `{value, confidence, source}` contract through recall-safe reranking and selective prediction. Supplied at an oracle level, the address lifts pool Top-1 from 4.9% to **78.5%** at zero recall cost (an upper-bound ceiling, all 60 cases); the one realized extractor — a deterministic position-slot detector — lifts the addressable filler subset ($n = 35$) from a 6.6% floor to **58.9%** end-to-end (coarse fields also extracted; a 67.6% upper bound if they are supplied at oracle level); its confidence passes an ECE calibration gate (AUROC 0.80); and deferring the least-confident fifth raises answered-set Top-1 to **73.4%**. The result is an auditable spatial multimodal grounding layer for AEC: ontology is the guardrail, topology is the discriminator, and the system knows when to abstain.

Keywords: AEC, spatial intelligence, neuro-symbolic grounding, IFC, vision-language models

1 Introduction

1.1 Motivation: bridging physical site evidence and digital truth

AEC coordination moves constantly between two incompatible views of the same building. On site, evidence is egocentric and fragmented: a phone photograph, a marked floorplan crop, a short chat message from a front-line worker. In the project record, truth is allocentric and structured: a BIM

model, IFC element types, topology, and GUIDs. The traceability gap is the missing translation between these two views. A coordinator does not merely need to know that an image contains *a window*; they need to know which of the many near-identical windows in the BIM cloud should receive the issue, handoff record, or BCF link. Before an AI system can answer higher-level AEC questions — what happened, who is responsible, what action should follow — it must solve the foundational grounding question: *where exactly is it in the model?*

The setting is deliberately cold-start. A construction or facilities team has a complete IFC/BIM model but no history of labelled site imagery; on day one they want to point a phone at an element, add a short note, and be told *which* modelled element it is. The final step is a selection from a retrieved candidate pool in which the ground truth is present (median 76, in-pool 100%), so the difficulty is not retrieval but **discrimination among visually identical siblings**, in a regime with no paired supervision to learn that discrimination from. Our strongest learned configuration — a fine-tuned-VLM extraction pipeline feeding deterministic retrieval, with no within-pool reranking — plateaus at Top-1 6.7% with the answer already in the pool, because the signal that separates two identical windows is not in either window's pixels but in the building's relational structure, which a black-box matcher would have to reconstruct internally and unreliably (we develop this design rationale in §2.1 [12]).

1.2 Research questions

This paper makes one contribution — an **ontology-grounded, topology-derived spatial address** and the calibrated routing that makes it usable. Ontology supplies the floor and the guardrail: the vocabulary, storey, class, and schema-compliant execution boundary that prevent hallucinated element IDs. Topology supplies the discriminator: the local relational structure that separates visually identical siblings. The contribution is developed through three research questions:

- **RQ1 — Representation (§3.2).** *What is the minimal sufficient spatial address?* We show it is **type-conditional**: a coarse ontological prefix (storey + class) that is necessary but saturated, completed by a class-specific topological body — the position-slot (i, M) for fillers, a connectivity fingerprint for walls (Figure 1 gives one worked address per class). At an oracle level it reorders the pool from Top-1 4.9% to **78.5%**, with the gain partitioned by element type (fillers 91%, walls 64%).
- **RQ2 — Mechanism (§3.5).** *How is a noisily-recovered address used without losing recall, and how much of the ceiling is realizable?* Hard filtering destroys recall, so the address is a **soft prior in a recall-fixed pool**; one real extractor lifts the addressable Top-1 from 6.6% to **58.9%** end-to-end (a 67.6% upper bound if the coarse fields are supplied at oracle level), its confidence passes an ECE gate (AUROC 0.80), and — since reweighting a finest-grained prior is a no-op — its payoff is **selective prediction**: deferring the least-confident ∼20% raises answered-set Top-1 to **73.4%** [6, 7].
- **RQ3 — Architecture (§3.6).** *Where should the relational context live — extracted deep at inference, or compiled into the node?* A **depth law**: under an oracle the confusable set is already a singleton at one hop ($|C|$ median $13 \to 1$), so deeper hops add nothing in principle, and the realized gain saturates at one hop for an *informational* reason — the model reads deep relation *types* reliably (96.7/93.9/69.6% at hop 1/2/3), but those types are homogeneous (every window `FILLS` a wall, every wall `CONNECTS` a wall) and so carry no discriminating power; the discriminating signal (neighbour identity) is image-recoverable only as the filler

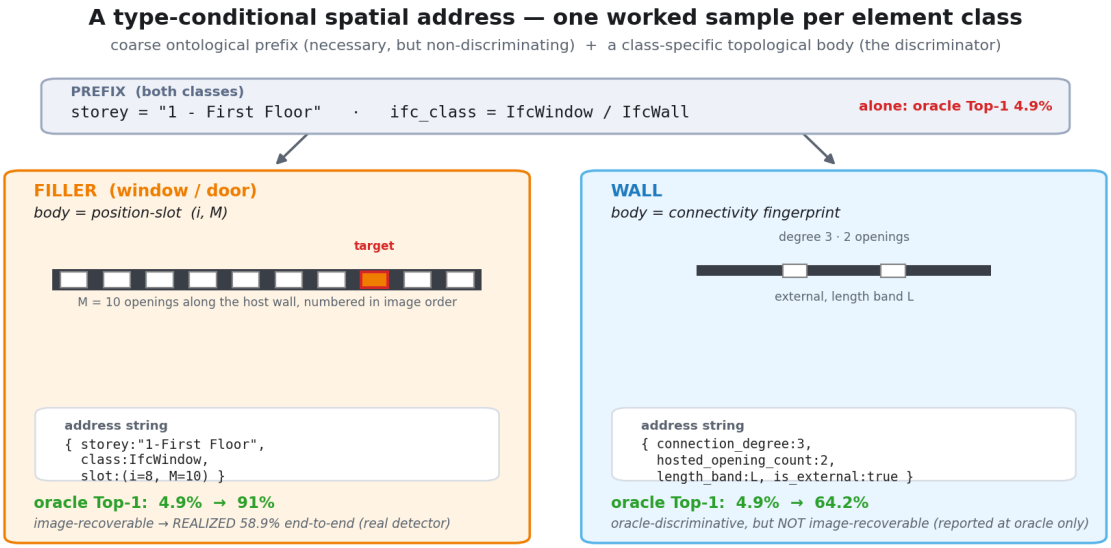

Figure 1: What a **type-conditional spatial address** is, by class — one worked sample each. A coarse ontological prefix (storey + IFC class) is shared but *non-discriminating* (oracle Top-1 4.9% alone). The discriminating body is class-specific: a position-slot (i, M) for fillers ("the i-th of M openings along the host wall") and a connectivity fingerprint for walls. Each reorders the *same* retrieved pool to the per-class oracle Top-1 shown; the filler slot is image-recoverable and realized (58.9% end-to-end), the wall fingerprint is reported at oracle level only.

position-slot (Figure 2). The architecture therefore compiles depth into the node and extracts at depth ≤ 1, placing learning at the neural→symbolic interface [10].

1.3 Related work and positioning

Prior work approaches the image-to-building-model problem from two directions that do not meet in the middle.

Vision-language spatial perception. One line endows VLMs with spatial reasoning by training on synthetic spatial-QA data; SpatialVLM demonstrates that metric and relational spatial judgments are learnable from generated supervision [3], and VLM-based scene-graph extraction recovers object-relation structure from images [14]. These systems answer in *image* space — a region, a relation, a distance — and stop at the image boundary: none of them names an element in a building's authoritative model.

LLM-driven graph retrieval. A second line connects language models to graph stores. Document-level GraphRAG builds its graph by LLM extraction from text [5], which is noisy in AEC because element attributes and spatial relationships are rarely verbalized; text-to-Cypher generation hallucinates node and relationship names and requires LLM-supervised verification scaffolding to approach reliability [8, 13]. IFC-native systems avoid the noisy-graph problem by building on the schema directly: IFC-graph converts the model into a queryable property graph [15], Graph-RAG pipelines extract information from IFC data [9], and MCP4IFC drives IFC-based design through LLM tool

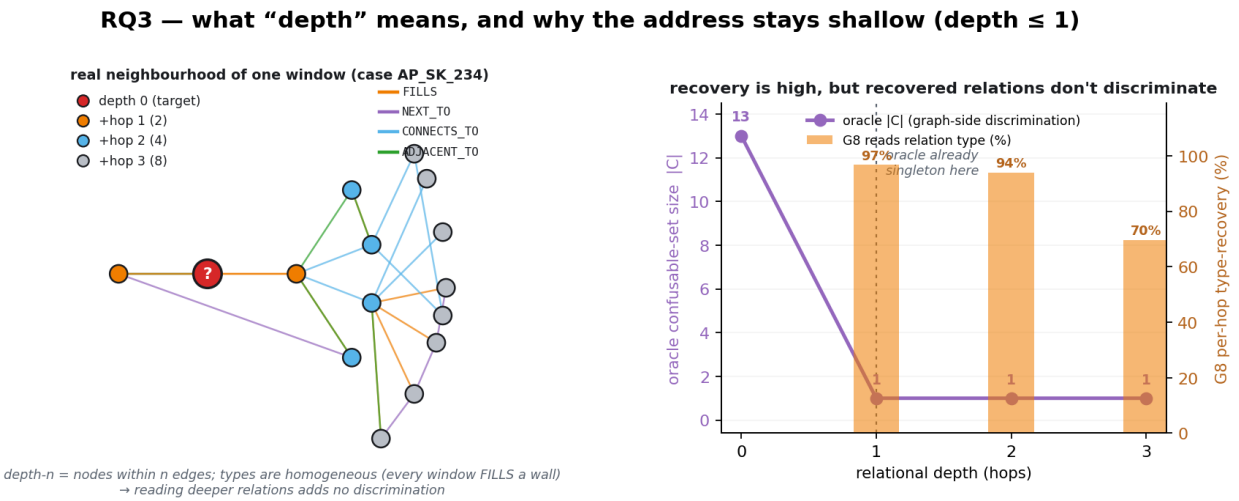

Figure 2: RQ3 made concrete. *Left:* a depth-n neighbourhood is every node within n edges of the target; each added hop (FILLS / NEXT_TO, then CONNECTS_TO, then ADJACENT_TO) pulls in more relational context. The model *reads* these relation types reliably (96.7/93.9/69.6% at hop 1/2/3 on the held-out set), but the types are homogeneous (every window `FILLS` a wall) so they do not discriminate. *Right:* the oracle confusable set $|C|$ is already a singleton at one hop, so deeper hops add nothing in principle; realized discrimination saturates at one hop for an informational reason, not an extraction-reliability cascade. This is the depth law that keeps the address shallow.

calls [11]; Text2BIM generates building models from language [4]. All of these, however, operate on *structured or textual* queries — none grounds unstructured multimodal site evidence to an element.

