formal300: Evaluating Local-to-Global Procedure Reconstruction from FAA Approach Charts

Anonymous authors
Paper under double-blind review

Abstract

Professional charts and machine records encode the same procedure at different representational levels: charts distribute intent across text, symbols, and spatial geometry, whereas machine records require ordered and normalized fields. We introduce formal300, a benchmark for recovering canonical missed-approach procedure structure from 300 FAA approach charts with 964 normalized legs. Rather than treating extraction as independent field prediction, formal300 evaluates a hierarchy from atom correctness to exact legs and exact procedures under a frozen canonicalization and scoring contract. A chart-blind majority prior attains 58.12% atom accuracy but recovers no exact leg or procedure, showing that correct individual scoring atoms need not form a correct ordered procedure. A layout-aware domain-rule pipeline uses OCR layout and public procedure rules to assemble ordered legs, while an auxiliary-trained 2B-parameter vision-language system predicts canonical structures from chart images. Their errors on individual scoring atoms overlap only partially, and both achieve exact recovery for only a small fraction of complete procedures. Reference recoverability is supported by a blind 200-atom re-audit and a separate blind trained-reader calibration yielding 59/60 exact submissions on 30 charts. formal300 provides an auditable, controlled testbed for studying how distributed professional-chart evidence is converted into canonical ordered structure.

1 Introduction

Professional charts and machine records often describe the same process in different forms. FAA approach charts communicate procedures through text, symbols, layout, and spatial geometry (Federal Aviation Administration, 2026a;c); ARINC 424/CIFP and navigation databases encode related machine semantics as fielded records (Airlines Electronic Engineering Committee, 2022; Federal Aviation Administration, 2026b). These representations are not field-by-field translations: charts compress procedure intent for human readers, while machine records require explicit ordered legs, fields, and value domains. This paper asks whether current models can recover such a machine-comparable structure from chart-side evidence.

We use FAA missed-approach procedures as a controlled, public, safety-relevant instance of this broader evaluation problem. A reader must recover where the missed approach begins, which fixes or tracks are followed, which altitude or distance constraints apply, when turns occur, and when holding begins (Federal Aviation Administration, 2017). The evidence is distributed across prose, plan view, profile view, local symbols, and public charting conventions rather than laid out as a JSON form. Related benchmarks study document-to-JSON extraction, map and chart QA, engineering diagrams, and safety-domain structured reasoning (Ferguson et al., 2026; Barzelay et al., 2026; Chang et al., 2022; Cheng et al., 2026; Rahul et al., 2019; Gupta et al., 2025; Sima et al., 2024); our target is different: ordered procedure structure aligned with machine-side semantics, not a local answer, generic field set, or unstructured explanation.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We introduce formal300 to evaluate ordered procedure reconstruction under a canonical representation and a three-level metric hierarchy. Here, local-to-global reconstruction gap denotes the observed disparity between partial atom correctness and complete ordered-structure recovery. Concretely, we contribute (1) a curated FAA chart-to-structure benchmark with chart-side recoverability checks, a frozen scorer, and an explicit canonicalization contract; (2) an evaluation protocol that separates atom correctness, exact-leg reconstruction, and exact-procedure reconstruction while quantifying chart-blind priors; and (3) a cross-protocol analysis that organizes diagnostic failures into evidence access, semantic normalization, and field-to-leg binding.

2 Related Work

Documents to Structured Representations. DocVQA, InfographicVQA, and ChartQA bring text, layout, and visual encodings from document pages, infographics, and charts into multimodal QA evaluation (Mathew et al., 2021; 2022; Masry et al., 2022). Recent professional document extraction work pushes the target from short answers to structured records, including PDF-to-JSON extraction, schema validity, and field-level scoring (Ferguson et al., 2026; Barzelay et al., 2026). Our benchmark inherits this structured-output form, but the output is not a generic field set; it is an ordered procedure structure organized by missed-approach leg and aligned with machine-side procedure semantics.

Professional Charts and Rule-Governed Diagrams. MapQA and ENC-Bench show that maps and navigational charts require specialized handling of symbols, scale, spatial relations, and task variation (Chang et al., 2022; Cheng et al., 2026). Engineering-diagram extraction provides a neighboring paradigm: P&ID tasks move from symbol and text recognition toward component association and graph-like structure generation (Rahul et al., 2019; Gupta et al., 2025). Missed-approach extraction shares with these tasks the difficulty that local elements must be bound to the correct structural position, but the target is not local QA or generic graph recovery; it is ordered missed-approach procedure semantics from FAA charts.

Safety-Domain Structured Reasoning. DriveLM formulates autonomous-driving scenes as graph visual question answering, using dependencies among perception, prediction, and planning questions to evaluate driving reasoning (Sima et al., 2024). This kind of safety-domain work motivates auditable structured semantics rather than unstructured answers. Our main task remains extraction of machine-side procedure semantics from public FAA charts; candidate-record consistency checking is a downstream extension rather than the main leaderboard target.

3 Task and Representation

We evaluate chart-to-machine-structure recovery rather than the reverse direction. Generating charts from machine records mainly tests controlled generation, charting standards, and readability. Our direction asks whether a model can recover implicit structure using only chart-side evidence. This recovery process must handle three structural mismatches: granularity mismatch, where one chart sentence or graphical element may correspond to multiple machine fields; location mismatch, where evidence for one leg may be distributed across missed-approach prose, plan view, profile view, and holding symbols; and expression mismatch, where charts may use rounding, abbreviations, layout proximity, and domain conventions while machine representations require explicit legs, fields, and values.

