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

Chia Hui Yen Joshua Bard
Computational Design Laboratory, Carnegie Mellon University
huiyenc@alumni.cmu.edu

2026

Abstract

Architecture, Engineering, and Construction (AEC) projects hinge on coordination across many stakeholders and heterogeneous data formats. Despite open digital standards (OpenBIM/IFC), day-to-day coordination between physical and digital models still depends on manual interpretive labour: practitioners translate unstructured site evidence (photographs, chat notes, and marked floorplan patches) into model-linked, schema-compliant records. This manual translation is error-prone, causing information loss across handoffs, ambiguous evidence-to-element links, delayed update of site information, and regulatory non-compliance. We propose a neuro-symbolic **interpreter middleware** that maps on-site evidence to the digital model automatically by separating probabilistic perception from deterministic graph retrieval. We target the **spatial grounding** problem at its core: linking a piece of multimodal site evidence to the single correct IFC element, by GUID, amidst extreme geometric repetition, where dozens of near-identical elements on a floor share every queryable attribute and only relational spatial position separates them. The system uses a fine-tuned VLM to emit typed spatial constraints, an IFC parse engine to build a normalized and topology-enriched project graph, and a symbolic query cascade that returns auditable candidate pools instead of hallucinated element IDs. This project-level grounding layer is the core contribution; a type-conditional spatial address is an enhancement module for the repeated-element ranking gap. On the held-out synthetic benchmark, the interpreter retains the target in the candidate pool throughout, while learned extraction alone reaches Top-1 6.7% because many elements remain visually and semantically identical. The enhanced spatial address reorders the same pool from Top-1 4.9% to **78.5%** at the oracle ceiling; this ceiling is realizable only in part, since the recoverable signal is the filler position-slot, where one OpenCV detector lifts the addressable filler subset ($n = 35$) from a 6.6% floor to **58.9%** end-to-end (21.2% on the realistic cluttered-floor cases) and supports selective prediction (AUROC 0.80, 95% CI $[0.58, 0.99]$ at $n = 35$; deferring the least-confident fifth raises answered-set Top-1 to **73.4%**). The outcome is an auditable site-to-BIM interpreter: neural models parse evidence, IFC topology constrains retrieval, and calibrated confidence decides when the system should defer to a human reviewer.

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

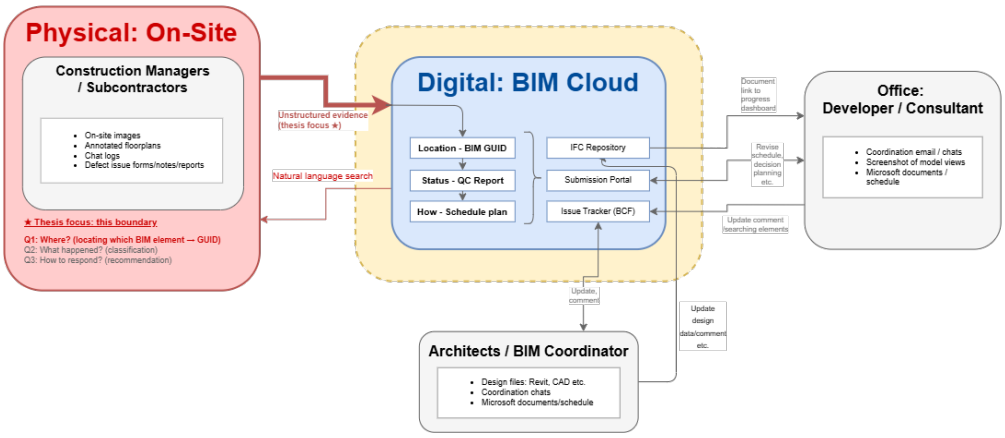

Figure 1: The traceability gap and the interpreter-middleware vision. Site evidence is unstructured and egocentric; the BIM/IFC record is structured and allocentric. A coordinator translates between them by hand, introducing semantic drift and context decay. The interpreter middleware inserts a machine layer that answers the field questions (*where?*, then *what?*, *how?*), anchoring every observation to a regulator-grade IFC GUID. This work targets the foundational *where*.

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?*

Why the grounding problem is genuinely hard is best seen at the element level. Industrialized (prefabricated, volumetric) construction produces extreme geometric repetition: a single storey can hold dozens of windows that share every queryable attribute, so a pure attribute filter is at chance. With $\sim$46 identical windows on a floor, attribute-only matching is $\approx$2.2% (1 of 46); the only signal that separates two siblings is their *position in the building's relational topology* (Figure 2). This "attribute-entropy deadlock" is the empirical reason the system is neuro-symbolic rather than a single learned matcher.

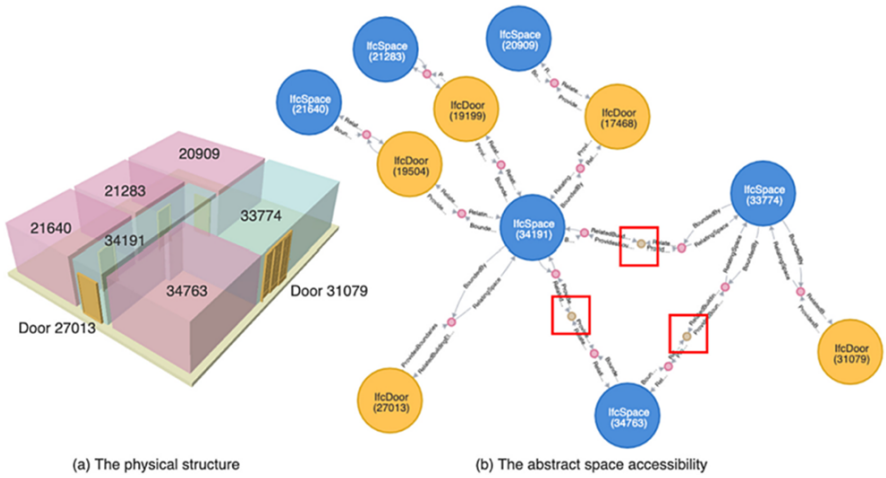

Figure 2: Why topology, not attributes. (a) A physical building module and (b) its abstract space/opening graph. Visually and semantically identical elements are separated only by their relational position (highlighted), which attribute filtering cannot express. Abstract graph representation after Zhu et al. [15].

The setting is 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. On site, attaching an observation to the right element requires more than keying its type: labelling it a "window" or "door" returns a pool of identical elements, so the task reduces to **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 reports the project as an **interpreter middleware**: a neuro-symbolic layer that turns mixed site evidence into typed constraints, grounds those constraints against an enriched IFC graph, and returns traceable GUID candidates for human verification. Ontology supplies the execution boundary that prevents hallucinated element IDs. Topology supplies the relational structure needed to separate visually identical siblings. The spatial address and calibration results are an enhancement inside this larger system, not the whole system. The contribution is developed through three questions:

- **RQ1: Multimodal grounding under ambiguity (§3.3).** *How can heterogeneous multimodal site evidence (a site photo, a marked floorplan crop, and a free-text note) be fused and reliably grounded into structured IFC search constraints despite geometric repetition and noisy input?* The neural layer is treated as a parser, not a GUID generator: it jointly reads the visual and textual channels and emits typed fields with confidence and provenance. Fine-tuning saturates the coarse IFC fields (storey and class) and improves spatial-relation extraction, while exposing which fields require deterministic support.

- **RQ2: Ontology-driven, hallucination-resistant retrieval (§3.2, §3.4).** *How can IFC ontology and topology turn probabilistic neural outputs into deterministic, auditable retrieval that cannot return an element absent from the model?* The IFC parse engine builds a project graph, the planner compiles constraints into a fixed Cypher cascade over only the model's real GUIDs, and the retrieval backend preserves the target through recall-safe union rather than brittle intersection.
- **RQ3: Spatial-address enhancement (§3.5, §3.2.1).** *What additional topology-derived signal resolves the repeated-element ambiguity left by the base interpreter?* A type-conditional spatial address reorders the same recall-safe pool from Top-1 4.9% to **78.5%** at oracle level; one realized position-slot detector lifts the addressable filler subset to **58.9%** end-to-end and supports selective prediction [6, 7].

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, or 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 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. Our work occupies a quadrant that these two lines leave open: it couples 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 middleware with a typed intermediate layer** that is hallucination-resistant and auditable by construction. Unstructured AEC evidence (photo +

note + plan, with optional workflow metadata) is mapped to a per-field `{value, confidence, source}` contract: the neural layer keeps the generalization needed to describe open-vocabulary multimodal evidence, while the symbolic layer alone names GUIDs, so the design balances multimodal flexibility against hallucination control rather than trading one for the other (§2).

2. **A searchable representation of building topology**: an IFC-native graph, a deterministic query cascade, and a type-conditional spatial-address format. Together they answer how relational structure should be encoded so that a building knowledge graph is searchable for element-level grounding: raw IFC becomes a normalized, topology-enriched graph; typed constraints compile into deterministic Cypher plans with recall-safe fallback and candidate-pool fusion; and an ordinal position-slot (openings) plus a connectivity fingerprint (walls) reorder the retrieved pool from Top-1 4.9% to 78.5% at oracle level, with the realized opening-slot detector reaching 58.9% end-to-end and supporting defer-on-low-confidence routing (§2.3, §2.6, §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).