Positioning. To our knowledge, this is the first work to combine open-vocabulary multimodal perception with IFC-native deterministic graph retrieval for *element-level* grounding in construction site workflows. Prior IFC-based systems [9, 11] operate on structured queries; prior VLM-based spatial systems [3] lack IFC-grounded retrieval. This work bridges the gap, placing the system in the high-determinism, multimodal-input quadrant of the landscape: flexible vision-language perception on the input side, ontology-constrained graph execution on the output side, with a calibrated routing layer between them. The neuro-symbolic precedent — neural perception feeding a symbolic executor, with learning at the interface — is the concept-learner line [10].

1.4 Contributions

1. **A neuro-symbolic interpreter architecture** that maps unstructured multimodal AEC evidence (photo + note + plan) to IFC element GUIDs via a structured intermediate representation — a per-field `{value, confidence, source}` contract — achieving hallucination-resistant retrieval: the neural layer describes evidence, but only the symbolic layer names GUIDs (§2).

2. **A type-conditional spatial address** — the ordinal position-slot for openings and a connectivity fingerprint for walls — computable from the raw IFC model with no human labels, that reorders the retrieved pool from Top-1 4.9% to 78.5% at an oracle level. The opening slot is image-recoverable and we realize it (58.9% end-to-end on the filler subset); the wall

fingerprint contributes the oracle gain but is *not* image-realizable and is reported at oracle level only (§3.2, §3.5).

3. **An oracle-based evaluation methodology** that decomposes retrieval failure into the symbolic-layer ceiling versus neural extraction quality — demonstrating 100% ground-truth retention, isolating extraction as the bottleneck, and yielding a recall-safety principle (union, never intersection, when combining mature and immature signals) and a measured depth law for relational context (§3).

4. **An empirical analysis of where learned and deterministic components belong**: a per-field profile of a LoRA fine-tuned VLM showing saturation on coarse fields and failure on discriminating ones, motivating delegation to deterministic visual specialists with calibrated, selectively-predictive routing (ECE gate, defer-on-low-confidence) rather than further fine-tuning (§3.3, §3.5).

1.5 Scope, stated up front

Two honesty boundaries frame every result. First, the diagnostic ceilings (RQ1, and the oracle rows of the evaluation) assume perfect extraction; the realizable numbers come from one deterministic extractor (the position-slot), with the wall fingerprint and remaining descriptors demonstrated at oracle level only — full realization is future work. Second, all measurements are on a single synthetic project (cold-start by design); the *form* of the contribution — a type-conditional, shallow, image-recoverable, calibrated-soft address — is general, but the specific reliabilities are project-specific. To keep the claim anchored beyond our own ablations, we add standard lexical and dense retrieval baselines and a triage-effort proxy: generic retrievers plateau at Top-1 1.7% / Top-10 15.6–25.0%, the realized position-slot specialist closes the filler subset to 58.9% Top-1 end-to-end, and the perfect-address ceiling reduces expected inspection effort from a median 38 candidates to 0.5. We do not claim to beat an end-to-end matcher in the large-data limit; we claim that in the cold-start regime we can actually measure, the structured address is what makes grounding realizable, auditable, and transferable.

2 System Design

2.1 Design rationale: four decisions, each against a measured alternative

The architecture is best explained by the alternatives it rejects. Each key decision below names the rejected alternative, the evidence against it, and the trade-off accepted; Figure 3 collects the four head-to-head measurements.

Why neuro-symbolic over end-to-end neural matching? The obvious design — train a single cross-attention model to compare the site evidence against the candidates and rank them end-to-end — is, in its strongest "bitter-lesson" form [12], the objection the design must answer, and we answer it empirically. First, the candidates are graph nodes, not images: they carry no photograph, and an image-space grounding result (a region in a photo or plan) does not reach a GUID, because image space and graph space are not co-registered. The only bridge is a representation independent of the image's coordinate frame — a coordinate-free relational address. Second, the discriminative signal is relational, not pixel-local: what separates two visually identical windows is graph topology ("the third of five fillers along an external wall"), which a pixel matcher would have to re-derive inside its own weights — unreliably, as the classical opening-count baseline (∼27%) illustrates. Third,

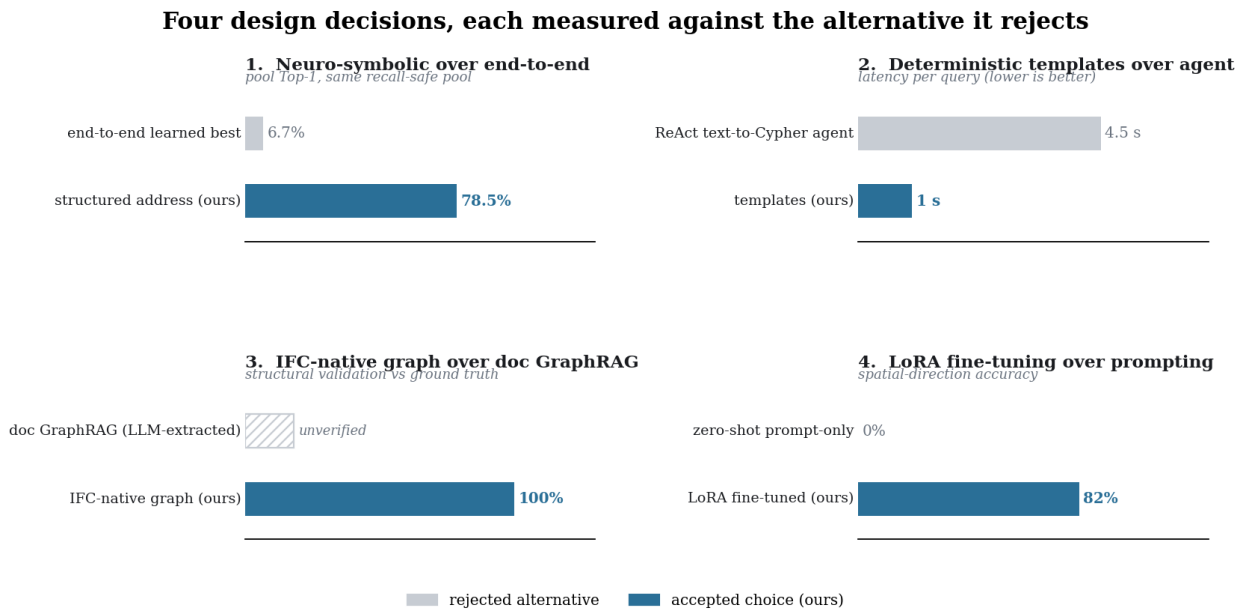

Figure 3: Four design decisions, each against the alternative it rejects, on a single measured quantity (grey = rejected, colour = accepted). (1) On the same recall-safe pool, the strongest learned end-to-end configuration ranks the target at Top-1 6.7% while the structured address reaches 78.5%. (2) Deterministic query templates run at ∼1 s with 100% syntax correctness and exact repeatability, versus a ∼4.5 s ReAct text-to-Cypher agent that picks different tools per run. (3) The IFC-native graph is validated against ground truth (wall `connection_degree` 14/14), whereas a document-GraphRAG graph is LLM-extracted and unverified. (4) LoRA fine-tuning lifts spatial-direction accuracy from ∼0% to 82% (coarse fields 30–63% → 100%).

our strongest *learned* configuration plateaus on this discrimination: the fine-tuned-VLM extraction pipeline (*G8*), feeding deterministic retrieval with no within-pool reranking, reaches Top-1 6.7% with the answer in the pool 100% of the time, and generic lexical/dense retrievers over the same pool reach Top-1 $\leq$ 1.7% (§3.2); the same pool reorders to 78.5% when the relational address is supplied. We do not train a dedicated end-to-end cross-attention matcher over the pool: its inputs would be the site image against text-identical sibling candidates (the graph nodes carry no images), so it would have to recover the same relational signal internally — the configuration we argue against; a zero-shot VLM reranker as the off-the-shelf upper-bound of that family is left as a baseline to add (§4.2). Finally, construction compliance requires repeatability and audit: a scalar matching score is not auditable, not correctable, and cannot be gated by per-field calibrated confidence. Trade-off accepted: less end-to-end flexibility, in exchange for recall guarantees, auditability, and zero-shot transfer of the address to any IFC model.