Task Interface. The unified input object is an FAA approach chart image. Different protocols may receive OCR text, the full image, image plus OCR, or task-relevant crops and question sequences; all inputs come from the same public chart or allowed independent preprocessing. Normalized reference JSON, scorer outputs, CIFP/424 records, and human error labels are not model inputs. The output is ordered leg-level canonical JSON containing chart_id, procedure metadata, missed_approach.leg_count, and an ordered missed_approach.legs[] array. Each leg is scored on six fields: Q_terminator, Q1_fix_ident,

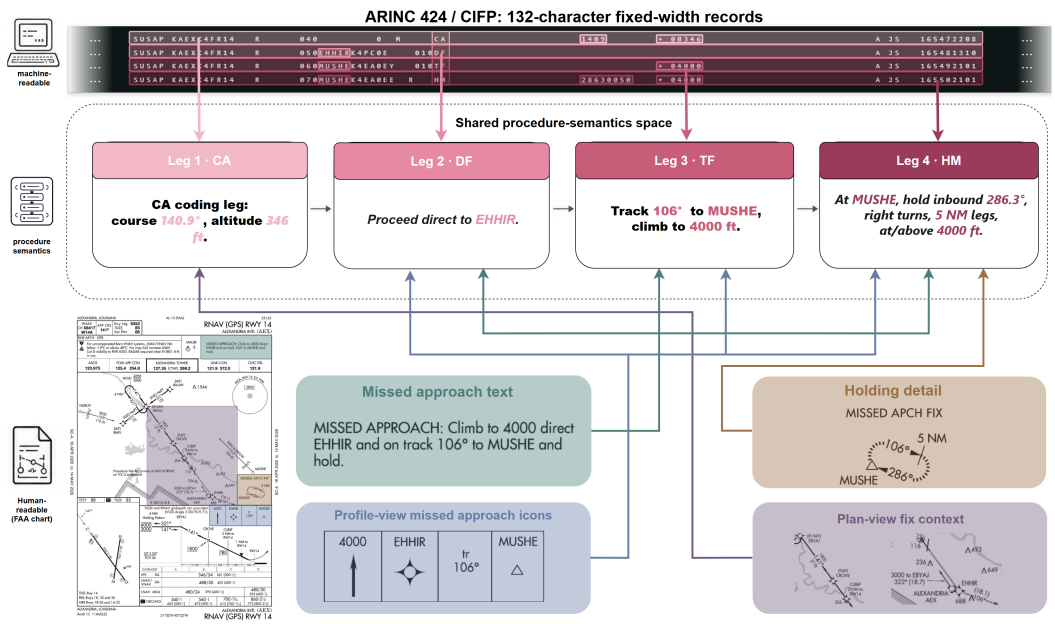


Figure 1: Two representations of the same instrument procedure. ARINC 424/CIFP records store procedure legs as fixed-width machine records, while FAA charts present the same procedure through text, symbols, and spatial geometry. The intermediate leg cards represent the shared procedure-semantics space targeted by this benchmark. This illustrative four-leg figure is separate from the three-leg KACQ_R33 worked example below.

Q2_altitude_constraint, Q3_turn, Q4_course_or_radial, and Q5_hold_params. The leg-count atom and each leg's answers.Q* field are status/value objects under the provided schema. Appendix A.1 gives a compact chart-to-JSON example. In that KACQ_R33 sample, chart evidence maps to three ordered legs: a CA climb leg, a DF leg to WITUD, and an HM hold at WITUD; text and plan-view evidence supply fix, altitude, inbound, and hold-length values, while the field-legality contract supplies legal completions such as unrestricted turn.

Table 1: Method input boundaries. Each row states test-time inputs and any declared training or example condition.

Method	Visible input and condition
Original A/B text protocols	Registered OCR strings and protocol-permitted candidates or links; original A1/A2 do not receive A1+'s two-dimensional OCR boxes.
Original C vision protocols	Full chart image or registered crop/QA inputs; only C4 uses its registered image-plus-OCR combination. Inputs differ across C rows.
A1+	PaddleOCR text and boxes/two-dimensional layout, public FAA/ARINC rules, and legal defaults; no target-procedure CIFP, reference, or scorer feedback.
D1-aux500	Test chart image and fixed prompt at inference; the 500/100 auxiliary package supplies training/development supervision, not extra test-chart input.
0/1/2-shot probes	Test chart and 0, 1, or 2 protocol-compliant in-context examples from airports disjoint from test airports; comparisons hold within model family and access channel.

Per-procedure CIFP/424 records, canonical references, scores, audit/error labels, and other methods' predictions remain evaluation-side; public domain rules and legal conventions may be used by methods.

Canonical JSON is designed as an evaluation-layer representation, not as a replacement for FAA or industry navigation databases. It preserves the missed-approach leg order,

terminator type, fix/navaid, turn, course/radial, altitude/distance, and holding fields needed for extraction, while omitting procedure-coding details that are outside the scored task. This choice prevents the task from collapsing into natural-language QA while also avoiding a requirement that models emit full ARINC 424 records. In other words, the required answer is not an explanation; it is a program-parseable intermediate structure that can be aligned to a reference and decomposed diagnostically.

formal300 is deliberately a reconstruction benchmark rather than a transcription benchmark. Some target values are directly expressed on the chart, whereas others are canonical completions fixed by public charting and procedure-coding conventions. We treat this canonicalization contract as part of the task definition and report chart-blind priors to contextualize aggregate correctness, which includes both chart-supported targets and convention-inferable completions.

Scoring Object. The scorer first checks whether the output is schema-valid canonical JSON; after validation, it aligns prediction and reference by ordered leg position. The formal test split denominator is 4052 equally weighted scoring atoms: 200 procedure-level leg-count atoms plus six per-leg fields over 642 reference legs. Missing legs, extra legs, invalid answer objects, and parse failures are retained as errors. The reference never uses unknown as a target, so model outputs of unknown are scored as ordinary values and do not receive wildcard credit. not_applicable or null can score only when that field is inapplicable for the leg type. The released comparison policy allows only chart-display equivalences, such as fix/navaid spelling variants and integer display of course/radial or holding inbound values; it does not infer missing fields or repair predictions from context. Procedure exact match is the strict endpoint for complete reconstruction; leg exact match measures whether an ordered structural unit is fully correct; and atom accuracy diagnoses partial field recovery. Because atom accuracy can remain high under chart-blind priors, we do not interpret it as a standalone measure of chart understanding.