1.5 Scope, stated up front

Two scope boundaries frame every result. First, the base project is the interpreter layer: VLM constraint extraction, IFC graph construction, deterministic retrieval, traceability, and workflow handoff. The spatial address, specialist detector, and calibration gate are an enhanced module added to reduce the repeated-element ranking gap. The deployed ranking is deliberately distilled: the fine-tuned VLM supplies the typed constraints that build the recall-safe pool, a deterministic opening-slot specialist supplies the within-pool discriminator, and a calibrated gate decides when to defer. We also engineered and evaluated a fuller stack, adding a learned size-band filter and a Gemini Graph-RAG reranker over the shortlist; it did not improve on the distilled path and is reported as an ablation (§3.6), so the distillation is a measured result, not a convenience. Second, all measurements are on a single synthetic project (cold-start by design). The system form is transferable, but the reported reliabilities are project-specific until validated on real site traces.

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 direct alternative, a single cross-attention model that ranks candidates end-to-end, is the strongest "bitter-lesson" objection [12], and we answer it empirically. The candidates are graph nodes carrying no image, so an image-space result does not reach a GUID: the only bridge is a coordinate-free relational address, and the discriminating signal is relational, not pixel-local ("the third of five fillers along an external wall"), which a pixel matcher would have to re-derive internally and unreliably. The evidence is that

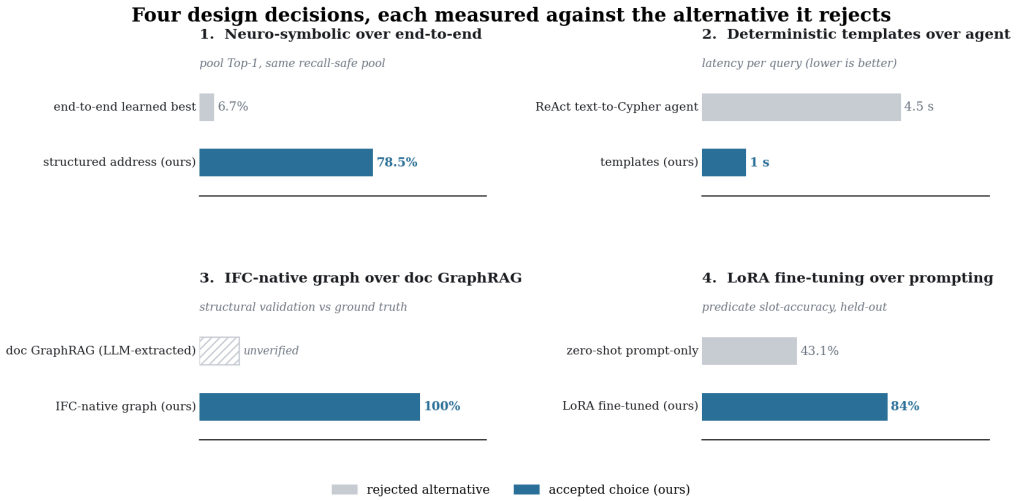

Figure 3: Design rationale summary. The accepted design uses structured neuro-symbolic parsing, deterministic query templates, an IFC-native graph, and LoRA adaptation.

our strongest *learned* configuration plateaus on exactly this discrimination: our fine-tuned-VLM pipeline with no within-pool reranking reaches Top-1 6.7% (lexical/dense retrievers $\leq 1.7\%$) on a pool that reorders to 78.5% once the relational address is supplied (§3.2). Finally, construction compliance requires repeatability and audit, which a scalar matching score cannot provide. 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 $\sim$4.5 s versus $\sim$1 s for the constrained pipeline and chose different tool sequences for the same input, which is unsuitable 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], a noisy process 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 and cannot improve over a project's lifecycle. LoRA fine-tuning accumulates field feedback as supervised examples for periodic retraining, and runs locally at $\sim$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 project-level pipeline has five stages: multimodal input (photo + note + plan crop, with optional workflow metadata) → neural extraction into typed constraints → deterministic query compilation against the IFC graph → ranked GUID candidates → validation, visualization, and workflow handoff. Figure 5 traces the enhanced version of this pipeline on one case. The evidence flow is intentionally ordered from raw observation to auditable execution: a vocabulary-constrained VLM reads the coarse semantic prefix; deterministic specialists recover 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.** The deployed pipeline is five stages:

1. **Multimodal input:** a site photo, a short note, and a marked floorplan crop.
2. **Neural perception:** the fine-tuned VLM reads all three and emits a typed constraint record (storey, IFC class, and typed spatial relations), each field carrying a value, a confidence, and a source tag; it describes, it never names a GUID.
3. **Symbolic retrieval:** the constraints compile into a fixed Cypher cascade over the enriched IFC graph, and a recall-safe union returns a candidate pool with the target retained in ∼100% of cases. The relational predicates do pool formation here, not within-pool ranking.
4. **Deterministic discrimination and soft rerank:** the OpenCV opening-slot specialist reads the position-slot (i, M) from the plan, and the pool is reordered by a soft score (storey + class + confidence-weighted slot match) that never removes a candidate.
5. **Calibrated answer/defer:** a temperature-scaled confidence commits the top GUID when it clears the threshold and otherwise defers the candidate pool to a human reviewer.

A fuller stack, a ResNet size-band filter and a Gemini Graph-RAG reranker, was evaluated and not adopted (§3.6); the deployed ranking is therefore deterministic end to end apart from the perception front end. Figure 4 gives the complete project view: an offline path that compiles raw IFC into the enriched Neo4j graph and mines paired training data (with a feedback loop), and an online inference path from multimodal input through the neural and symbolic layers to a BIM-linked output that supplies the element GUIDs required to populate a BCF/CDE issue-package schema (a fuller engineered stack, adding a learned size-band filter and an LLM Graph-RAG reranker, is evaluated as an ablation in §3.6 but not adopted). The spatial-address and calibration components sit inside this larger middleware as one ranking-and-routing stage.

At runtime, the boundary distinguishes dynamic site evidence from static reference resources. Inputs are site photos, marked floorplan crops, chat-style notes, and optional 4D context. Reference resources are the IFC model and schema-like output contracts. Outputs include structured constraints, candidate pools, selected or deferred GUIDs, execution traces, and handoff records for downstream coordination. The spatial-address module below is therefore one routing and ranking component within the broader interpreter middleware.

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

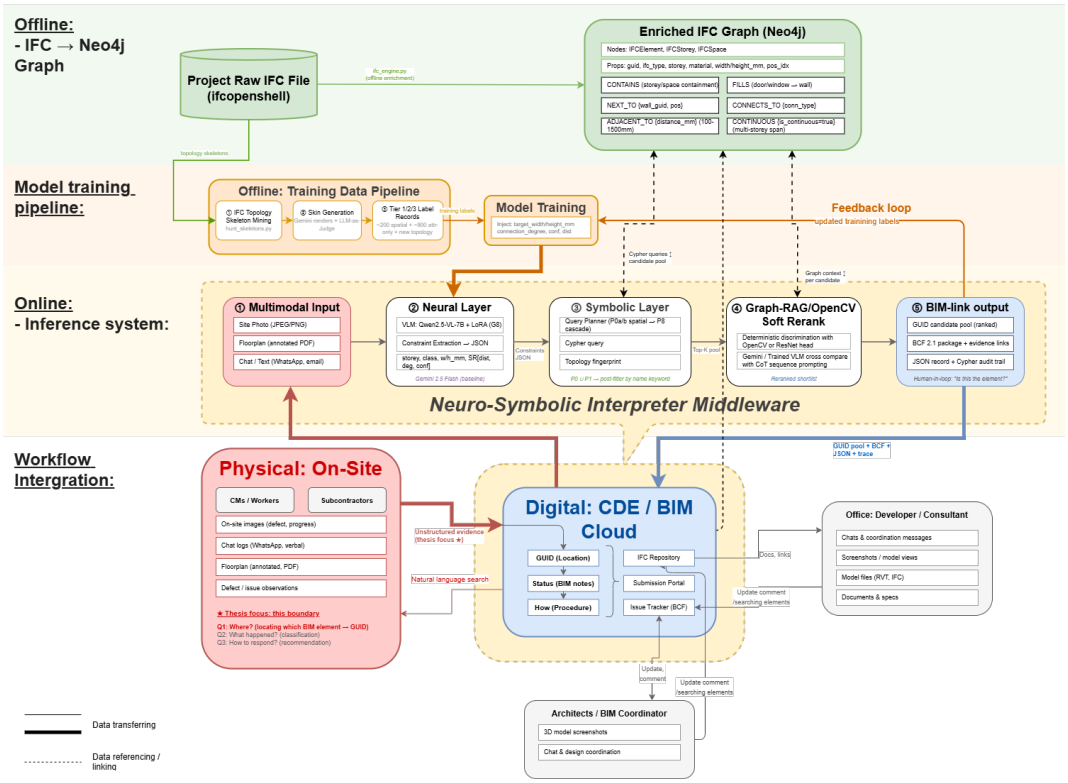

Figure 4: The neuro-symbolic interpreter middleware, end to end. *Offline:* raw IFC is parsed and topology-enriched into a Neo4j graph, from which skeleton-first synthetic cases are mined for training (feedback loop). *Online:* multimodal input → neural VLM layer → symbolic Cypher cascade → BIM-linked output (GUID pool, JSON record, BCF → CDE); an LLM Graph-RAG reranker was evaluated but not adopted (§3.6).