Why deterministic query templates over text-to-Cypher? LLM-generated Cypher produces syntax errors and hallucinates node and relationship names; even dedicated systems need an LLM-supervised generation-verification loop to approach reliability [8, 13]. Our query patterns are enumerable (nine strategies) and domain-specific, so a fixed template with parameter slots gives 100% syntax correctness. The agentic alternative was tested directly (timings from prior thesis work): a ReAct-style agent ran at ∼4.5 s versus ∼1 s for the constrained pipeline and chose different tool sequences for the same input — unacceptable where the same photo must always produce the

same result. Trade-off accepted: only predefined query patterns are covered, not arbitrary questions.

Why an IFC-native graph over document-level GraphRAG? Document GraphRAG builds its graph by LLM extraction from text [5] — noisy in AEC, where element attributes and spatial relationships are rarely verbalized. Our graph is compiled from the IFC schema directly [1, 15], so the graph itself is zero-error ground truth. Trade-off accepted: the system requires a project IFC file and cannot answer questions about things absent from the model.

Why fine-tuning over prompting-only? Few-shot prompting is stateless — it cannot improve over a project's lifecycle. LoRA fine-tuning accumulates field feedback as supervised examples for periodic retraining, and runs locally at ∼1 s with no per-call API cost versus 3–5 s for a prompted frontier model. Trade-off accepted: a synthetic data pipeline and retraining infrastructure are required (§2.5).

2.2 Architecture overview

The pipeline has five stages: multimodal input (photo + note + plan crop) → neural extraction into a typed record → a structured constraint contract → deterministic query compilation against the IFC graph → a ranked GUID shortlist or an explicit deferral. Figure 4 traces a single case through the spine, with the confidence-routing path — from the extractors' `{value, confidence, source}` records to the answer/defer gate — highlighted; that path is the system's determinism↔adaptivity mechanism. The evidence flow is intentionally ordered from raw observation to auditable execution: a vocabulary-constrained VLM reads the coarse semantic prefix; deterministic specialists recover the fields the VLM does not carry reliably; the IFC graph supplies the legal candidate universe and precomputed topology; and the routing layer decides whether the recovered address is confident enough to answer or should surface a candidate set for human verification. This ordering is the hallucination control: **the neural layer describes evidence, but the symbolic layer alone names GUIDs.**

2.3 IFC parse engine: from a linear file to a query-ready graph

The symbolic substrate is built once per project by compiling the raw IFC file into a property graph (IfcOpenShell → Neo4j), following the IFC-as-graph line [15]. Beyond the schema's containment hierarchy, the engine enriches the graph with the spatial topology the address needs: `ON_STOREY` from spatial containment, `FILLS` from `IfcRelFillsElement`/`IfcRelVoidsElement` chains, `CONNECTS_TO` from `IfcRelConnectsPathElements` wall connectivity, and `ADJACENT_TO` as generic proximity (centroid distance under 1500 mm) [1]. A graph database rather than a vector store is a requirement, not a preference: the discriminating signal is multi-hop topology with a deterministic schema, which embedding similarity cannot express. The reconstruction is validated against the dataset's own skeleton — the recovered wall `connection_degree` matches ground truth in 14 of 14 checked cases — so the graph exists, correct and complete, on day one with no human annotation: the cold-start property the system claims. Figure 6 traces this compilation: a flat STEP file whose only navigable structure is containment, the per-relation enrichment rules with their edge counts on the studied project, and the resulting dense, query-ready graph.

2.4 Neural layer: a fine-tuned VLM as the perception engine

The perception engine is Qwen2.5-VL-7B with LoRA adapters (rank 32 in the strongest configurations, targeting Q/K/V/O and MLP layers across both the vision encoder and the lan-

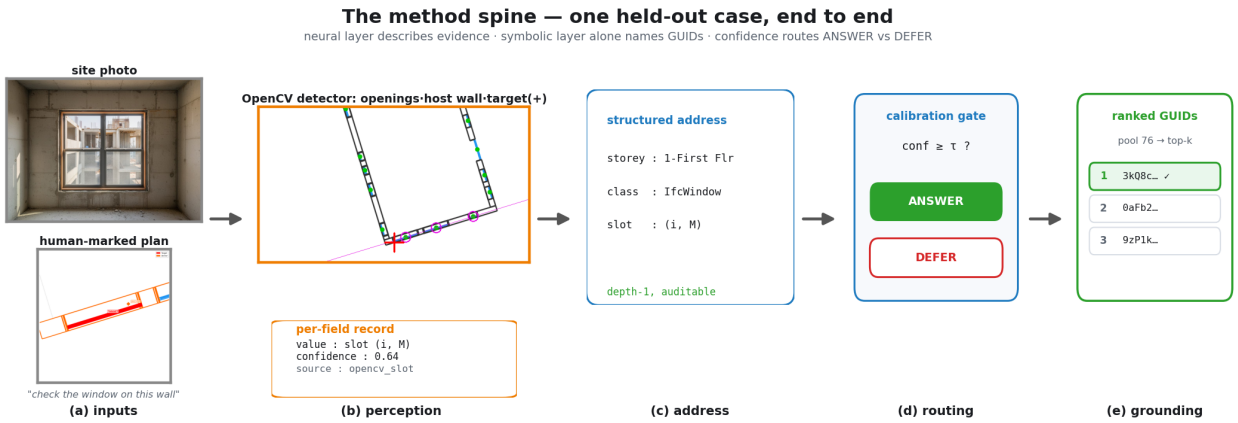

Figure 4: The method spine, on one held-out case (AP_SK_107), with the real artefacts as the centrepiece. *(a)* A site photograph, a human-marked plan (target wall in red, anchor in orange), and a short note enter. *(b)* The OpenCV slot detector's actual overlay reads the openings (green) along the host wall (magenta) and locates the target (red +), emitting a `{value, confidence, source}` record. *(c)* The depth-1 structured spatial-address record is assembled. *(d)* A calibration gate routes confident cases to ANSWER and the rest to DEFER. *(e)* The address is matched against the IFC-graph pool to a ranked GUID shortlist. The neural layer describes evidence; the symbolic layer alone names GUIDs.

guage model), chosen for stable JSON emission and native dynamic resolution. It reads the photo, the note, and the plan crop and emits a structured contract — `{storey_name, ifc_class, spatial_relations:[{predicate, object_type, direction}]}` — and nothing else: *it describes, it never queries* (Figure 7). Fine-tuning on the synthetic corpus is what makes the spatial fields exist at all: spatial-direction accuracy reaches 82% where zero-shot prompting scores ∼0% (this 82% and the G-series ablation below are prior thesis measurements, not re-run on this repo's harness), and the coarse fields (storey, `ifc_class`) saturate at 100% versus 30–63% zero-shot. The G-series ablation also exposes a multi-task capacity trade-off under a fixed adapter rank: under-resourced adapters collapse on spatial extraction while the coarse fields stay saturated, so topology-specific supervision and added position-context capacity are needed to recover spatial fingerprints — and, as the per-field profile in §3.3 shows, the finest discriminating fields do not survive fine-tuning at all, which is why they are delegated to the specialists below.

2.5 Synthetic data pipeline: supervision with zero real labels

Cold-start means no real paired data, so supervision is manufactured from the model itself in three stages. *Stage 1 — IFC structure mining:* IfcOpenShell traversal plus spatial-relationship computation yields, for every element, its ground-truth address and relational context. *Stage 2 — visual synthesis:* Blender renders scenes with deliberate hard negatives — identical-looking adjacent rooms — so the model cannot shortcut on appearance. *Stage 3 — text generation:* an LLM writes site-note queries at varying specificity, and an LLM-as-judge filter removes degenerate cases. The output is 990 training cases on the studied project and a 60-case held-out benchmark (the evaluation set throughout §3).

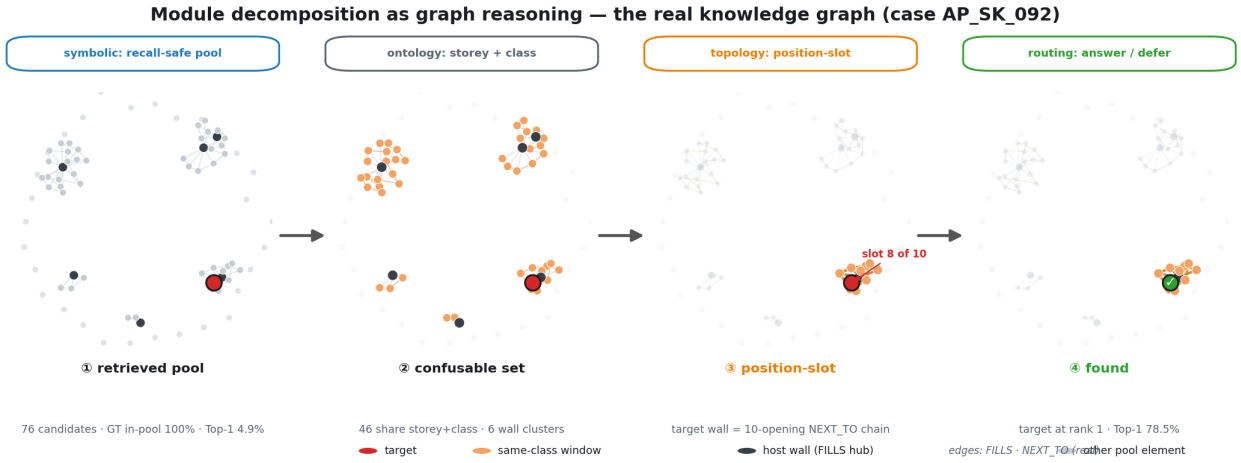

Figure 5: Module decomposition told as graph reasoning on one real held-out case (AP_SK_092, a window). The layout is the *real* knowledge graph — a force-directed layout on the actual `FILLS` (window→host-wall hub) and `NEXT_TO` (consecutive openings) edges, held fixed across all four panels so the same graph is progressively filtered. **(1)** The symbolic layer returns a recall-safe pool of 76 candidates (ground truth in-pool 100%) where Top-1 is only ∼5%. **(2)** 46 of the 76 share the target's storey and IFC class, forming 6 host-wall clusters — ontology cannot separate them, so ranking, not retrieval, is the problem. **(3)** The target's host wall is a 10-opening `NEXT_TO` chain; the position-slot (8 of 10) re-weights the siblings (a soft prior — no candidate removed). **(4)** The target lands at rank 1 (oracle address Top-1 78.5%). The target node (red) is tracked throughout.