4 Dataset and Reference Construction

A formal benchmark sample corresponds to one FAA approach chart image and its target missed-approach procedure. We build samples from public FAA d-TPP cycle 2604 terminal-procedure PDFs (Federal Aviation Administration, 2026c), retaining the source PDF, rendered PNG, FAA URL, sample manifest, canonical reference JSON, and raw CIFP/424 per-procedure records. The formal300 set contains 300 charts, covering 174 airports and 964 normalized missed-approach legs. It is a fixed selected benchmark sample assembled from two documented construction sources: an initial 100-chart set and a 200-chart expansion set recorded in the release manifest. It should not be interpreted as a random or frequency-representative sample of all FAA approach procedures. The paper-facing 50/50/200 split is generated from the fixed pool using a deterministic, score-blind, airport-grouped metadata stratification over procedure type, construction source, sample mode, and leg count. Samples are primarily RNAV but also include LOC and ILS, and cover holding, multi-leg, and simple missed-approach patterns with two to five legs. A separate 500/100 SFT-only auxiliary train/dev package is explicitly marked as outside formal300 and is not used in formal evaluation.

Table 2: Dataset statistics. Sample modes are mutually exclusive: holding means that the reference contains a holding leg; multi-leg means no holding and three or more missed-approach legs; simple means no holding and two missed-approach legs.

Category	Count	Notes
Formal charts	300	public FAA approach charts
Evaluation split	50 / 50 / 200	core train / validation / test
Airports	174	unique airport identifiers
Normalized missed-approach legs	964	after reference projection
Procedure families	262 / 25 / 13	RNAV / LOC / ILS
Sample modes	211 / 62 / 27	holding / multi-leg / simple
Leg counts	2–5	per missed-approach procedure
Auxiliary SFT package	500 / 100	SFT train / dev, outside formal300

The normalized missed-approach reference representation is derived from evaluation-side CIFP/ARINC 424-related records. CIFP is an FAA raw ARINC 424 data product updated on a 28-day cycle (Federal Aviation Administration, 2026b). For each sample, we align chart airport, procedure identifier, and runway context with CIFP/424 records, then select the ordered missed-approach transition records associated with the target procedure. The projection step converts these records into ordered canonical JSON while preserving leg order, path terminator, fix/navaid, course/radial, turn, altitude, distance, and holding fields. Raw 424 fields outside the scored task remain evaluation-side provenance rather than model-visible input.

This construction deliberately separates reference construction from model input. CIFP/424 records provide the evaluation-side answer key, but the scored target is not the raw machine record. A field is admitted to the canonical scorer only when it is visible in the public chart or inferable from public charting and procedure-coding conventions; raw sequence numbers, transition identifiers, and other machine-only metadata remain unscored provenance. Models in the main extraction protocols may inspect only public chart-side clues. For post-hoc explanation, we also store evidence boxes linking missed-approach prose, plan-view regions, profile-view evidence, holding symbology, relevant tables, or local symbols to canonical legs and fields. These artifacts help determine whether a model failed because it did not see evidence, saw evidence but assigned it incorrectly, or failed while converting evidence into machine structure; they do not alter scoring and are not main-method inputs.

Training-data policy. formal300 is the official benchmark set and fixes the 50/50/200 split used for development, validation, and formal test reporting. The 500/100 auxiliary SFT package is provided as a separate task-specific training resource: it contains chart-image to canonical-JSON paired supervision, but not evidence-region annotations, recoverability labels, or field-level rationales. It is outside the formal300 denominator and must be declared separately by future systems. We therefore separate trained SFT diagnostics into a formal-train-only D-SFT-formal50 raw-format diagnostic and an auxiliary-trained D1-aux500 learnability baseline. Both use the same released target and scorer. D-SFT-formal50 asks whether 50 formal training charts are enough to produce strict JSON/schema-valid formal-test outputs; D1-aux500 asks whether the fixed D1 output interface and representation become learnable with the released auxiliary supervision.

Recoverability contract. The construction-team author recoverability audit addresses the main risk in using CIFP/424-supported references: the scored answer must not require information hidden from the public chart. On the formal200 denominator, 2197/4052 atoms are directly chart-supported, 1855/4052 are convention-inferable or legality-constrained, and 0/4052 scored atoms require CIFP-only metadata. This distinction is central to interpretation: chart-supported atoms test evidence recovery, while convention-inferable atoms test whether that evidence can be completed into the benchmark's canonical procedure representation. Some fields are therefore scored as legal completions rather than visually read values. For example, unrestricted CF/DF turn is represented as BOTH, and DF course/radial is not_applicable. Appendix B gives the field-family audit and examples. To assess reference

reliability and human recoverability, two external re-auditors independently reviewed a risk-stratified sample of 200 atomic targets from 40 formal-test charts. Before adjudication, the review agreed with the reference on 197/200 atoms. The remaining three disagreements arose from direction normalization in a single reciprocal-course case and were resolved using the prespecified inbound-course convention. No sampled atom was judged unrecoverable from the source chart and public conventions. This audit provides direct chart-side support for the sampled reference targets and shows that trained readers can determine them under a shared annotation protocol (Appendix Table 7).

Quality control checks the reference projection, evidence links, schema validity, and input boundary. Three domain-trained annotators curated formal300: one prepared each chart, two others checked it, and roles rotated across the set. All formal300 charts were checked, with about 80 chart-level adjudication fixes across reference projection, evidence checks, and schema/legality rules. The annotators had flight-procedure-design training; one also had graduate-level aeronautical-information background. A pre-curation timed expert exercise was used only as an internal feasibility check and is not reported as a human-performance baseline because the participants later contributed to reference curation. A separate 30-chart blind expert calibration with two independent domain-trained readers achieved 1211/1212 pooled scored atoms and 59/60 exact procedures, comprising 30/30 for E2 and 29/30 for E3 under the strict scorer; details are in Appendix B.3. The re-audit and blind trained-reader calibration address distinct validity questions: the former assesses support for sampled reference fields, while the latter assesses completion of the ordered canonical task on 30 charts. Main extraction methods may read only the chart-derived inputs registered for their protocol; CIFP/424 records, reference JSON, validation reports, scorer outputs, error labels, and previous outputs are forbidden inputs.