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 necessary because 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. Thus the graph exists, correct and complete, on day one with no human annotation: the cold-start property the system claims. Figure 7 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 language 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,`

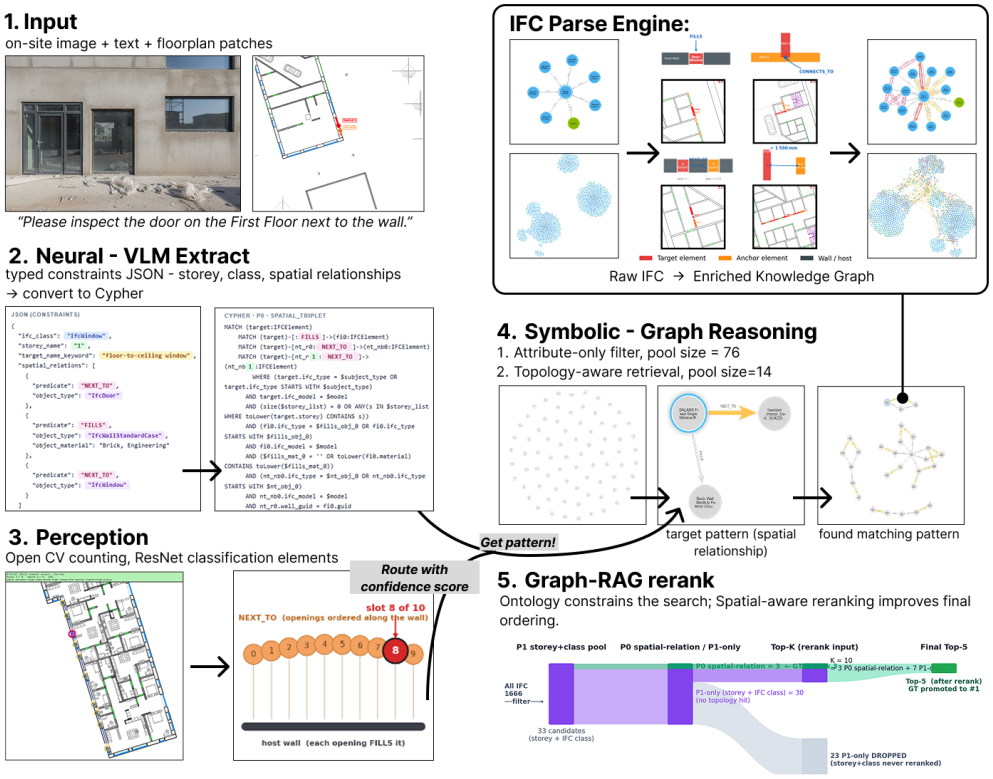

Figure 5: Interpreter pipeline on one held-out case. Mixed evidence becomes typed constraints, address fields, a calibrated route, and a ranked IFC GUID shortlist.

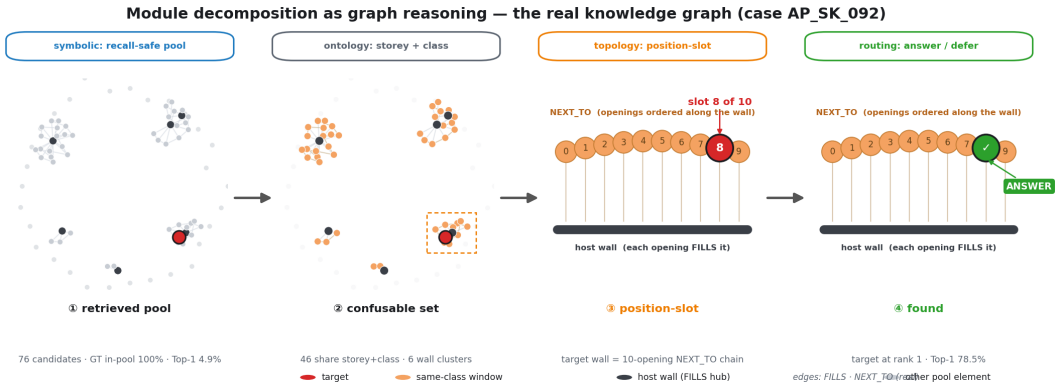

Figure 6: Graph reasoning for a repeated-window case. The interpreter first preserves a recall-safe pool; the enhanced position-slot signal then ranks siblings on the same host wall.

`spatial_relations:[{predicate, object_type, direction}]}`, and nothing else: *it describes, it never queries* (Figure 8). 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 adapter-configuration ablation below come from the project's full fine-tuning sweep, detailed in Appendix A.3), and the coarse fields (storey, `ifc_class`) saturate at 100% versus 30–63% zero-shot. This adapter sweep also exposes a multi-task capacity trade-off under a fixed adapter

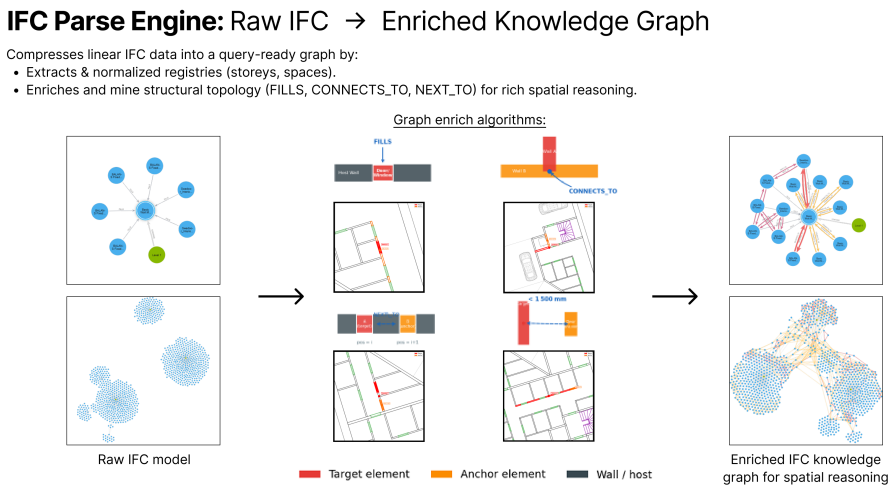

Figure 7: IFC parse engine. Raw IFC is normalized and enriched with retrieval-oriented topology before symbolic query execution.

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. 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.

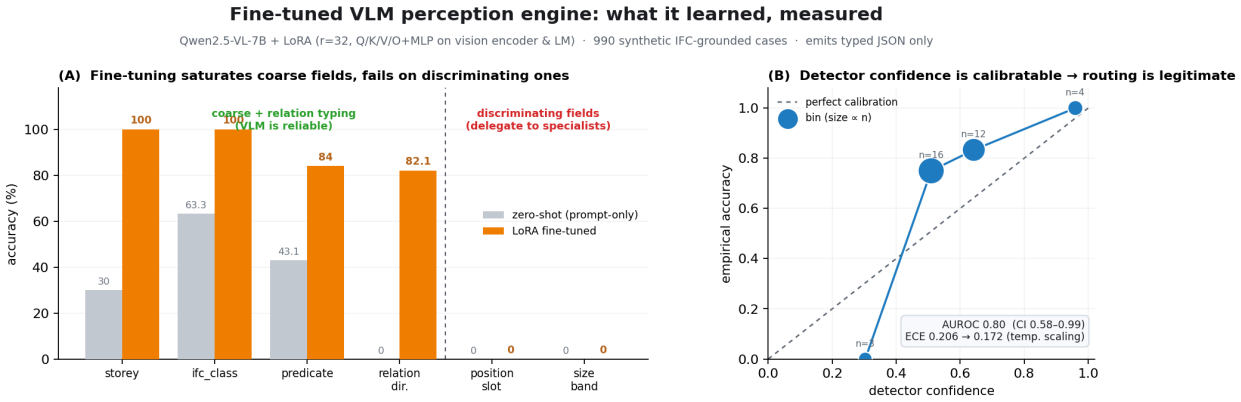

Figure 8: Neural extraction and routing diagnostics. Fine-tuning improves coarse fields and relation typing, while discriminating address fields motivate deterministic specialists and calibration.

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 hard negatives, including 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), plus

a 116-case cross-IFC set for generalization checks. Because every case is built skeleton-first from IFC semantics before stochastic visual evidence is added, the ground-truth address is fixed prior to augmentation (Figure 9).