2.6 Symbolic layer: deterministic query compilation with a recall-safe cascade

The constraint contract compiles into Cypher through nine predefined query strategies (spatial triplet, continuous span, space+type, . . .); the VLM never generates Cypher (§2.1). Execution follows a cascade that is recall-safe by construction: strict constraints first, then loosen one constraint at a time, with a storey+type safety net guaranteeing a non-empty pool. The pool is then *never* hard-filtered further — the measured reason is in §3.4 — and ranking inside the pool is delegated to the spatial address.

2.7 The spatial address, deterministic specialists, and calibrated routing

The element that closes the loop is the **type-conditional spatial address**: the minimal descriptor set that identifies an element within its confusable set $C(e)$ — the pool elements sharing its storey and IFC class — subject to two deployment constraints: it must be **IFC-computable** (derivable from the model with no labels) and **image-recoverable** (estimable from photo and plan). The address is class-conditional. *Fillers* (windows, doors) are identified by the **position-slot** (i, M): the i-th of M openings ordered along the host wall, numbered in an image-coordinate convention (a representational prerequisite quantified in §3.5). *Walls* are identified by a **connectivity fingerprint** (`connection_degree`, `hosted_opening_count`, `length_band`, `is_external`). Relational context is *compiled into the node* at ingestion — the depth-1 sub-graph and the distilled node attributes carry the same information — so the runtime extractor recovers at most one hop (the depth law, §3.6).

Recovery of the discriminating fields is delegated to **deterministic visual specialists**: an OpenCV detector that color-segments the plan's openings and orders them along the host wall to read the position-slot, and a ResNet size head for dimension bands — both of which outperform

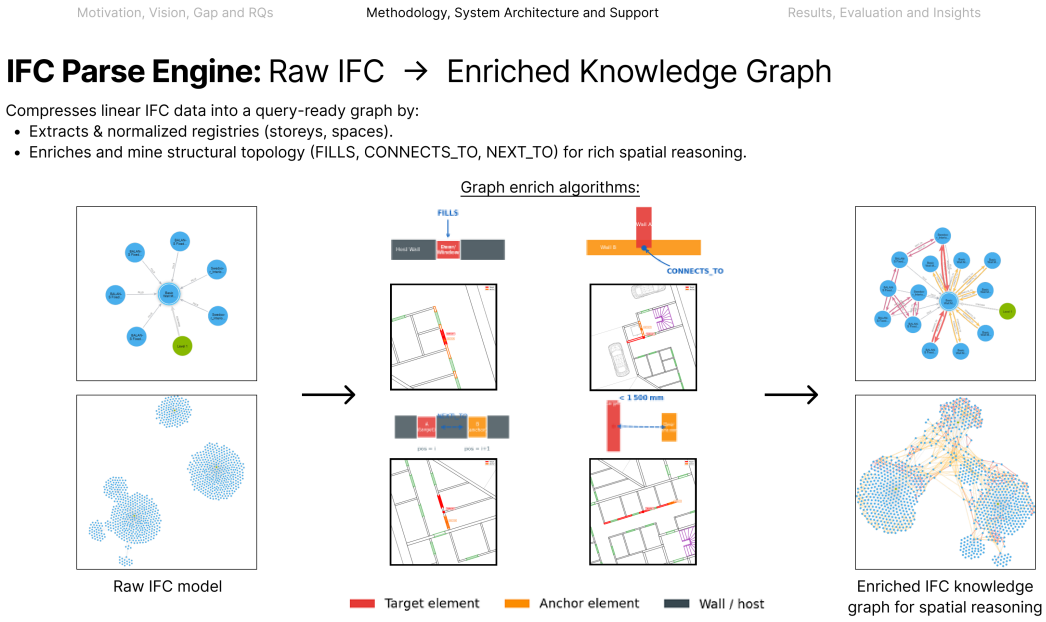

Figure 6: The IFC parse engine: from a raw IFC model (left) to an enriched, query-ready knowledge graph (right). The engine extracts and normalises the registries (storeys, spaces) and then mines structural topology — `FILLS`, `CONNECTS_TO`, `NEXT_TO` — shown by the per-relation enrichment insets (target element red, anchor orange, host/wall dark). The local containment-only graph cannot separate identical siblings; the enriched graph carries the relational structure the spatial address queries.

the trained VLM on exactly these sub-tasks. Every extracted field, from every source, is emitted as a uniform record `{value, confidence, source}` — the contract invariant. A routing policy consumes the contract: the recovered slot enters the rerank as a confidence-weighted prior inside the recall-fixed pool, and a low-confidence extraction routes to **defer** ("here are the candidates") rather than to a confident wrong answer. The premise that confidence may route at all is gated on calibration [7] and measured, not assumed (§3.5).

3 Evaluation & Results

3.1 Setup, metrics, and staging

All numbers are measured on the held-out benchmark of the studied project (Tier-3; $n = 60$ cases / 59 elements, of which 35 are addressable fillers). The split is region-disjoint by construction, but a data audit found **element-level leakage**: 12 of the 59 held-out target elements (9 fillers, 3 walls) also appear as training targets under a different render, because an element can sit in two regions. We disclose this directly and note that it cannot affect the headline results by construction: the oracle diagnostics (§3.2, §3.6) read only the IFC graph and learn nothing, and the realized position-slot result (§3.5) uses a *deterministic* OpenCV detector with no trained weights — both are leakage-proof. The only learned component touched by the split is the per-field VLM profile (§3.3); re-profiling on the leakage-clean 48-case subset leaves it unchanged (storey/class 100%,

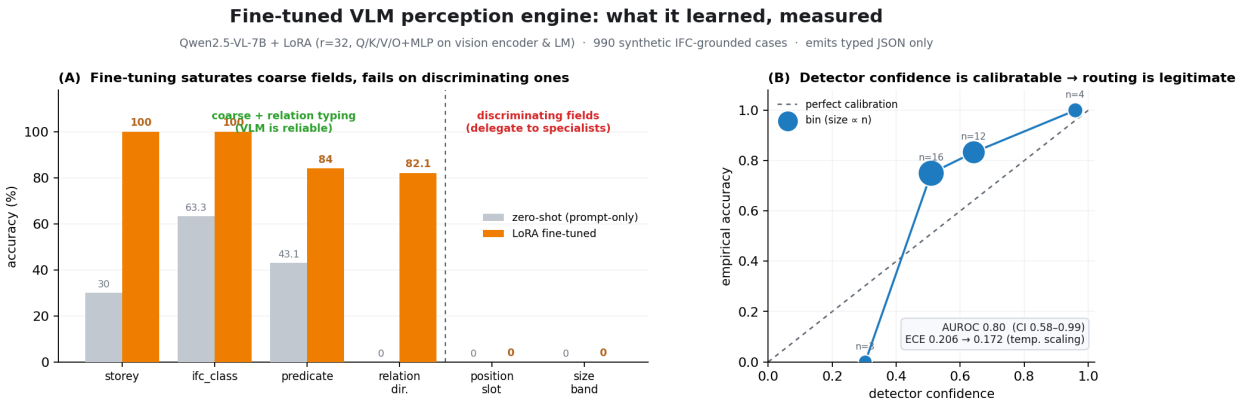

Figure 7: The fine-tuned VLM perception engine, measured. *(A)* Per-field accuracy, ordered coarse→discriminating, zero-shot prompt-only versus LoRA fine-tuned: fine-tuning saturates the coarse prefix (storey, `ifc_class` from 30/63% to 100%) and learns relation typing (direction ∼0% → 82%), but the *discriminating* address fields (position-slot, size band) stay at 0% even fine-tuned — the evidence that motivates delegating them to deterministic specialists. *(B)* The realized detector's confidence is calibratable (reliability diagram; AUROC 0.80, ECE 0.206→0.172 after temperature scaling), which is what makes confidence-routing legitimate. (No training-step loss curve is shown: per-field gain, held-out generalisation, and calibration are the meaningful evidence that a structured-extraction adapter has learned.)

position-slot/size 0%), so the architectural conclusion does not depend on the leaked cases. We structure the evaluation in stages: first, an *oracle* experiment establishing the symbolic-layer ceiling under perfect extraction (§3.2); second, neural extraction accuracy per field (§3.3); third, how noisy signals combine without destroying recall (§3.4); fourth, the realized extractor with calibrated, selectively-predictive routing (§3.5); fifth, the depth analysis (§3.6); and last, failure decomposition and workflow-facing cost (§3.7). This staging separates symbolic retrieval quality from neural extraction quality — the two failure modes require different fixes, and conflating them produces misleading aggregate metrics. We report **GT-in-Pool** as the primary retrieval metric rather than Top-1, because Top-1 conflates retrieval failure (element never in the pool) with ranking failure (element in the pool but ranked poorly); guaranteed retrievability is the architectural contribution, and ranking is where the spatial address acts. Table 1 consolidates the headline numbers; the remainder of this section derives and explains each block.