Artifact availability. An anonymized supplementary archive is provided with this submission. It contains benchmark manifests, formal200 references, the schema and frozen scorer/comparison policy, method code, prompts and configurations, reported outputs and per-chart scores, validation summaries, and reproduction scripts. It supports deterministic re-scoring and re-aggregation. Hosted-model outputs are recorded evaluation snapshots, not guarantees of exact future API replay. With formal200 references included, formal200 should be treated after release as a fixed audit and development benchmark rather than a hidden leaderboard for post-release state-of-the-art claims.

5 Evaluation Protocols

All main results are reported on the same formal200 charts, references, schema, and frozen scorer. The original A1/A2 comparison fixes deterministic rules and changes only the OCR source, isolating text-recognition differences. The B-series uses OCR text, flat candidates, or automatic links as text-side inputs, testing text LLM conversion and candidate binding. The C-series uses full images, QA aggregation, structured questionnaires, or image+OCR fusion, testing visual evidence access and elicitation protocol. D-SFT applies LoRA SFT to Qwen/Qwen2-VL-2B-Instruct (Wang et al., 2024); at inference it receives only the full chart image and a fixed prompt. D1 is the fixed output adapter for D-SFT; its logged transformations use no reference JSON, scorer outputs, OCR, CIFP/424 records, field candidates, or other method outputs (Appendix C.3). A1+ retains PaddleOCR text boxes and combines text-unit isolation, two-dimensional label–value pairing, holding geometry, FAA/ARINC instruction grammar, deterministic leg assembly, and domain defaults; its inference does not read CIFP, references, or scorer feedback.

The protocol matrix spans complementary extraction routes: OCR-plus-rules, text-LLM conversion, direct visual prompting, image–text fusion, supervised chart-to-canonical learning, and a layout-aware domain pipeline. A1+ changes multiple pipeline components relative to A1 and is therefore treated as a separate pipeline rather than a single-factor ablation. All formal200 results share the chart set, reference, schema, and scorer; fuller protocol rows appear in Appendix Table 9.

Every method outputs the same canonical JSON and is evaluated by the same deterministic scorer. Schema-invalid outputs, parse failures, and coverage failures are retained as end-to-end system behavior; the reported scores therefore reflect both reconstruction and output-interface reliability. The main table uses the full formal200 denominator; methods with incomplete coverage receive zero correct for missing charts while preserving those charts' reference totals. Displayed confidence intervals are recomputed from saved per-chart scores under the corrected reference using 10,000 chart-level bootstrap resamples (seed 20260504); they condition on the frozen method snapshot and do not estimate training-seed variance. The C2 batched-leg QA supplement is used to show protocol sensitivity and is not included in the frozen main protocol comparison.

6 Experiments and Results

Table 3 follows the metric hierarchy defined above. Exact-match columns summarize structural recovery, while atom accuracy reports field-level diagnostic behavior under the same scorer.

Table 3: Main formal200 results under a shared reference, schema, and scorer. Rows differ in input interface, supervision, and domain knowledge (Table 1); cross-row differences are therefore descriptive protocol comparisons rather than controlled causal ablations. Exact-match columns measure structural recovery, while atom accuracy is a field-level diagnostic.

Role	Row	Condition	Proc. exact	Leg exact	Atom acc.
Prior	All-NA+count	no chart input	0/200	0/642	39.96
Prior	Formal majority	train-set majority	0/200	0/642	58.12
Oracle	Count+majority	true leg count	0/200	0/642	62.41
OCR/rules	A1	OCR + rules	0/200	0/642	28.90
Text LLM	B1	OCR text + LLM	0/200	0/642	22.83
Image VLM	C1	full-image VLM	0/200	0/642	37.69
Image+OCR	C4	image + OCR VLM	0/200	1/642	37.19
Aux. SFT	D1-aux500	Qwen2-VL + D1	8/200	198/642	75.20
Rules	A1+	layout-aware domain rules	24/200	344/642	87.81

Among the evaluated protocol snapshots, the layout-aware A1+ pipeline attains 3558/4052 correct atoms (87.81%), 344/642 exact legs, and 24/200 exact procedures, while auxiliary-trained D1 reaches 3047/4052 (75.20%), 198/642, and 8/200. The chart-blind majority prior obtains 58.12% atom accuracy but no exact leg or procedure, while A1+ recovers 24/200 complete procedures at 87.81% atom accuracy. These metrics distinguish partial correctness from complete reconstruction; their difference includes error accumulation under strict matching and does not by itself isolate binding-specific failures. This 58.12% value is a measured chart-blind reference level, not an optimal chart-blind baseline, a mathematical lower bound, or a chart-understanding score. The formal-train-only D1 diagnostic scores 10/4052 atoms; only 1/200 raw outputs are schema-valid (Appendix C.3).

D1 is a fixed, reference-blind output adapter applied before scoring; its deterministic canonicalization actions are logged and use no reference or scorer feedback (Appendix C.3).

In-context examples. Within a later C1 access channel, 0/1/2-shot settings score 39.88/68.21/61.85% atoms, 0/62/46 exact legs, and zero exact procedures; the original main-table C1 (37.69%) used a different channel and is not a paired zero-shot anchor. GPT-5.4 scores 31.81/42.23/47.56% atoms, at most one exact leg per condition, and zero exact procedures. The matched frozen, unfinetuned Qwen2-VL-2B stack scores zero at all three levels under the tested interface in all three shot settings. Of C1's 1148 net atom gains from zero to one shot, 721 (62.8%) concern explicit-value atoms, so the gain is not solely a format or status-vocabulary correction. These matched probes show that in-context examples can

substantially change atom- and leg-level recovery within some access channels, but the effect is model- and protocol-dependent and does not translate into exact-procedure recovery in the tested settings; no shot condition matches D1's 8/200 exact procedures (Table 4; matched setup and gain decomposition in Appendix C.4).