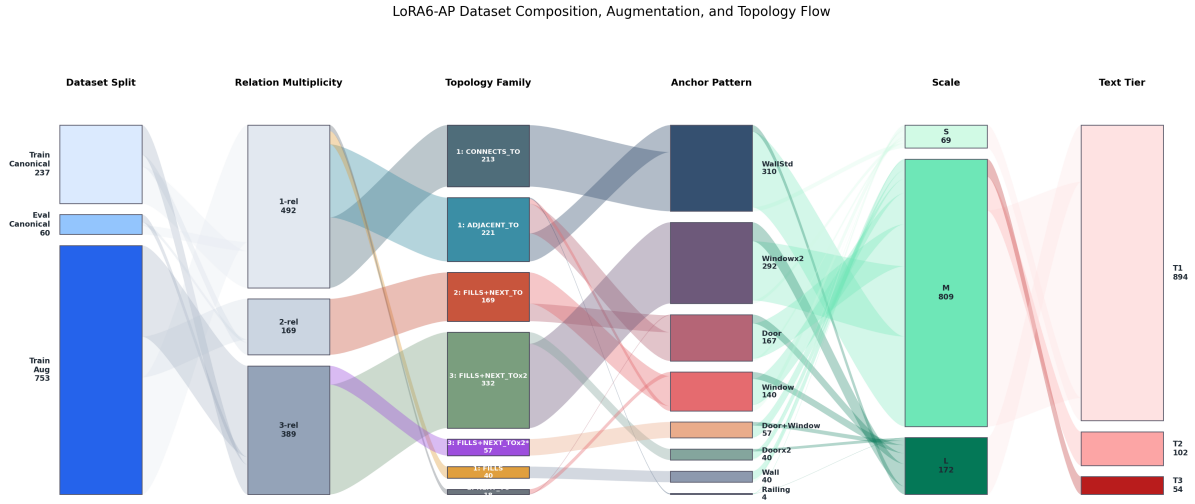

Figure 9: Skeleton-first synthetic curation. Each case is traced from the dataset split through relation multiplicity (1/2/3-relation), topology family, anchor-element pattern, evidence scale, and text tier; the dominant 3-relation family is `FILLS+NEXT_TO+NEXT_TO` (a window anchored between two openings on the same wall).

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; §3.4 reports the measurement behind this choice. Ranking inside the pool is delegated to the spatial address.

2.7 The spatial address, deterministic specialists, and calibrated routing

The enhanced ranking mechanism 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). Formally, for graph nodes V, storey $z(\cdot)$, and IFC class $y(\cdot)$, the first ambiguity class is

$$C(e) = \{u \in V : \; z(u) = z(e) \;\wedge\; y(u) = y(e)\}.$$

The address then augments this coarse prefix with the smallest class-specific discriminator:

$$a(e) = \begin{cases} (z(e), y(e), i(e), M(e)), & e \in \mathcal{F}, \\ (z(e), y(e), d(e), o(e), \ell(e), \chi_{\text{ext}}(e)), & e \in \mathcal{W}, \end{cases}$$

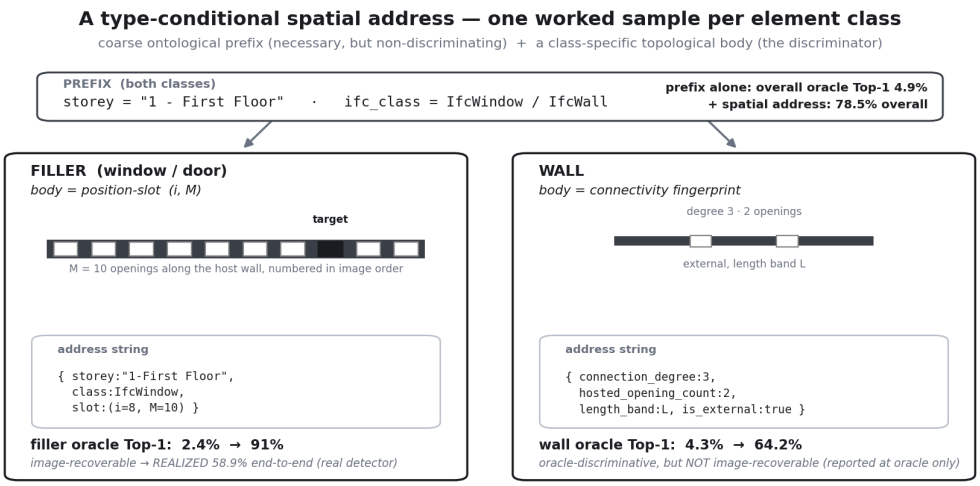

Figure 10: Type-conditional spatial addresses, one worked sample per class. A coarse storey+class prefix is necessary but non-discriminating (overall oracle Top-1 4.9%); a class-specific topological body completes it. Fillers use an ordinal position-slot (i, M) – the i-th of M openings along the host wall – lifting filler oracle Top-1 from 2.4% to 91.0% and, because the slot is image-recoverable, 58.9% realized end-to-end with the real detector. Walls use a connectivity fingerprint (degree, hosted-opening count, length band, external), lifting wall oracle Top-1 from 4.3% to 64.2%; it is oracle-discriminative but not image-recoverable, so it is reported at oracle only. Combined, the address lifts overall oracle Top-1 to 78.5%.

where $\mathcal{F}$ denotes fillers, $\mathcal{W}$ denotes walls, (i, M) is the ordinal opening slot, d is wall connection degree, o is hosted-opening count, ℓ is length band, and χ_{ext} is the exterior-wall indicator. 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, §3.2.1).

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 outperform the trained VLM on exactly these sub-tasks. Every extracted field, from every source, is emitted as a uniform record `{value, confidence, source}`, preserving 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. For a candidate c in the recall-fixed pool R, the deployed ranking score is a soft agreement score, not a filter:

$$s(c) = \lambda_z \mathbf{1}[z(c) = \hat{z}] + \lambda_y \mathbf{1}[y(c) = \hat{y}] + \lambda_a \mathbf{1}[a_{\text{disc}}(c) = \hat{a}_{\text{disc}}], \qquad c \in R,$$

where a_{disc} is the class-specific discriminating body of the address, e.g., the filler slot. The confidence

gate decides whether the top ranked candidate is actionable:

$$\text{decision}(x) = \begin{cases} \text{ANSWER}, & \hat{p}_T(x) \geq \tau, \\ \text{DEFER}, & \hat{p}_T(x) < \tau. \end{cases}$$

The premise that confidence may route at all is gated on calibration [7] and validated empirically (§3.5).

3 Evaluation & Results

3.1 Setup, metrics, and staging

All numbers are measured on the held-out benchmark of the studied synthetic project ($n = 60$ cases over 59 distinct IFC 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, §3.2.1) read only the IFC graph and learn nothing, and the realized position-slot is recovered by a *deterministic* OpenCV detector with no trained weights, so the slot itself is leakage-proof. The end-to-end filler number (58.9%) additionally consumes the VLM's extracted storey and `ifc_class`; re-profiling those coarse fields on the leakage-clean 48-case subset leaves them unchanged (storey/class saturated), so the leaked cases cannot account for the result. 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%, 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, including the depth analysis of where discrimination lives (§3.2, §3.2.1); 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); and last, failure decomposition and workflow-facing cost (§3.7). This staging separates symbolic retrieval quality from neural extraction quality because 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). The base interpreter is evaluated by target retention, traceable pool formation, and extraction quality; the enhanced module is evaluated by within-pool ranking and selective prediction. 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. This establishes target retention; 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

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
Extraction → deterministic retrieval, realized end-to-end ($n = 60$)				
Foundational VLM (Gemini 2.5, prompt-only)	91.7	1.7	18.3	0.056
Fine-tuned VLM (ours) → 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	18.1	76.3	0.352
+ structured spatial address	100	**78.5**	**98.1**	**0.854**
Realized address, addressable fillers ($n = 35$)				
Modal-prior slot (floor)	100	6.6	—	—
VLM coarse + OpenCV slot (deployed)	100	$\mathbf{58.9}_{[43.2,74.6]}$	67.1	—
Oracle coarse + OpenCV slot (upper bound)	100	$67.6_{[53.6,81.6]}$	80.9	—

Table 1: Held-out retrieval and ranking results. The base interpreter preserves the target pool; the enhanced address improves within-pool ranking, with bootstrap 95% CI shown in subscript. The bold *VLM coarse + OpenCV slot* row is the full realized neuro-symbolic system (fine-tuned VLM supplies the coarse fields, the deterministic detector supplies the position-slot); the rows above it are external baselines and oracle/upper bounds, not the deployed pipeline.

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 oracle Top-1 to 18.1% and Top-10 to 76%, but uniquely identifies only 2 of 60 targets. What remains is the irreducibly *relational* residue.

The type-conditional address resolves the residual ambiguity. 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. This 78.5% is the *overall* oracle ceiling, and it aggregates two class-specific addresses of unequal realizability: the filler position-slot (oracle Top-1 91.0%), which we realize from an image in §3.5, and the wall connectivity fingerprint (oracle Top-1 64.2%), which is oracle-discriminative but *not* image-recoverable. The realizable ceiling is therefore filler-weighted and below 78.5%; we treat 78.5% as an information bound, not a deployable target. 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 11). A strong general VLM cannot recover the relational slot from the image alone; the structured address, not the backbone, is what discriminates identical siblings.