3.2 Oracle experiment: the symbolic ceiling, and where discrimination lives

Under perfect extraction the symbolic layer retains the target in **100% of cases** (GT-in-Pool) and compresses the live pool along the fingerprint ladder — from 1,233 elements (whole-model) to 46 (storey+class) toward single digits as spatial constraints are added. The architecture is therefore provably sound; the question is where, inside the pool, the discriminating information lives.

The coarse floor saturates. Restricting to the target's storey and IFC class leaves a median confusable set of 46 (mean 112) and an oracle Top-1 of only **4.9%** — *below* the realized end-to-end reranker (6.7%) and equal to it on Top-10 (31.5% vs. 30.0%). Ontology is necessary but cannot separate same-class siblings: it is the floor, not the discriminator. Adding the richest attribute (the family/type string `object_type`) shrinks the confusable set to a median of 13 (3.8×) and lifts

Method	GT-in-Pool	Top-1	Top-10	MRR@10
External baselines, full held-out pool ($n = 60$)				
Lexical (BM25)	100	1.7	15.6	0.058
Dense (MiniLM)	100	1.7	25.0	0.108
Zero-shot VLM — full pool	100	3.3	15.0	0.086
Zero-shot VLM — sibling shortlist	100	3.3	20.0	0.097
Our pipeline, realized end-to-end ($n = 60$)				
G8 extraction → deterministic retrieval	∼100	6.7	30.0	0.110
Oracle ceiling, same pools ($n = 60$)				
Coarse: storey + class	100	4.9	31.5	0.137
+ `object_type`	100	—	76.0	—
+ structured spatial address	100	**78.5**	**98.1**	**0.854**
Realized address, addressable fillers ($n = 35$)				
Modal-prior slot (floor)	100	6.6	—	—
OpenCV slot, *extracted* coarse (end-to-end)	100	$\mathbf{58.9}_{[43.2,74.6]}$	67.1	—
OpenCV slot, *oracle* coarse (upper bound)	100	$67.6_{[53.6,81.6]}$	80.9	—

Table 1: Consolidated held-out results. Every method ranks the *same* retrieved pools with the ground truth in-pool throughout (GT-in-Pool, %), so the table isolates *ranking*, not retrieval. Top-1/Top-10/MRR in %. External text retrieval and a zero-shot Qwen2.5-VL-7B reranker (base model, no adapter) stay at single-digit Top-1 — the VLM at chance — while the oracle structured address reorders the identical pool to 78.5%. The realized filler block shows how much survives a real detector: one deterministic specialist lifts Top-1 from the 6.6% floor to **58.9%** end-to-end (with the coarse fields *also* extracted; 67.6% if storey/class are supplied at oracle level, the upper-bound row), bootstrap 95% CI in subscript; deferring the least-confident ∼20% further lifts the answered subset to 73.4% (80.6% in the oracle-coarse variant) (§3.5). Dashes mark quantities not measured for that row.

oracle Top-10 to 76%, but uniquely identifies only 2 of 60 targets. What remains is the irreducibly *relational* residue.

The type-conditional address closes it. Supplied at an oracle level, the class-conditional address of §2.7 takes the real retrieved pool from a coarse Top-1 of 4.9% to **78.5%**, and Top-10 from 31.5% to **98.1%** (MRR 0.137 → 0.854), at zero recall cost. The gain decomposes exactly along the type-conditional split — fillers reach Top-1 **91.0%** (position-slot), walls **64.2%** (connectivity fingerprint; within same-storey walls the fingerprint collapses a median confusable set of 110 → 2, uniquely identifying 10 of 22 walls where `object_type` identifies none). No uniform descriptor achieves this: the position-slot is meaningless for a wall, the fingerprint degenerate for a window. Generic retrieval baselines confirm the gap is not an artifact of our own pipeline. Lexical and dense retrievers plateau at Top-1 1.7% (Top-10 15.6% / 25.0%) on the same pools, and a *zero-shot* Qwen2.5-VL-7B reranker — the same backbone our extractor fine-tunes, given the marked site photo, the marked floorplan, and the candidate pool — stays at chance: Top-1 3.3% over the full pool and 2.9% on fillers even when the pool is pre-filtered to the same-storey, same-class confusable set (per-case chance 1.8–2.4%; candidates are shuffled so that copying the list order scores at chance, which is what the model does). On those same pools the oracle spatial address reaches 78.5% (Figure 8). A strong general VLM cannot recover the relational slot from the image alone; the structured address, not the backbone, is what discriminates identical siblings.

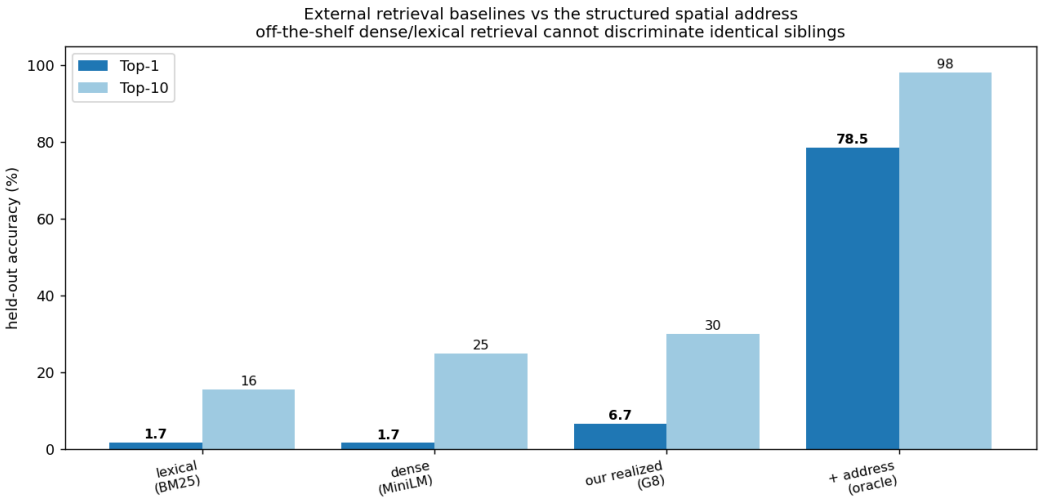

Figure 8: Baselines against the oracle ceiling on identical pools. Lexical and dense text retrieval, a zero-shot Qwen2.5-VL-7B reranker (base model, no adapter; both the full pool and the same-storey/class sibling shortlist), and the deployed G8 pipeline all plateau in single-digit Top-1 — the zero-shot VLM at chance — with the ground truth in-pool throughout; the oracle spatial address reorders the same pool to 78.5%/98.1%. The bottleneck is ranking inside the pool, and the missing signal is the relational address.

3.3 Neural extraction quality: what the fine-tuned VLM does and does not recover

Per-field evaluation shows fine-tuning works where it was aimed: against zero-shot prompting (`ifc_class` 63.3%, storey 30.0%), the fine-tuned models saturate both at **100%**, reach predicate slot-accuracy 82.8–86.2% versus 43.1%, and lift relation direction to 82.1% versus ∼0%. End-to-end, however, the best configuration delivers Top-10 30.0% / Top-1 6.7% / MRR@10 0.110 — with the target in the pool ∼100% of the time. The decisive diagnostic is the held-out per-field re-profile (Figure 9): the VLM is saturated on the *coarse prefix* (storey, class — exactly the fields the oracle shows are non-discriminating) and recovers the *discriminating* structured fields at **0%** (the position-slot and the size band), and emits a relation-direction field in 57% of held-out cases (this 57% is an *emission rate* — whether any direction is produced — not the 82% direction *accuracy* reported above on the thesis evaluation set; the two numbers measure different quantities and are not comparable). The conclusion is architectural, not "fine-tune harder": delegate the slot and size to deterministic visual specialists, keep the VLM on the coarse prefix and relation typing where it is reliable. This is the neuro-symbolic interface in practice — learn where the network is reliable, compute deterministically where it is not [10].