Table 4: Matched 0/1/2-shot results within each family on formal200. Each slash-separated triple corresponds to 0, 1, and 2 examples. GPT-5.4 recovers at most one exact leg in each condition.

Family	Atom acc. (%)	Exact legs	Exact procedures
C1 later channel	39.88 / 68.21 / 61.85	0 / 62 / 46	0 / 0 / 0
GPT-5.4	31.81 / 42.23 / 47.56	≤1 / ≤1 / ≤1	0 / 0 / 0
Qwen2-VL-2B, frozen no SFT	0 / 0 / 0	0 / 0 / 0	0 / 0 / 0

Stability and harder subsets. With data, epoch count, hyperparameters, and formal200 scoring fixed, three 2B training seeds yield 76.18±1.36% atom accuracy, 30.27±3.51% leg exact match, and 4.67±2.02% procedure exact match (mean ± sample SD). The submitted D1 run, 75.20/30.84/4.00%, lies within one SD for each metric. Harder non-three-leg and non-holding subsets still recover approximately 11–13% exact legs while the chart-blind prior recovers none; they do not recover full procedures (Appendix D.3).

Selected cross-cycle probes. To test whether D1's structural recovery is confined to FAA cycle 2604 formal300, we evaluated two selected probes under default strict comparison. In cycle 2603, D1 recovers 23/77 exact legs and 1/30 exact procedures across 30 charts, including at least one exact leg on a VOR-family chart. In cycle 2607, it recovers 14/177 legs and 0/53 procedures across 53 charts, scoring 59.01% atoms against a 49.60% chart-blind reference. These small, selected probes document some exact-leg recovery outside cycle 2604, rather than a controlled estimate of temporal transfer. The 2607 set enriches revised charts and lacks formal200-level independent validation; both probes use strict comparison rather than formal200 display equivalence and are reported separately (Appendix D.4).

7 Diagnostic Discussion

We organize the diagnostics around evidence access, semantic normalization, and field-to-leg binding. Source-view variants examine sensitivity to the presented chart views; reviewed-evidence conversion compares converters under specified evidence inputs; and residual categories summarize observed mismatches. These diagnostics characterize protocol-dependent recovery rather than provide a controlled causal decomposition of the three error sources (Appendix Tables 12 and 13 and Appendix E.3).

With the full chart minus missed-approach prose, D1 retains 68.39% atom accuracy, whereas the cropped MA-text-only input yields 11.82% (Appendix Table 12). This comparison shows input-view sensitivity, not the causal contribution of individual regions, because cropping changes evidence availability and input presentation.

The same reviewed MA text yields substantially different accuracy under rule-based and LLM conversion (Appendix Table 13), showing converter-dependent recovery even with reviewed text. These differences combine converter behavior with normalization and evidence granularity, as detailed in Appendix E.2. The diagnostic target is successful conversion into ordered canonical structure, not candidate-value presence.

A1+ has 494 atom errors under the corrected reference, spanning representation/normalization, OCR-character, and text-to-leg mapping failures. A1+ and D1 exhibit non-nested error patterns: D1 correctly predicts 226 of A1+'s 494 erroneous atoms, while A1+ correctly predicts 737 of D1's 1005 errors. These non-nested correctness patterns quantify complementary coverage of the shared canonical target across the rule-based and learned extraction routes. A reference-visible diagnostic union reaches 3784/4052 atoms (93.4%) and

44/200 procedures only as a diagnostic upper bound, quantifying complementarity under the shared canonical target (Appendix E.3).

8 Limitations

This paper evaluates normalized missed-approach extraction from public FAA approach charts, not SID, STAR, enroute charts, full procedure coding, obstacle assessment, database packaging, vendor loading formats, or operational navigation. Canonical JSON is an evaluation-layer representation and cannot replace FAA or industry records. The diagnostics explain benchmark behavior and do not constitute deployment-level safety evidence.

A further limitation is that aggregate atom accuracy mixes two capabilities: recovery of chart-expressed evidence and canonical completion under public procedure conventions. Nearly half of the scored atoms are convention-inferable or legality-constrained, which is why atom accuracy should be interpreted together with chart-blind priors and exact-structure metrics rather than as a pure perception score.

Responsible use is therefore part of the benchmark contract. The intended positive use is auditable research on professional-chart understanding, cross-representation structure recovery, and evaluation design. The main negative risk is overinterpretation: a benchmark score or model output could be mistaken for procedure-design, flight-planning, avionics-loading, database-update, or operational-navigation evidence. The release documents the sources, provenance, and non-operational scope, and all model outputs must be treated as research artifacts only.

formal300 is a fixed selected benchmark over FAA approach charts rather than a frequency-weighted sample of all FAA procedures or other chart systems. Formal train, validation, and test splits are airport-disjoint, while the separate auxiliary-training setting is additionally analyzed on airport-overlap and airport-disjoint subsets (Appendix D.2). Hosted-model rows are frozen evaluation snapshots, and the selected cross-cycle probes are reported separately because they use different chart selections and comparison conditions. These scope boundaries motivate future evaluation on additional procedure families, chart systems, and temporal cycles.

9 Conclusion

formal300 shows that chart-to-structure extraction is not well characterized by flat field accuracy alone. On a controlled FAA missed-approach task, a chart-blind prior recovers many individual atoms but no complete structural unit, while domain-rule and learned systems recover different subsets of exact legs and procedures. Together with the recoverability and reader audits, these results support an auditable testbed for evaluating how chart-side evidence is normalized into ordered leg-level structure, with complete reconstruction assessed separately from partial field correctness. formal300 is intentionally narrow: its role is to provide an auditable testbed for studying this cross-representation reconstruction problem, rather than to establish operational aviation capability or broad transfer across chart systems.

AI Use Statement

Large language models were used as experimental systems as described in the protocol sections. An AI coding and writing assistant was also used to organize existing study materials, revise the manuscript, prepare tables, and check compilation and consistency. This assistance does not constitute independent validation of experimental findings or adjudication of reference semantics. The authors are responsible for the submitted content.