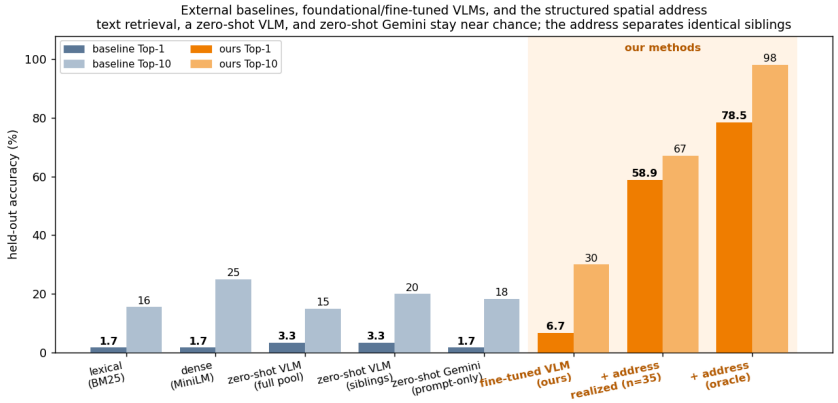

Figure 11: Ranking on identical recall-safe pools. Generic retrievers (BM25, MiniLM), a zero-shot VLM reranker, and zero-shot Gemini (prompt-only extraction → retrieval) all stay near chance on identical siblings. Our fine-tuned VLM is the perception engine and leads every learned method here (Top-1 6.7% vs. $\leq 1.7\%$ for the foundational and zero-shot baselines), but ranking pixel-identical siblings is not what extraction is for; that separation comes from the structured spatial address. The one realized detector reaches Top-1 58.9% end-to-end on the addressable filler subset (the $n = 35$ bar; its matching oracle is the filler 91.0%), and the full-pool oracle address is 78.5% (an information bound; all other bars are $n = 60$). The VLM's contribution is the field extraction the address consumes (Figure 12), not within-pool ranking.

3.2.1 How deep the discrimination lives: the depth law

The residual ambiguity is also *shallow*. Under an oracle the confusable set is already a singleton at one hop (median $|C|$ 13 → 1, Table 2), so deeper hops add nothing even in principle and the only question is whether that one hop is image-readable. Two measured facts answer it. First, **recovery does not collapse with depth**: the deployed model extracts the depth-k relation (predicate + object type + direction) correctly in 96.7% / 93.9% / 69.6% of cases carrying a hop at $k = 1/2/3$ ($n = 60/33/23$). Second, **what is readable is not discriminative**: the predicates are type-homogeneous (all 389 `FILLS` are window/door→wall, every `CONNECTS_TO` is wall↔wall), so a correctly-read hop matches every sibling and shrinks $|C|$ by nothing; the discrimination the oracle exploits is the specific neighbour's *identity*, which is not image-recoverable (the lone exception is the filler position-slot, §3.5). The depth law is thus *informational*, not a reliability cascade [2]: compile depth into the node and extract at depth ≤ 1.

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%. The gap carries downstream: on the same enriched graph, the foundational VLM (Gemini 2.5, prompt-only)

relational depth	oracle median $\|C\|$	VLM type-recovery (n)
attributes only (depth 0)	13	—
+ 1 hop	**1**	96.7% (60)
+ 2 hops	1	93.9% (33)
+ 3 hops	1	69.6% (23)

Table 2: Depth law. Oracle discrimination saturates at one hop; deeper recovered relation types are mostly homogeneous.

lands at Top-1 1.7% / MRR@10 0.056, whereas our fine-tuned VLM reaches Top-1 6.7% / MRR@10 0.110 (Table 1), confirming that domain fine-tuning, not the foundation model, is what makes structured extraction usable. End-to-end, 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 per-field accuracy comparison against zero-shot Gemini (Figure 12) makes the division of labour explicit: fine-tuning saturates the *coarse prefix* (storey, class, exactly the fields the oracle shows are non-discriminating) and recovers the spatial predicate and relation direction where the foundational model is at or near chance, yet *both* models recover the *discriminating* structured fields at **0%** (the position-slot and the size band), which is why those are delegated to deterministic specialists. The 82.1% direction figure is an *accuracy*; separately, the fine-tuned model *emits* a direction field in only 56.7% of held-out cases (an *emission rate*, whether any direction is produced; 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 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 preserves recall while intersection amplifies extraction error.

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 empirical floor is low: the fine-tuned reranker emits position as free text

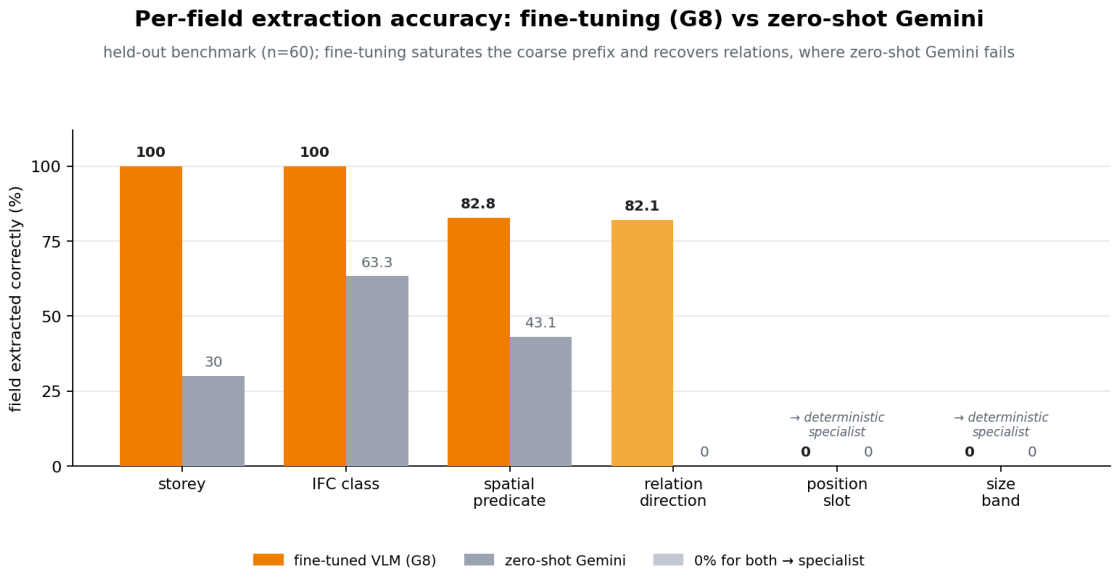

Figure 12: Per-field extraction accuracy, fine-tuned VLM (ours) versus zero-shot Gemini, on the held-out benchmark ($n = 60$). Fine-tuning saturates the coarse prefix (storey, IFC class at 100% versus 30.0% / 63.3% zero-shot) and recovers the spatial predicate (82.8% vs 43.1%) and relation direction (82.1% vs ∼0%). The discriminating position-slot and size band sit at 0% for *both* models and are delegated to deterministic specialists. The 82.1% here is a direction *accuracy*; it is distinct from the 56.7% rate at which the fine-tuned model *emits* any direction field, a different quantity we do not conflate.

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* from the model's own extraction (not oracle), it lifts filler Top-1 from the 6.6% floor to **58.9%** (95% CI $[43.2, 74.6]$, $n = 35$, Top-10 67.1%) at zero recall cost; supplying oracle coarse fields instead raises this to **67.6%** ($[53.6, 81.6]$, Top-10 80.9%), an upper bound on a perfect coarse extractor. The residual error is in the ordering index, not opening perception, and is not threshold-tuned: sweeping each detector threshold ±25% keeps Top-1 in $[55.1, 83.1]$, never near the floor.

(1b) The aggregate is composition-sensitive: we report the split. The 35 fillers are not uniform in difficulty. Eighteen sit on upper storeys re-rendered in isolation (host wall plus its openings, without surrounding corridors and walls), so their plans are far sparser than a realistic cluttered floor; the other seventeen are cluttered. Split both ways (Table 3), the cluttered subset reaches Top-1 **21.2%** end-to-end (39.1% with oracle coarse fields) versus 94.6% on the sparse subset, so the aggregate is flattered by its easier half. The deployable figure is therefore the cluttered-floor **21.2%** (a ∼8.8× lift over the 2.4% chance floor); 39.1% is the oracle-coarse upper bound. Closing the gap needs upper-floor plans rendered with full context (future work, §4.3).

Filler subset	Top-1 (end-to-end)	Top-1 (oracle coarse)	exact M
Realistic cluttered floors ($n = 17$)	**21.2**	39.1	12/17
Sparse host-wall renders ($n = 18$)	94.6	94.6	17/18
Aggregate ($n = 35$)	58.9	67.6	29/35
Oracle slot (ceiling)	—	91.0	35/35

Table 3: Realized position-slot split by plan difficulty, for the deployed end-to-end path (extracted coarse fields) and the oracle-coarse upper bound. The aggregate is composition-sensitive: realistic cluttered floors score Top-1 21.2 end-to-end (39.1 oracle-coarse) while sparser re-rendered upper-floor plans score 94.6 on both paths. The deployable figure is the cluttered-floor end-to-end result (21.2), not the aggregate. `exact` M is the detector's opening-count accuracy (independent of the coarse path); "oracle slot" is the perfect-detector ceiling.

(2) The address must be image-recoverable. The ground-truth slot is numbered along each wall's IFC local-X axis, a modelling frame invisible in any image: scored against it, detector and truth disagree on 16 of 35 fillers, *all* exact mirrors $i \mapsto M-1-i$, and an image-coordinate convention removes the disagreement (joint accuracy 34% $\rightarrow$ 74%). A field is realizable only if defined in a frame the evidence carries. The same constraint kills the wall fingerprint: its load-bearing fields need IFC wall-*instance* boundaries that collinear instances erase in the floorplan poché (length-band recovery 5/17), so the wall address is oracle-discriminative but not image-realizable, and realization is scoped to the filler slot.