3.4 Combining signals: union, never intersection

How should a mature signal (storey+type) combine with an immature one (spatial topology)? The strategy ablation is unambiguous: the **union** of constraint sets is the only strategy that preserves full GT-in-Pool while still gaining the spatial constraints' compression and ranking benefit; hard intersection over-prunes under noisy extraction, and spatial-only is too brittle. The measurement says why: treating each extracted field as a hard filter multiplies per-field recalls. On this model's

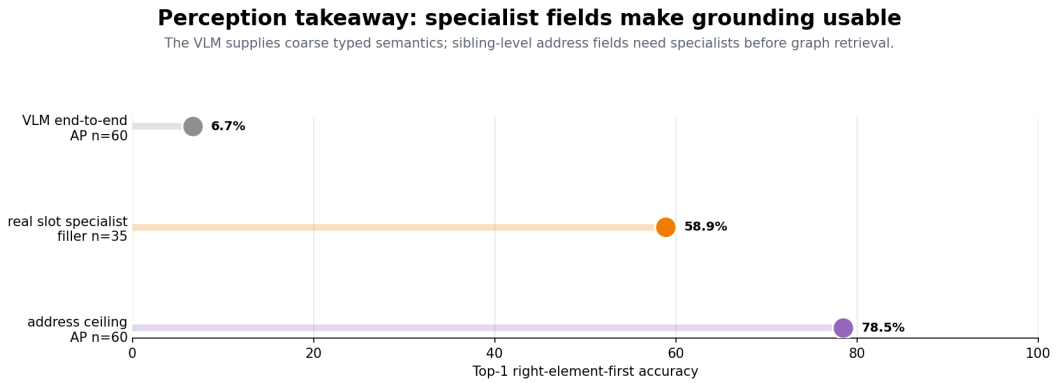

Figure 9: Perception takeaway: per-field re-profile of the fine-tuned VLM on the held-out set. The coarse prefix (storey, class) saturates; the discriminating fields (position-slot, size) are recovered at 0% — motivating delegation to deterministic specialists.

held-out extraction the coarse fields are reliable (storey 100% under leading-storey normalisation, 51.7% under exact string match; `ifc_class` 81.7%), but the discriminating fields are not (position-slot and size recovered at 0% by the VLM, §3.3), so a hard conjunction that *includes* the immature fields drives joint recall toward the product of the weak terms — each unreliable field the filter touches evicts the ground truth more often than it removes a distractor. The resolution is to relocate determinism: **the executor is deterministic given the structured record; the ranking is a calibrated soft prior inside a recall-fixed pool.** The address never removes a candidate; it re-weights one. Recall stays at 100% by construction, and the unreliable signal is spent only on ordering. As a design principle this generalizes: when combining mature and immature signals, union — intersection amplifies extraction error and destroys recall.

3.5 Realizing the address: one specialist, calibrated, selectively predictive

The oracle ceiling is a statement about information; this section measures how much survives a real, imperfect detector. The honest floor is stark: the fine-tuned reranker emits position as free text and recovers a usable slot in **0 of 35** held-out fillers; even a modal-prior slot guess reaches only 6.6%. We report four results.

(1) One realizable extractor closes most of the gap. The OpenCV position-slot detector recovers the opening count exactly in 83% of cases (29/35) and the full slot in 74% (joint (i, M)). Fed into the soft rerank with the coarse fields *also* taken from the model's actual extraction (storey and `ifc_class`, not oracle), it lifts filler Top-1 from the 6.6% floor to **58.9%** (bootstrap 95% CI $[43.2, 74.6]$, $n = 35$) and Top-10 to 67.1%, at zero recall cost — a genuinely end-to-end number. Supplying the coarse fields at oracle level instead (storey extracts at 100% under leading-storey normalisation, `ifc_class` at 97% on fillers) raises this to **67.6%** (CI $[53.6, 81.6]$) / Top-10 80.9%, an upper bound on what a perfect coarse extractor would add. The residual error concentrates in the ordering index, not in the perception of the openings. The wide interval is a direct consequence of the small addressable set and is reported, not smoothed over. The detector's three pixel thresholds are not a knife-edge fit: sweeping each ±25% leaves realized Top-1 in $[55.1, 83.1]$ (never near the 6.6% floor), two of the three are inert, and only the same-wall perpendicular band matters — and the deployed value is mid-range, not tuned to the peak.

(2) A representational prerequisite: the address must be image-recoverable. The ground-truth slot in the source data is numbered along each wall's IFC local-X axis — a modelling artefact invisible in any image. Scored against that convention, detector and ground truth disagree on 16 of 35 fillers, *all* exact mirrors $i \mapsto M-1-i$; under an image-coordinate convention the disagreement vanishes and joint accuracy rises from an apparent 34% to the true 74%. An address field is realizable only if it is defined in a frame the evidence carries — a substantive design constraint, not bookkeeping. The same constraint explains a negative result: the wall fingerprint's load-bearing fields depend on IFC wall-*instance* boundaries that collinear instances erase in the floorplan poché (length-band recovery 5/17, realized Top-1 ≈ floor), so the wall address is oracle-discriminative but *not* image-realizable, and realization is scoped to the filler slot.

(3) The confidence is calibratable — and reweighting is a measured no-op. Routing is only legitimate if confidence tracks correctness [7]. On the held-out fillers the raw detector confidence is positively discriminative (AUROC **0.80**, bootstrap 95% CI $[0.58, 0.99]$) and moderately mis-calibrated (ECE 0.206, 95% CI $[0.09, 0.32]$); temperature scaling reduces it to 0.172 (Figure 10). The intervals are wide — at $n = 35$ the calibration evidence is suggestive, not conclusive, and we treat the gate as a measured operating point rather than a precise estimate. But the calibrated weight does *not* pay off as a rerank weight: because the slot is the finest tiebreaker inside a storey×class bucket, any strictly-positive weight induces the identical ordering — hard match, raw weight, and calibrated weight all yield the same Top-1. We report this no-op rather than bury it.

(4) The payoff is selective prediction. The calibrated confidence pays off as a deferral threshold: sweeping it traces a coverage–accuracy curve (Figure 11), and deferring the least-confident ∼20% of cases lifts answered-subset Top-1 from 58.9% to **73.4%** end-to-end (from 67.6% to 80.6% in the oracle-coarse variant) [6]. A worked case (Figure 12) shows the mechanism: the detector predicts slot 1 of 10 where the truth is 8 of 10 — an error — but its calibrated confidence (0.05) is below threshold, so the case routes to *defer* and surfaces the candidates instead of a confident wrong GUID. For a triage tool with downstream consequence, this is the correct behavior, and the coverage–accuracy curve, not a point estimate, is the honest report.

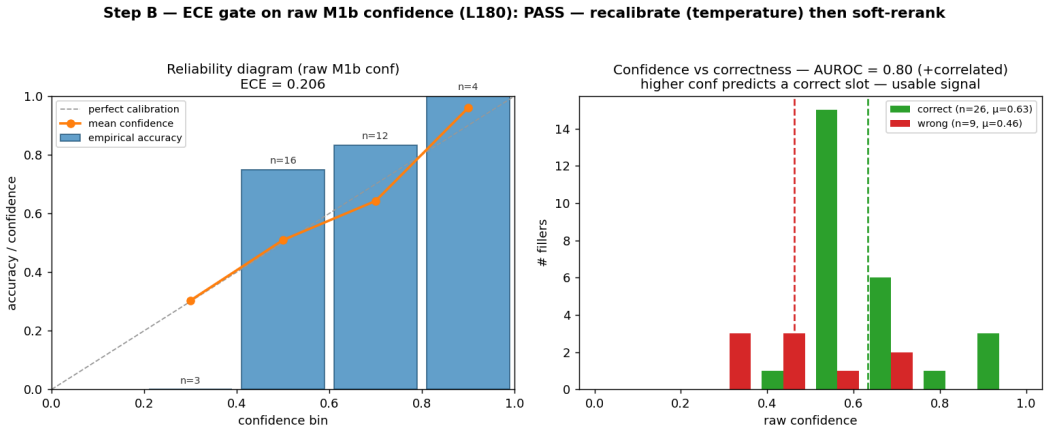

Figure 10: Calibration gate. Raw detector confidence is positively discriminative (AUROC 0.80) and moderately mis-calibrated (ECE 0.206); temperature scaling reduces ECE to 0.172.

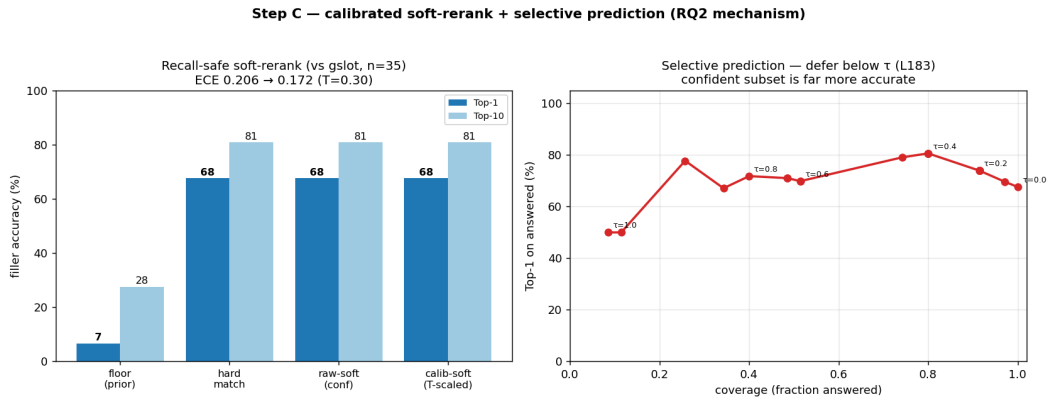

Figure 11: Selective prediction (oracle-coarse variant shown). Sweeping the deferral threshold traces a coverage–accuracy curve; deferring the least-confident ∼20% lifts answered-set Top-1 from 67.6% to 80.6% (the end-to-end variant with extracted coarse fields moves 58.9%→73.4%).