Ethics Statement

The benchmark uses public FAA chart material and is intended for research on document understanding. Predictions and canonical outputs must not be used for flight planning, procedure certification, or navigation-database loading. Reader results are reported using study-local aliases and aggregate scores; the original 30-chart calibration protocol and its interpretive limits are disclosed in Appendix B.3.

Reproducibility Statement

An anonymized supplementary archive is provided with this submission. It contains the benchmark manifests, canonical references, schema and scorer, method code, prompts and configurations, retained outputs and per-chart scores for the reported runs, validation summaries, and reproduction scripts. Section 3 specifies the evaluation denominator and comparison policy. Appendix A gives a worked output and the scoring contract; Appendix B reports reference recoverability and independent reader checks; Appendix C provides baseline reproducibility details, including the protocol matrix, model interfaces, and output handling; Appendix D gives prior and slice checks together with the selected seed and cross-cycle protocols; Appendix E contains the source-view, evidence-conversion, and residual-error diagnostics. The supplementary archive supports deterministic re-scoring and re-aggregation under the frozen comparison policy. Hosted-model scores are recorded snapshots rather than guarantees of exact future API replay.

References

Airlines Electronic Engineering Committee. ARINC Specification 424-23: Navigation System Database. SAE Industry Technologies Consortia IA, 2022. Published 2022-03-30.

Udi Barzelay, Ophir Azulai, Inbar Shapira, Idan Friedman, Foad Abo Dahood, Madison Lee, and Abraham Daniels. VAREX: A benchmark for multi-modal structured extraction from documents. arXiv preprint arXiv:2603.15118, 2026.

Shuaichen Chang, David Palzer, Jialin Li, Eric Fosler-Lussier, and Ningchuan Xiao. MapQA: A dataset for question answering on choropleth maps. arXiv preprint arXiv:2211.08545, 2022. TRL @ NeurIPS 2022 poster.

Ao Cheng et al. ENC-Bench: A benchmark for evaluating multimodal large language models in electronic navigational chart understanding. arXiv preprint arXiv:2603.22763, 2026. CVPR 2026 accepted.

Federal Aviation Administration. Instrument Procedures Handbook. Federal Aviation Administration, 2017. FAA handbook.

Federal Aviation Administration. Aeronautical Chart Users' Guide. Federal Aviation Administration, 2026a. Effective 22 January 2026.

Federal Aviation Administration. Coded Instrument Flight Procedures. https://www.faa.gov/air_traffic/flight_info/aeronav/digital_products/cifp/, 2026b. Accessed 2026-05-04.

Federal Aviation Administration. Digital Terminal Procedures Publication. https://www.faa.gov/air_traffic/flight_info/aeronav/digital_products/dtpp/, 2026c. Accessed 2026-05-04.

Nick Ferguson, Josh Pennington, Narek Beghian, Aravind Mohan, Douwe Kiela, Sheshansh Agrawal, and Thien Hang Nguyen. ExtractBench: A benchmark and evaluation methodology for complex structured extraction. arXiv preprint arXiv:2602.12247, 2026.

Mohit Gupta, Chialing Wei, Thomas Czerniawski, and Ricardo Eiris. PIDQA—question answering on piping and instrumentation diagrams. Machine Learning and Knowledge Extraction, 7(2):39, 2025. doi: 10.3390/make7020039.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In Findings of the Association for Computational Linguistics: ACL 2022, pp. 2263–2279. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.findings-acl.177.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. DocVQA: A dataset for VQA on document images. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 2200–2209, 2021. doi: 10.1109/WACV48630.2021.00225.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and C. V. Jawahar. InfographicVQA. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 1697–1706, 2022.

R. Rahul et al. Automatic information extraction from piping and instrumentation diagrams. In Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods, pp. 163–172, 2019. doi: 10.5220/0007376401630172.

Chonghao Sima et al. DriveLM: Driving with graph visual question answering. In Proceedings of the European Conference on Computer Vision, 2024.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191, 2024.

A Task Object and Scoring Contract

A.1 Worked Chart-to-JSON Example

A sample pairs a public FAA approach chart with its target missed-approach procedure. The evaluated object has missed_approach.leg_count and ordered missed_approach.legs[]; each leg's six scored fields are expressed as status/value objects in the released schema. KACQ_R33 illustrates the distinction between reading chart values and producing a legal canonical target.

The following contiguous prefix is copied from the released KACQ_R33 reference JSON; it shows the actual hierarchy and one leg's first answer fields rather than reconstructing a schema from the readable table.


```
"missed_approach": {
  "leg_count": {
    "status": "present",
    "value": 3
  },
  "legs": [
    {
      "leg_index": 1,
      "answers": {
        "Q_terminator": {
          "status": "present",
          "value": "CA"
        },
        "Q1_fix_ident": {
          "status": "not_applicable",
          "value": null
        },
```


Table 5: KACQ_R33 ordered legs, selected chart evidence, released-reference excerpts, and evidence roles. The exact JSON hierarchy is quoted above.

Ordered leg	Selected chart evidence	Canonical reference excerpt	Evidence role
1 CA	MA text; plan-view track	AT_OR_ABOVE 1480 ft; course 332.2 deg	text and track; display rounding is scorer-governed
2 DF to WITUD	WITUD and 3000-ft labels	AT_OR_ABOVE 3000 ft; turn BOTH	chart values; BOTH is legality completion
3 HM at WITUD	MA text and plan view	AT_OR_ABOVE 3000 ft; inbound 152.1 deg; 4.0 NM; RIGHT	holding fields supported jointly

For CF/DF legs, an unrestricted turn is represented by BOTH. DF has no independent legal course/radial target, so that field is not_applicable; a TF course/radial counts only when the normalization rule explicitly adopts the charted value. The released reference has no open-ended unknown targets. The formal scorer applies limited display equivalence, while missing, invalid, or unparseable outputs remain on the full denominator. Schema, scorer, method code/configurations, and per-sample JSON are in the supplementary archive.

B Reference Recoverability and Human Checks

B.1 Author Audit of All Scored Atoms

References are constructed from corresponding CIFP/424 records, which are not model inputs. Only targets recoverable from chart evidence or public coding conventions enter the

scored object. Table 6 classifies all 4052 formal200 atoms by field family in the construction-team/author audit.