(3) Confidence is calibratable; reweighting is a no-op. The raw detector confidence is positively discriminative (AUROC **0.80**, 95% CI $[0.58, 0.99]$) and moderately mis-calibrated (ECE 0.206 $\rightarrow$ 0.172 after temperature scaling, Figure 13); at $n = 35$ this is suggestive, not conclusive, and we treat the gate as an operating point [7]. As a *rerank* weight it does nothing: the slot is the finest tiebreaker in a storey$\times$class bucket, so any positive weight yields the identical Top-1. We report the no-op rather than bury it.

(4) The operational role is selective prediction. Sweeping the deferral threshold traces a coverage–accuracy curve (Figure 14): deferring the least-confident $\sim$20% lifts answered-subset Top-1 from 58.9% to **73.4%** end-to-end (67.6% to 80.6% oracle-coarse) [6]. A worked case (Figure 15) shows the mechanism: a wrong slot (predicted 1 of 10, truth 8) carries confidence 0.05, below threshold, so it defers and surfaces candidates instead of a confident wrong GUID. The coverage–accuracy curve, not a point estimate, is the appropriate report for a triage tool.

3.6 Does a learned reranker help? A redundancy ablation

The deployed ranking is deterministic; we built the fuller alternative to test whether a learned reranker beats it. The most engineered configuration adds two stages on the symbolic pool: a ResNet size-band classifier that filters by predicted dimension, and a Gemini Graph-RAG reranker that describes each shortlisted candidate by its graph fingerprint and reorders the top ten behind a deterministic fusion pre-rank. The result (Table 4, full pool $n = 60$) is a clean negative: on the topology-filtered pool the reranker *lowers* Top-1 from 6.7% to 1.7%, because the symbolic layer has already spent the relational signal and the LLM re-scores near-identical descriptions; it helps only on an unfiltered coarse pool (0% $\rightarrow$ 8.3%). Stacking the size-band filter on top reproduces the regression and evicts one ground-truth element, breaking the 100% retention guarantee (only a

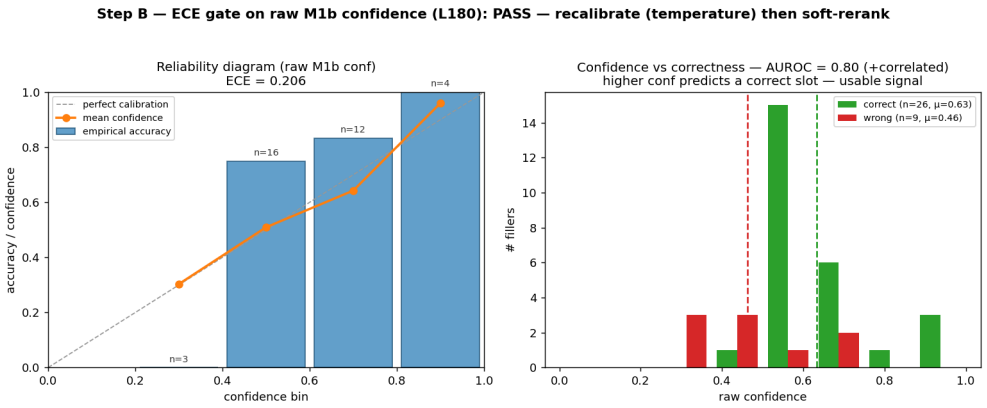

Figure 13: Calibration gate for the position-slot detector. Temperature scaling reduces ECE from 0.206 to 0.172.

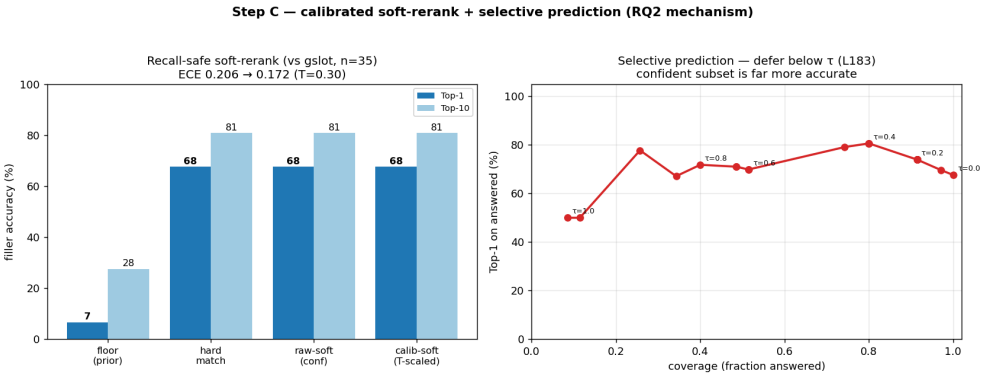

Figure 14: Selective prediction curve. Deferring low-confidence cases raises answered-set accuracy while reducing coverage.

score-fused variant is marginally positive, within noise).

We therefore keep ranking deterministic: the signal that survives pool formation is the filler position-slot (§3.5), and a calibrated gate decides when to defer. This is the union-over-intersection lesson (§3.4) applied to ranking: once the symbolic layer is doing the work, a black-box reranker trades auditability for noise.

3.7 Failure decomposition, latency, and candidate-pool collapse

Because the pipeline localizes errors to named fields, the residual loss decomposes cleanly: it is dominated by *extraction* errors (missing or wrong spatial predicates, and the lowest-accuracy fields: subtype, material, position-context), not retrieval logic, so every failure has a traceable stage and the roadmap is "improve the named weak fields," not "redesign the planner." Latency favours the constrained pipeline (∼1 s vs. ∼4.5 s for a ReAct agent, with exact repeatability). The workflow payoff is set-size reduction (Figure 16): over the same recall-safe pools the median candidate set collapses 76 → 46 (storey+class) → 1 (spatial address). This is a measured pool-size reduction, not a labour-time or user-study claim.

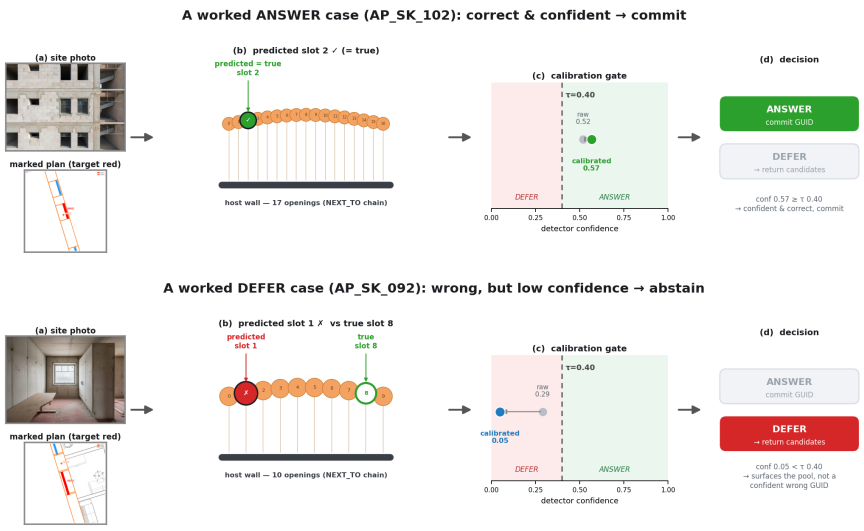

Figure 15: Answer/defer examples. The same confidence gate commits a high-confidence correct slot and defers a low-confidence slot error.

Post-retrieval ranking (full pool, $n = 60$)	GT-in-Pool	Top-1	Top-10	MRR@10
Topology retrieval, no rerank (Table 1)	100	**6.7**	30.0	0.110
+ Graph-RAG rerank, topology-filtered pool	100	1.7	30.0	0.082
+ Graph-RAG rerank, coarse storey+class pool	100	8.3	20.0	0.101
+ size-band filter + Graph-RAG (fully engineered)	98.3	3.3	26.7	0.092
+ size-band filter + score-fused Graph-RAG	98.3	8.3	26.7	0.124

Table 4: LLM-reranker ablation on the full held-out pool. A Gemini Graph-RAG reranker over the shortlist does not beat deterministic topology retrieval: on the topology-filtered pool it lowers Top-1 (6.7 → 1.7); it helps only on an unfiltered coarse pool, whose no-rerank baseline is Top-1 0%, where it adds the first discriminating signal (→ 8.3). The fully engineered variant (adding a ResNet size-band filter) regresses and evicts one ground-truth element (100 → 98.3 GT-in-Pool). The deployed system keeps ranking deterministic via the position-slot specialist (§3.5).