3.6 The depth law: relational context saturates at one hop

Where should relational context live — extracted deep at inference, or compiled into the node? Under an oracle the confusable set is already a singleton at one hop (median $|C|$ 13 → 1; Table 2, Figure 13), so the graph carries enough discrimination at depth 1 and deeper hops add nothing even in principle. The realizability question is therefore whether that one-hop discrimination can be read from an image, and the answer decomposes the depth law into two *measured* facts. First, **recovery does not collapse with depth**: on the held-out set the deployed model extracts the depth-k relation (predicate + object type + direction) correctly in 96.7% / 93.9% / 69.6% of cases at $k = 1/2/3$ — reading deep relation *types* is not the bottleneck. Second, **what is readable is not discriminative**: those predicates are type-homogeneous — all 389 `FILLS` are window/door→wall and every `CONNECTS_TO` is wall↔wall — so a correctly-extracted hop-2 "`CONNECTS_TO` a wall" matches every wall in the pool and shrinks $|C|$ by nothing. The discrimination the oracle exploits is the *identity* of the specific neighbour, which is not image-recoverable; the one exception we realize is the filler position-slot (74% joint, §3.5). The depth law is thus *informational*, not a reliability cascade: deeper relational context is either non-discriminative (type-homogeneous) or non-recoverable (named-neighbour identity), which is the structural form of the error-compounding documented for multi-hop reasoning over knowledge graphs [2]. The prescription is unchanged — compile depth into the node and extract at depth ≤ 1 — and the training side corroborates it: under a fixed LoRA rank, supervising depth-≥ 2 chains cost coarse `ifc_class` capacity for ≈0 realizable gain (prior thesis measurement).

3.7 Failure decomposition, latency, and triage effort

Because the pipeline localizes errors to named fields rather than opaque scores, the residual end-to-end loss decomposes cleanly: it is dominated by *extraction* errors — missing or incorrect spatial predicates and the lowest-accuracy fields (subtype, material, position-context) — not by retrieval logic. Every failure has a traceable stage, which is the operational differentiator over black-box matching: the roadmap is "improve the named weak fields," not "redesign the planner." Latency favors the constrained pipeline (∼1 s vs. ∼4.5 s for the ReAct agent, with exact repeatability). The workflow-facing cost metric makes the ceiling concrete (Figure 14): the expected number of candidates a coordinator must inspect falls from a median 38 under manual scan to 0.5 under a

relational depth	oracle median $\lvert C\rvert$	G8 type-recovery
attributes only (depth 0)	13	—
+ 1 hop	**1**	96.7%
+ 2 hops	1	93.9%
+ 3 hops	1	69.6%

Table 2: The depth law, from measured quantities (held-out set). The oracle confusable set collapses to a singleton at one hop, so depth ≥ 2 adds no discrimination even in principle. The deployed model *recovers* the depth-k relation type reliably (right column: exact predicate + object type + direction match), yet realized discrimination still saturates at one hop — the recoverable relation types are homogeneous (non-discriminative), and the discriminating signal (neighbour identity / position-slot) is image-recoverable only for fillers.

perfect address — with 15 seconds per inspection an explicit proxy assumption, not a user-study result.

4 Discussion

4.1 Workflow impact: a before/after walkthrough

The system's value is best seen as a concrete walkthrough rather than a claim.

Before (manual workflow). A site worker photographs a cracked column near a stairwell on floor 12 and writes a chat message: "crack near stairs, floor 12, block A." The coordinator reads it at end of day, goes on a site walk, identifies the element by sight, and manually types the IFC GUID into the issue tracker. Average lag: 1–3 days. If the coordinator leaves the project, the knowledge leaves with them. The element lands in a flat spreadsheet, not linked to the BIM record, and downstream compliance requires manual re-entry.

After (system-assisted). The worker sends the same photo and message. The system returns a compact ranked shortlist within ∼1 second — the pool compressed from ∼1,200 elements to a median of a few dozen, with the correct element in the pool ∼100% of the time and, on the addressable subset, at rank 1 in 58.9% of cases end-to-end (73.4% when the system is allowed to defer the least-confident fifth). The worker taps the correct element; a BCF package is generated and the element is linked to the BIM record. The coordinator's role shifts from data entry to verifying and escalating a shortlist, and no longer requires being on-site to resolve the link.

The claim is deliberately bounded: this does not replace the coordinator. It eliminates the manual data-entry loop, surfaces the correct element inside a short ranked shortlist, abstains explicitly when unsure, and improves over the project lifecycle as confirmed GUIDs accumulate as supervision.

4.2 Limitations

We state the boundaries directly. *Synthetic-to-real gap (central limitation):* all training and evaluation data are synthetic; the numbers are optimistic until validated on real site photographs, and a small pilot (50–100 labelled real cases from a live project) is the single most informative next measurement. *Single project:* the ceilings and reliabilities are measured on one synthetic

model; the form of the contribution (type-conditional, shallow, image-recoverable, calibrated-soft) is general, the specific numbers are not. *Statistical power:* the calibration and coverage–accuracy estimates rest on 35 addressable fillers; we report the operating point (defer $\sim$20% $\rightarrow$ 73.4%), not a precise optimal threshold. *Class coverage:* the realized address covers fillers; the wall fingerprint is demonstrated at oracle level and is blocked from realization by the image-recoverability constraint itself (§3.5); "other"-class and room-relative addressing remain open. *Multi-hop:* joint multi-hop extraction collapses under error cascade; practical use is single-hop plus human review, by design. *IFC dependency:* the system retrieves against the model as given; if the IFC is outdated, retrieval is against stale ground truth — mitigated by versioning the graph and surfacing low-confidence results for manual check.

4.3 Future work, prioritized

1. **Real project traces.** Validate on live coordination data; the synthetic-to-real shift is the most credible threat to the numbers and the first thing to measure.
2. **Remaining address classes.** Size-band realization (the ResNet head), relation-direction improvement via subtype-contrastive augmentation, and room/space addressing where the export carries space boundaries.
3. **Learned spatial topology.** Replace the distance-threshold `ADJACENT_TO` heuristic with a learned spatial graph model to generalize beyond standard geometries.
4. **Write-back integration.** BCF and BIM-cloud integration, so the grounded GUID lands in a live workflow rather than a report; the human feedback loop then converts confirmed GUIDs into supervised updates.
5. **4D temporal context.** Schedules, timestamps, and progress updates, making the IFC graph a dynamic project database rather than a static snapshot.

The long-term goal is not to replace construction managers but to remove the manual translation burden between physical on-site evidence and the digital record — supporting timely, traceable, and continuously updated project analytics.

5 Conclusion

We asked how unstructured, egocentric site evidence can be grounded to a unique element in an allocentric BIM model, in a cold-start regime with no real labels. The answer is a neuro-symbolic decomposition organized around an **ontology-grounded, topology-derived spatial address**. The oracle analysis proves the architecture sound — the symbolic layer retains the ground truth in 100% of cases and, supplied with the type-conditional address, reorders the pool from Top-1 4.9% to 78.5% — and isolates extraction, not retrieval design, as the bottleneck. The realization study shows how much of that ceiling survives a real detector and how to consume a noisy address: as a calibrated soft prior inside a recall-fixed pool (union, never intersection), with one deterministic specialist lifting the addressable subset from 6.6% to 58.9% end-to-end and selective prediction raising the answered subset to 73.4% — the system knows when to abstain. The depth analysis supplies the architectural principle both rely on: realizable relational discrimination saturates at one hop, so depth is compiled into the node and learning is placed at the neural$\rightarrow$symbolic interface.

Three findings generalize beyond this system. An address field is deployable only if it is defined in a frame the evidence carries — image-recoverability is a design constraint, not an afterthought. When

a mature signal combines with an immature one, union preserves recall and intersection destroys it. And calibration's payoff here is not reweighting (a measured no-op) but *selective prediction*: the honest deliverable of a grounding system with downstream consequence is a coverage–accuracy curve and an explicit defer route. Within the regime we can measure, the structured spatial address is what makes site-to-BIM grounding realizable, auditable, and transferable; validating it on real project traces is the immediate next step.

References

[1] buildingSMART International. *Industry Foundation Classes (IFC) 4.3.2.0 Specification*. buildingSMART International, 2024. URL https://ifc43-docs.standards.buildingsmart.org/. IFC4x3 HTML documentation; accessed 2026-06-11.

[2] Abir Chakraborty. Multi-hop question answering over knowledge graphs using large language models, 2024. URL https://arxiv.org/abs/2404.19234.

[3] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Danny Driess, Pete Florence, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. SpatialVLM: Endowing vision-language models with spatial reasoning capabilities, 2024. URL https://arxiv.org/abs/2401.12168.

[4] Changyu Du, Sebastian Esser, Stavros Nousias, and André Borrmann. Text2BIM: Generating building models using a large language model-based multi-agent framework, 2024. URL https://arxiv.org/abs/2408.08054.

[5] Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. From local to global: A graph RAG approach to query-focused summarization, 2024. URL https://arxiv.org/abs/2404.16130.

[6] Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. doi: 10.48550/arXiv.1705.08500. URL https://arxiv.org/abs/1705.08500.

[7] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017. doi: 10.48550/arXiv.1706.04599. URL https://arxiv.org/abs/1706.04599.

[8] Anton Gusarov, Anastasia Volkova, Valentin Khrulkov, Andrey Kuznetsov, Evgenii Maslov, and Ivan Oseledets. Multi-agent GraphRAG: A text-to-Cypher framework for labeled property graphs, 2025. URL https://arxiv.org/abs/2511.08274.

[9] Sima Iranmanesh, Hadeel Saadany, and Edlira Vakaj. LLM-assisted graph-RAG information extraction from IFC data, 2025. URL https://arxiv.org/abs/2504.16813.