Table 6: Construction-team/author recoverability classification for all 4052 scored formal200 atoms; this is not an external audit of all atoms.

Atom type	Count	Chart-supported	Convention-inferable/legality	CIFP-only scored
Leg count	200	200	0	0
Terminator	642	642	0	0
Fix/navaid	642	439	203	0
Altitude	642	445	197	0
Turn	642	52	590	0
Course/radial	642	221	421	0
Hold parameters	642	198	444	0
Total	4052	2197	1855	0

B.2 External Blind Re-Audit

Two external readers independently reviewed a risk-stratified sample of 200 atoms from 40 charts. They could use charts and public conventions, but not the original reference answers or model predictions. Agreement with the reference was measured before adjudication. The three reported differences concerned one reciprocal-orientation case and were resolved using the pre-specified inbound-course convention. Table 7 reports the aggregate agreement used for the recoverability analysis.

Table 7: Reported external blind re-audit of 200 risk-stratified atoms from 40 charts; 197/200 is agreement with the reference before adjudication.

Sample or outcome	Count
Leg count / terminator / fix / altitude	7 / 18 / 44 / 44
Turn / course / hold	29 / 21 / 37
Pre-adjudication agreement	197/200
Reciprocal-orientation disagreements, one case	3/200
Judged unrecoverable in sample	0/200

No sampled atom was judged unrecoverable from charts and public conventions. This sample result does not establish zero errors across all 4052 atoms.

B.3 Blind Expert Calibration

E2 and E3 did not construct the formal300 references. Independently trained readers saw the charts, task instructions, and schema, but not the references, CIFP/424 records, or model answers; each made one submission per chart under strict scoring. The 30-chart sample comprised 17 holding, 7 multi-leg, and 6 simple cases, and 25 RNAV, 4 LOC, and 1 ILS chart.

Table 8: Blind trained-reader calibration on the complete task for 30 charts under the released strict scorer; not a formal200 human leaderboard.

Reader	Scored charts	Atom accuracy	Leg exact	Critical-field leg exact	Procedure exact
E2	30/30	606/606 = 100.00	96/96 = 100.00	96/96 = 100.00	30/30 = 100.00
E3	30/30	605/606 = 99.83	95/96 = 98.96	96/96 = 100.00	29/30 = 96.67
Pooled submissions	60 on 30 charts	1211/1212 = 99.92	191/192 = 99.48	192/192 = 100.00	59/60 = 98.33

The sole KAXH_R27 leg-3 Q4_course_or_radial difference was interpretive: the frozen reference adopts the TF charted course_deg=200.4, while E3 submitted not_applicable.

This disclosure changes neither the reference, scorer, nor main result table. The calibration checks trained-reader completion of the full task; it is not an estimate of formal200 or general-population human success.

C Baseline Reproducibility Details

C.1 Frozen Run and Full Protocol Matrix

The frozen Group 1 run group1_formal_eval_50_200_50_seed20260437_20260430_r1 was created on 2026-04-30 UTC. Its inference runner reads only the method input manifest; the scoring manifest and reference are read after predictions finish. Text/VLM schema retry is limited to one attempt without gold or scoring feedback, and semantic repair is disabled. A1+ was later evaluated separately on the same test surface and is not a row from the original Group 1 manifest.

Table 9: Original formal200 protocol matrix, with A1+ separately identified on the same charts, reference, schema, and scorer. A1+ changes multiple components relative to A1 and is not a single-factor geometry ablation. Bootstrap intervals appear only where computed.

Row	Regime	Input/protocol	Charts	Atoms	Atom acc. [95% CI]	Δ maj.	Leg/proc exact
A1	no SFT	OCR-1 + rules	200/200	1171/4052	28.90 [26.58,31.24]	-29.22	0/642; 0/200
A2	no SFT	OCR-2 + same rules	200/200	906/4052	22.36 [19.85,24.92]	-35.76	0/642; 0/200
B1	no SFT	OCR text + LLM	200/200	925/4052	22.83 [20.80,24.91]	-35.29	0/642; 0/200
B1_prime	no SFT	OCR text + flat candidates + LLM	200/200	1057/4052	26.09 [24.22,27.99]	-32.03	1/642; 0/200
B1_prime_link	no SFT	OCR text + auto links + LLM	185/200	721/4052	17.79 [15.85,19.81]	-40.33	0/642; 0/200
C1	no SFT	full image → VLM JSON	200/200	1527/4052	37.69 [36.04,39.34]	-20.43	0/642; 0/200
C2	no SFT	full image → fixed QA aggregation	200/200	1044/4052	25.77 [24.39,27.19]	-32.35	0/642; 0/200
C3	no SFT	full image → questionnaire → JSON	196/200	1442/4052	35.59 [33.71,37.46]	-22.53	0/642; 0/200
C4	no SFT	full image + OCR text → VLM JSON	200/200	1507/4052	37.19 [34.59,39.89]	-20.93	1/642; 0/200
D-SFT-formal50 raw (no D1)	formal SFT	strict raw-format diagnostic	0/200	0/4052	0.00	-58.12	0/642; 0/200
D1-formal50 final-v2	formal SFT	final-v2/D1 diagnostic	1/200	10/4052	0.25	-57.87	0/642; 0/200
D1-aux500	aux. SFT	Qwen2-VL-2B + D1 interface	200/200	3047/4052	75.20 [71.85,78.50]	+17.08	198/642; 8/200
A1+	rules	layout-aware domain pipeline	200/200	3558/4052	87.81	+29.69	344/642; 24/200

C.2 Baseline Interfaces and Environment

A1/A2 hold the rules fixed and change OCR source. OCR-1 uses PaddleOCR 3.5.0, Paddle 3.2.2, PP-OCRv5, en, and engine-returned order; OCR-2 uses Tesseract 5.5.0.20241111, eng, OEM 3, PSM 6, and TSV block/paragraph/line/word order. B1 receives OCR text; B1_prime adds flat field candidates and B1_prime_link adds automatic links. C1/C2/C3 receive the full chart without OCR text, while C4 also receives OCR-1 text. B-series hosted text calls use temperature 0, at most 4096 output tokens, and schema-bound tool output. C-series image calls use claude-sonnet-4-5-20250929 at temperature 0, at most 4096 tokens for C1/C3/C4 and 1024 for formal C2 QA, with tool output and one schema retry where applicable. Hosted API results are recorded snapshots, not guaranteed exact future replays.