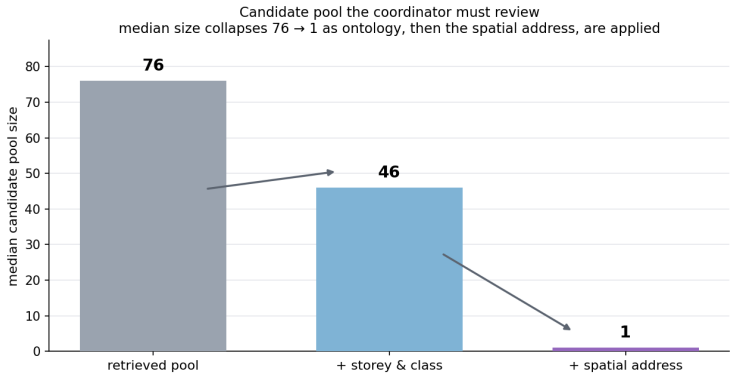

Figure 16: Candidate-pool collapse. The median number of candidate elements a coordinator must review falls from 76 (retrieved pool) to 46 (storey + class) to 1 (spatial address) over the same held-out cases.

4 Discussion

4.1 Workflow impact: a before/after walkthrough

The workflow effect can be described through a concrete walkthrough.

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 interpreter parses the evidence into constraints, executes the IFC graph cascade, and returns a compact ranked shortlist within ∼1 second. The correct element remains in the pool ∼100% of the time; on the addressable subset, the enhanced slot module places it at rank 1 in 58.9% of cases end-to-end (21.2% on realistic cluttered floors; 73.4% when the system is allowed to defer the least-confident fifth). The worker or coordinator confirms the element, and the output becomes a model-linked record rather than a loose image or chat message.

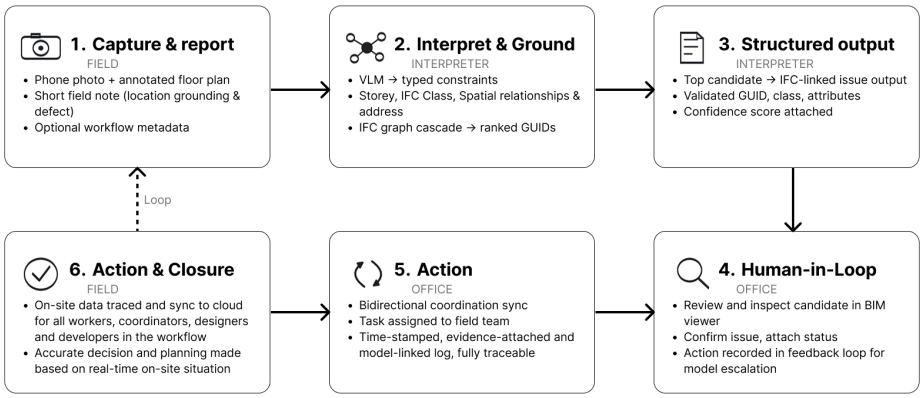

Figure 17: End-to-end workflow of the interpreter layer, shown as a closed loop: on-site capture → interpret and ground → IFC-linked issue (BCF) → office review and validation → action/update loop → verified on-site closure, which feeds the next observation. Each card is tagged with its actor (field, machine interpreter, office coordinator); the dashed return edge marks the continuous loop.

The claim is bounded: this does not replace the coordinator. It shifts the coordinator from manual search and transcription toward verification, correction, and escalation. Confirmed GUIDs can later serve as supervised feedback for project-specific fine-tuning.

From a single answer to institutional memory. Once evidence is anchored to a GUID, each element accumulates a traceable, time-stamped history rather than a loose chat thread, mitigating

the context decay that follows personnel turnover (Figure 18, left). Aggregated across a project, these element-level records support macro-scale analytics that are hard to see in siloed reporting: heatmaps of defect density, spatial trade conflicts, and recurring inspection or regulatory blind spots (Figure 19). The same neuro-symbolic logic extends toward continuous, autonomous on-site tracking as manual capture is replaced by a camera network. These are downstream implications of the grounding layer, not claims evaluated here; they motivate the project framing in which the spatial-address enhancement is one component.

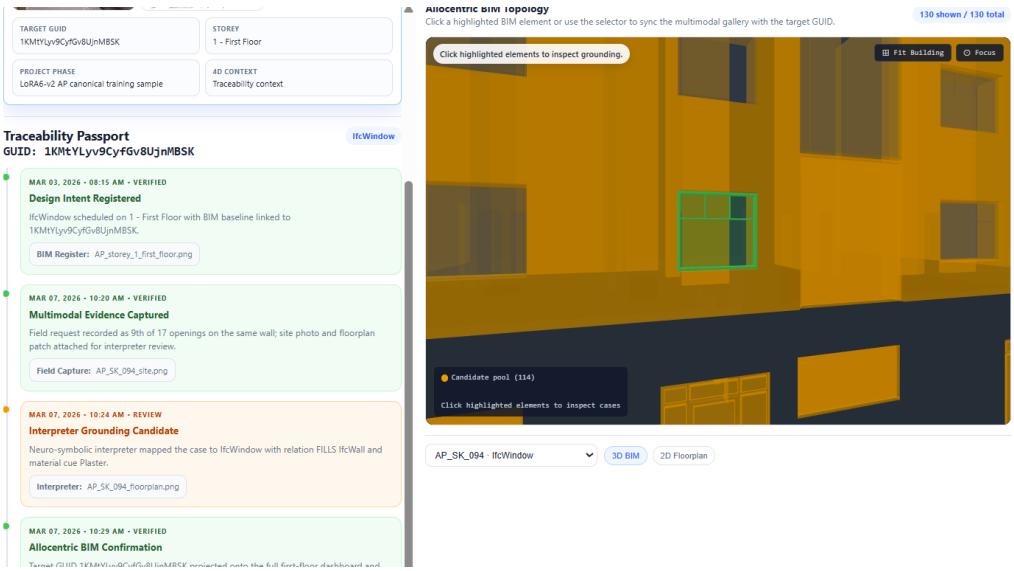

Figure 18: Element-level traceability "passport": a per-GUID audit trail (design intent registered → multimodal evidence captured → interpreter grounding candidate → allocentric BIM confirmation) synchronized with the 3D model.

Figure 19: Macro-scale traceability heatmaps aggregating element-level data to reveal defect density, trade conflicts, and regulatory blind spots that are difficult to detect in traditional siloed reporting.

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. *Test-set composition:* 18 of those 35 fillers are on sparse re-rendered upper-floor plans that are easier to segment than realistic cluttered floors, so the aggregate slot accuracy overstates deployment; the realistic figure is the cluttered-floor 21.2% end-to-end (Table 3), and we report the split throughout. *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, which can be mitigated by versioning the graph and surfacing low-confidence results for manual check.

4.3 Future work, prioritized

The priorities follow directly from the limitations. The first and most credible is **real project traces**: validating on live coordination data, since the synthetic-to-real shift is the single largest threat to the numbers. Next are the **remaining address classes**, which extend the realized address beyond the filler slot: size-band realization through the ResNet head, relation-direction improvement via subtype-contrastive augmentation, and room/space addressing wherever the export carries space boundaries. Beyond hand-tuned topology, a **learned spatial graph model** would replace the distance-threshold `ADJACENT_TO` heuristic and generalize past standard geometries. Closing the loop, **write-back integration** with BCF and the BIM cloud would land the grounded GUID in a live workflow rather than a report, turning confirmed GUIDs into supervised feedback. Finally, adding **4D temporal context** (schedules, timestamps, and progress updates) would make 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 an interpreter middleware: a VLM parses mixed evidence into typed constraints, an IFC-native symbolic layer retrieves legal GUID candidates, and a validation/handoff layer keeps the result traceable for human review. The oracle analysis supports the architecture: the symbolic layer retains the ground truth in 100% of cases and isolates extraction, not retrieval design, as the bottleneck. The spatial-address enhancement then addresses within-pool ranking: a type-conditional address reorders the pool from Top-1 4.9% to 78.5% at oracle level, and one deterministic specialist lifts the addressable subset from 6.6% to 58.9% end-to-end (21.2% on realistic cluttered floors). Selective prediction raises the answered subset to 73.4%. The depth analysis supplies the principle behind the enhancement: 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. First, site-to-BIM AI should be evaluated as coupled stages, not as a single model score: extraction, graph retrieval, ranking, and handoff fail in

different ways. Second, an address field is deployable only if it is defined in a frame the evidence carries; image-recoverability is a design constraint. Third, mature signals should preserve recall, while uncertain signals should rank or defer. Within the regime we can measure, the interpreter layer makes site-to-BIM grounding auditable, and the spatial-address enhancement makes repeated-element ranking substantially more usable. Real project traces are the immediate next validation step.

Acknowledgements

We thank Daniel Cardoso Llach (Computational Design Laboratory, Carnegie Mellon University) for his advisory guidance and feedback on this work as thesis reader.

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.

A Reproducibility Details

This appendix expands implementation details that are compressed in the main text. It is intended to make the arXiv version easier to audit without changing the paper's claims: all reported numbers remain synthetic/IFC-grounded and single-project unless explicitly stated otherwise.

A.1 Input protocol, data construction, and split

The input protocol is a cold-start site-to-BIM setting: a site image, a short text note, and a marked floorplan crop are grounded to an IFC element GUID. The current experiments do *not* use real construction photographs or live project traces. The site imagery is synthetic, and the floorplan crop is a designed human-marked input that indicates the local target/anchor region; this corresponds to a coordinator marking the relevant plan area before the system grounds the target element.