[10] Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B. Tenenbaum, and Jiajun Wu. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. In *International Conference on Learning Representations (ICLR)*, 2019. doi: 10.48550/arXiv.1904.12584. URL https://arxiv.org/abs/1904.12584.

[11] Bharathi Kannan Nithyanantham, Tobias Sesterhenn, Ashwin Nedungadi, Sergio Peral Garijo, Janis Zenkner, Christian Bartelt, and Stefan Lüdtke. MCP4IFC: IFC-based building design using large language models, 2025. URL https://arxiv.org/abs/2511.05533.

[12] Richard S. Sutton. The bitter lesson. http://www.incompleteideas.net/IncIdeas/BitterLesson.html, 2019. Online essay; accessed 2026-06-11.

[13] Aman Tiwari, Shiva Krishna Reddy Malay, Vikas Yadav, Masoud Hashemi, and Sathwik Tejaswi Madhusudhan. Auto-Cypher: Improving LLMs on Cypher generation via LLM-supervised generation-verification framework, 2024. URL https://arxiv.org/abs/2412.12612.

[14] Zuoxu Wang, Zhijie Yan, Shufei Li, and Jihong Liu. VLM-based scene graph generation for industrial spatial intelligence, 2024. URL https://ssrn.com/abstract=4926945. Preprint, submitted to Elsevier.

[15] Junxiang Zhu, Peng Wu, and Xiang Lei. IFC-graph for facilitating building information access and query. *Automation in Construction*, 148:104778, 2023. ISSN 0926-5805. doi: 10.1016/j.autcon.2023.104778.

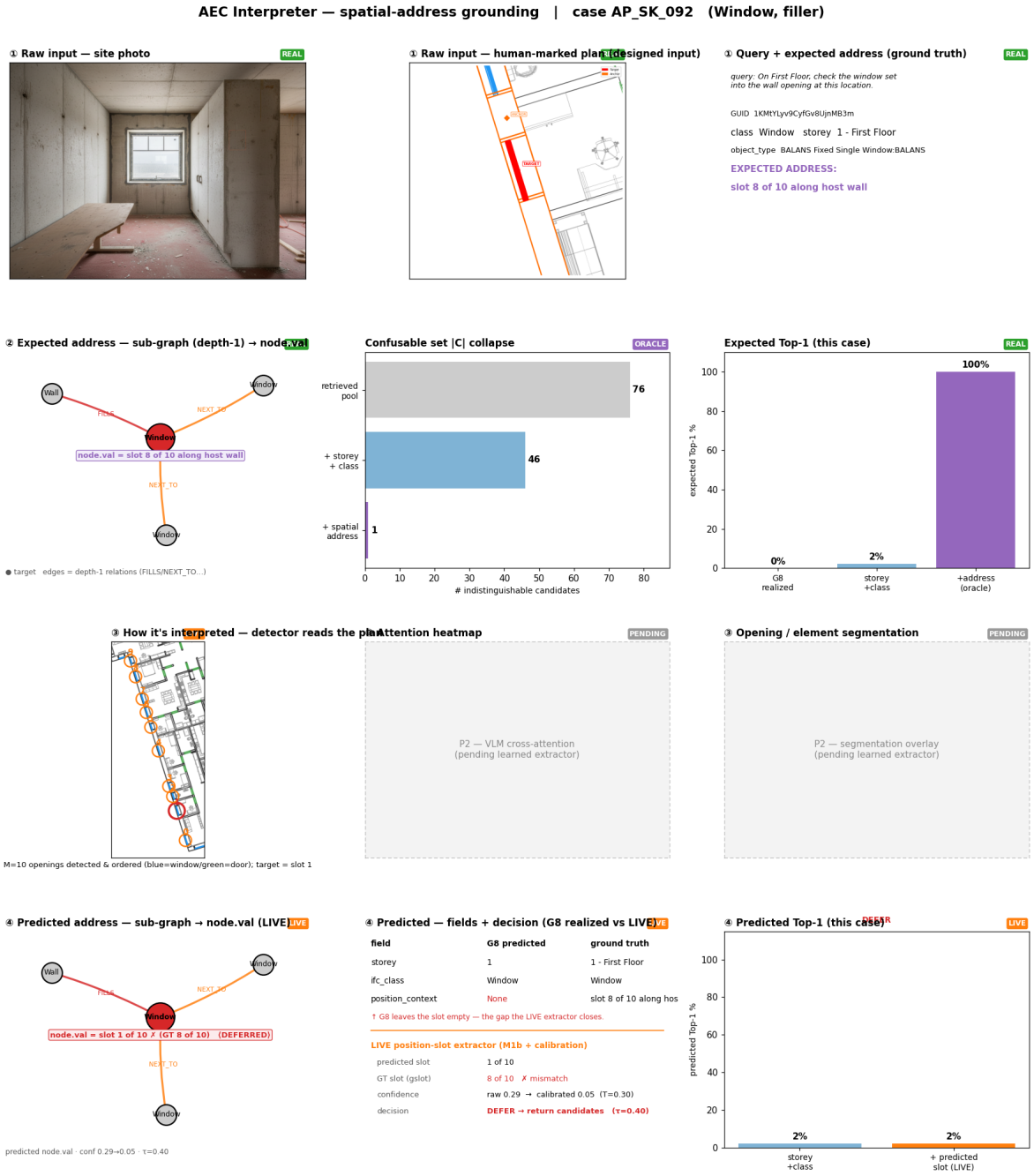

Figure 12: A worked DEFER case. The detector predicts slot 1 of 10 where the truth is 8 of 10, but its calibrated confidence (0.05) is below threshold: the case routes to defer and surfaces the candidate pool rather than emitting a confident wrong GUID.

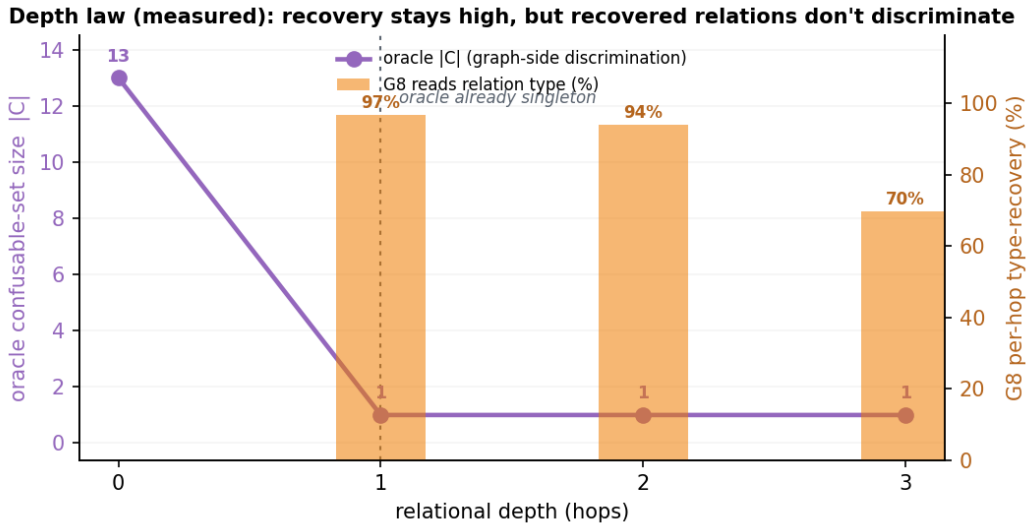

Figure 13: Information versus realizability. The oracle confusable set is a singleton by one hop, so deeper hops add nothing in principle; the model reads deep relation *types* reliably yet realized discrimination saturates at one hop because those types are homogeneous — the depth law is informational, not an extraction-reliability cascade.

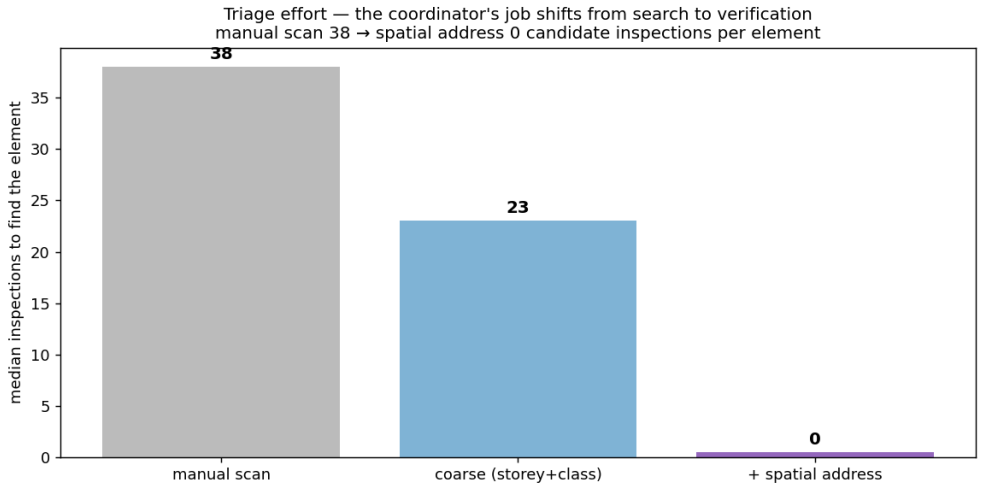

Figure 14: Triage-effort proxy on the same held-out pools. Expected inspections per case fall from a median 38 (manual scan) to 0.5 (perfect address); the realized specialist sits between, with success@1 78.5% under the oracle address.