Item	Value used in this paper
Primary IFC model	`AdvancedProject.ifc`, used for all headline measurements.
Indexed elements	1,233 elements across 10 storeys (9 containing elements).
Repeated classes	263 windows and 771 walls (`IfcWall` + `IfcWallStandardCase`); typical floors contain many visually similar windows.
Hierarchy depth	Deepest IFC hierarchy: 5 levels.
Retained skeletons	297 region-grounded skeletons mined from the IFC model.
Training split	237 canonical training skeletons plus 753 augmented variants, for 990 AP-only training cases.
Held-out split	60 held-out cases / 59 unique target elements; 35 cases are addressable fillers for the realized position-slot experiment.
Pattern families	`FILLS`: 69; `CONNECTS_TO`: 69; `NEXT_TO`: 97; `ADJACENT_TO`: 62; `CONTINUOUS`: 22 mined skeletons, no generated cases in the reported benchmark.
Leakage audit	12 of 59 held-out target elements also appear as training targets under different renders. Oracle graph diagnostics and deterministic OpenCV-slot results are leakage-proof; the learned VLM profile is unchanged on the leakage-clean 48-case subset.

Table 5: Dataset and split details for the reported benchmark.

A.2 Neural extractor and LoRA settings

The neural extractor is treated as a structured parser rather than a GUID generator. Its output is a JSON constraint contract consumed by deterministic planning and retrieval.

Component	Setting
Backbone	Qwen2.5-VL-7B-Instruct.
Deployment	4-bit quantized backbone via Unsloth on Modal A100 GPUs during the thesis experiments. Repository inference loads `Qwen/Qwen2.5-VL-7B-Instruct` with bitsandbytes NF4 4-bit quantization when an adapter is supplied.
Adapter family	LoRA adapters; the deployed configuration uses rank 32 and learning rate 1×10^{-4}. The fine-tuning sweep (Appendix A.3) varied rank (16 vs 32), learning rate, and supervision targets.
Target modules	Q/K/V/O projection layers and MLP projection layers, as used for the multimodal structured-extraction adapter.
Final parser role	Extract `storey_name`, `ifc_class`, optional `space_name`, optional descriptor keywords, and topology-bearing `spatial_relations`. The parser does not emit GUIDs and does not generate Cypher.
Prompt parity	LoRA inference uses the same short system prompt family as training: extract IFC search constraints from multimodal evidence and output valid JSON only. The richer prompt-only schema is reserved for zero-shot/prompt baselines, not LoRA inference.
Known limitation	Exact training seeds, wall-clock training time, and full optimizer step logs should be attached before a journal resubmission if the original training logs are recovered. They are not inferred here.

Table 6: Neural-extractor settings included for reproducibility.

A.3 Fine-tuning sweep and a worked extraction example

The deployed extractor was selected from a sweep of LoRA adapter configurations that varied data augmentation, adapter rank (16 vs 32), learning rate, and supervision targets. The progression moved from attribute-only extraction, to multi-hop spatial predicates, to position-context and relation-direction labels, and finally to absolute-dimension targets. Two findings fixed the deployed choice. First, the strongest combined Top-10 / MRR@10 on the held-out benchmark comes from the configuration that adds position-context and dimension supervision at rank 32 (the *position-context + dimension* adapter used throughout §3); the immediately preceding position-context-only configuration is competitive on MRR@10 but weaker on Top-10. Second, an under-resourced rank-16 control collapses on spatial extraction while the coarse fields (storey, `ifc_class`) stay saturated at 100%, which is the multi-task capacity trade-off noted in §2.4: topology-specific supervision, not raw scale, is what makes the spatial fields appear. A separate hybrid configuration pairs the VLM parser explicitly with the OpenCV slot detector and the ResNet size head, the division of labour realized in §3.5.

The extractor is a structured parser: it consumes the multimodal evidence and emits the JSON constraint contract, never a GUID or Cypher. For a window query *"On Level 1, check the window beside the door at this location,"* paired with a site photograph, a marked floorplan crop, and the OpenCV opening count (`2nd of 3 openings on the same wall`), the merged contract is:

```
{
  "storey_name": "1",
  "ifc_class": "IfcWindow",
  "target_name_keyword": "floor-to-ceiling window",
  "position_context": "2nd of 3 openings on the same wall",
  "position_context_confidence": 0.92,
  "position_context_source": "opencv",
  "spatial_relations": [
    {"predicate": "NEXT_TO", "object_type": "IfcDoor",
     "direction": "right", "confidence": 1.0},
    {"predicate": "FILLS",   "object_type": "IfcWallStandardCase",
     "confidence": 1.0},
    {"predicate": "NEXT_TO", "object_type": "IfcWindow",
     "direction": "left",  "confidence": 1.0}
  ],
  "confidence": 0.85,
  "source": "lora"
}
```

The `storey_name` and `ifc_class` are the coarse prefix the VLM saturates; `position_context` is supplied by the deterministic detector, not the VLM; and each `spatial_relations` triplet carries the typed topology consumed by the P0 cascade. Every field also carries `{confidence, source}` provenance, so the planner can route it as a hard filter, a soft prior, or a deferral signal.

A.4 Constraint contract and deterministic specialists

The contract invariant is per-field provenance: each extracted field is represented as `{value, confidence, source}` and later assigned a routing role such as hard filter, soft prior, drop, or clarify. The top-level fields are `storey_name`, `ifc_class`, `space_name`, `position_context`, `size_cluster`, `size_band`, `target_name_keyword`, and `spatial_relations`. Nested spatial triplets include at least a predicate and object type, with optional subtype, direction, material, and confidence.

Specialist	Implementation detail
OpenCV position-slot detector	Gaussian blur 5×5, $\sigma = 0$; Canny 60/160; Hough $\rho/\theta = 1$ px / $\pi/180$; Hough vote threshold 20; minimum line length / maximum gap 14 px / 18 px; HSV saturation mask $S \geq 45$; morphological open kernel 3×3; minimum component area / aspect 10 px^2 / 1.3; confidence override gate 0.80.
ResNet size-band classifier	ResNet-18 with ImageNet1K-V1 weights; Linear(512, 6) head; 192×192 RGB input with ImageNet normalization; rotations $\{0, 90, 180, 270\}$, ± 8 px translation, color jitter (0.15, 0.10); weighted cross-entropy; AdamW with learning rate 3×10^{-4} and weight decay 1×10^{-4}; cosine annealing $T_{\max} = 30$; 30 epochs, batch size 16; split 155/17/23; test accuracy 82.6%.
Prototype caveats	The OpenCV wall section is supplied by the marked-plan/region context in this prototype. The ResNet classifier consumes precomputed world coordinates rather than a fully end-to-end region proposal. These details bound the current claims and motivate the future real-data integration.

Table 7: Deterministic specialist settings.

A.5 Query cascade, ranking, and metrics

The planner compiles the structured contract to a fixed Cypher cascade, not to open-ended text-to-Cypher. Every strategy with non-empty required fields emits an auditable plan; the retrieval backend fuses candidate pools rather than hard-intersecting noisy fields.

Level	Evidence and role
P0a	Spatial triplet edge traversal using `spatial_relations`, `ifc_class`, and non-`CONTINUOUS` predicates such as `FILLS`, `NEXT_TO`, and `ADJACENT_TO`.
P0b	Continuous-span property filter for `CONTINUOUS` cases, using materialized node properties such as `is_continuous` and `top_constraint`.
P1	`space_name` + `ifc_class`.
P2	`target_name_keyword` name/subtype lookup.
P4	`storey_name` + `ifc_class`; also used as the P0 safety-net tier.
P5	`storey_name` only.
P6	`ifc_class` only.
P8	Broad capped fallback pool.
Default fusion	`p0_union_p1`: return topology matches first, then append storey/type mates as a recall-preserving safety net. The relaxation ladder drops brittle filters only when a rung returns an empty pool: full fingerprint → no position → topology only → no storey → relaxed lowest-confidence triplet → attribute fallback.

Table 8: Deterministic query cascade used by the symbolic layer.

Let n be the number of evaluated cases, GT_i the ground-truth GUID, $Pool_i$ the returned candidate pool, and $rank_i$ the rank of GT_i when present. Retrieval health and ranking quality are reported as

$$\text{GT-in-Pool} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}[GT_i \in Pool_i],$$

$$\text{Top-}k = \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}[rank_i \le k], \qquad \text{MRR@10} = \frac{1}{n}\sum_{i=1}^{n}\frac{\mathbf{1}[rank_i \le 10]}{rank_i}.$$

Pool-compression diagnostics use recall-weighted reduction,

$$\text{RWR} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}[GT_i \in Pool_i]\left(1 - \frac{|Pool_i|}{N}\right),$$

where N is the full indexed model size. Rate estimates and calibration quantities use case-level percentile bootstrap confidence intervals with 10,000 resamples and seed 0 unless otherwise stated.

A.6 Artifact availability

The development repository is https://github.com/hyChia88/aec-interpreter. The paper build lives under `paper/`; benchmark/test-set metadata live under `data/test_sets/`; prompts and schemas live under `prompts/` and `schemas/`; evaluation scripts and figure generators live under `eval/`. Large artifacts such as IFC models, synthetic render corpora, model adapters, and generated outputs are documented in the repository but are not all bundled in git because of size and licensing constraints. The current arXiv-style claim should therefore be read as a reproducible system description with released code/metadata and documented artifact paths, not as a fully packaged public benchmark with real-site photographs.