A1+ uses PaddleOCR text plus boxes and geometry, text-unit isolation, two-dimensional label–value pairing, holding geometry, FAA/ARINC instruction grammar, deterministic leg assembly, and domain defaults. All A1+ development was completed on the designated development split before its single formal200 evaluation. The subsequent reference correction was applied uniformly to retained frozen outputs for re-scoring and did not trigger model or rule re-tuning. Inference reads no CIFP/424, reference, or scorer feedback.

C.3 D1 Output Interface and Formal-Train-Only Diagnostic

D-SFT uses Qwen/Qwen2-VL-2B-Instruct with QLoRA 4-bit NF4, double quantization, rank 8, alpha 16, dropout 0.05, and q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj targets. The reported run uses seed 260428, one epoch, batch size 1, gradient accumulation 8, learning rate 2×10^{-4}, maximum sequence length 4096, image min/max pixels 3136/501760, greedy decoding, and at most 1536 new tokens. Auxiliary 500-train/100-dev chart-to-JSON supervision lies outside formal300 and has no chart, PDF, exact procedure key, family key, image-path, or reference-path overlap with it; airport disjointness from formal test is not claimed.

D1 strips surrounding whitespace/code fences, parses JSON, wraps short form in the canonical envelope, fills unscored identity fields from the manifest, validates legality, and applies an empty canonical fallback to unparseable outputs. Missing answer keys, if any, receive literal unknown/null values under the frozen policy, not inferred answers. D1 reads no reference, score, OCR, CIFP/424 record, field candidates, or other method outputs; its deterministic transformations are documented in the action log below. Raw aux500 outputs have 196/200 prediction files and 184/200 schema-valid records; after D1, all 200 enter full-denominator scoring at 3047/4052 atoms. The action log records 190 whole-output JSON parses, 4 envelope wraps, 61 chart-ID fills, 6 empty fallbacks that receive official zero, 6 invalid-fix replacements, and 0 missing-answer-key fills. These counts are actions, not mutually exclusive chart classes.

The formal-train-only checkpoint yields 32 parseable raw outputs and 1 schema-valid output among 200; after the released final-v2 full-denominator path it scores 10/4052 atoms, with no exact legs or procedures. This differs from D1-aux500 in training data and is not a controlled SFT mechanism ablation.

C.4 Matched In-Context Controls

Within each family in Table 4, 0/1/2-shot conditions share the test charts, schema, temperature, and scorer; exemplar airports do not overlap test airports. The later C1-channel 39.88% zero-shot score is not paired with the original formal C1 37.69%, and GPT-5.4's 31.81% zero-shot anchor is specific to the matched setup reported here. C1's 1148 net additional correct atoms from zero to one shot comprise 427 reference-inapplicable and 721 explicit-value atoms; pure status-vocabulary correction accounts for only 8.7%. In matched 2B one-shot outputs, 193/200 copy the exemplar fix. The full shot results remain in the main-text table.

C.5 Compute and Artifact Boundary

Final D-SFT training took about four hours and formal200 inference about two hours on one RTX 4060 Laptop GPU; development runs used about 22–30 GPU-hours. Hosted API compute is not locally measurable. The supplementary archive provides released data/manifests, prompts, method configurations, retained outputs, validation records, per-chart scores, and reproduction scripts. It supports offline re-scoring and re-aggregation under the frozen reference and scorer. Hosted-model outputs are recorded evaluation snapshots rather than guaranteed future API replays.

D Interpretation, Stability, and Scope

D.1 Chart-Blind Priors

The 58.12% chart-blind prior is a majority-answer/legality reference learned from the 50 formal training charts and applied to formal200. It matches many atoms but no complete legs or procedures.

Table 10: Chart-blind formal200 references derived from formal50 training; 58.12% is an observed reference level, not a mathematical floor or chart-reading score.

Baseline	Correct/Total	Atom acc.	Procedure exact	Leg exact
All not_applicable, majority leg count	1619/4052	39.96	0/200	0/642
Majority CA–DF–HM pattern + legality only	2307/4052	56.93	0/200	0/642
Formal-train majority answer by leg index	2355/4052	58.12	0/200	0/642

It is an observed chart-blind reference level, neither a mathematical lower bound for other methods nor a chart-understanding score.

D.2 Split and Structural Slices

Formal train, validation, and test are airport-disjoint. Auxiliary SFT has no duplicate chart/PDF/procedure/family/image/reference path with formal300, but shares airports with 37 of the 111 formal-test airports. The D1-aux500 slices below use the released scorer and each group's scored-atom denominator.

Table 11: Selected D1-aux500 formal200 slices under the released scorer; each atom accuracy uses that slice's atom denominator, not its chart count.

Method	Slice	Charts	Accuracy
D1-aux500	all	200	75.20
D1-aux500	auxiliary-airport-disjoint	123	78.87
D1-aux500	auxiliary-airport-overlap	77	69.67
D1-aux500	3-leg	157	84.75
D1-aux500	non-3-leg	43	48.55
D1-aux500	holding	141	84.77
D1-aux500	multi-leg non-holding	41	48.32

D1 also recovers fields on auxiliary-airport-disjoint charts; this argues against attributing all performance to shared-airport cases, without establishing structural-template generalization. The three-leg/holding versus harder-slice gap also prevents extending the aggregate 75.20% to every structure.

D.3 Training-Seed Stability

With data, epoch count, hyperparameters, D1 processing, and formal200 scoring fixed, three 2B seeds give 76.18±1.36% atoms, 30.27±3.51% exact legs, and 4.67±2.02% exact procedures (mean ± sample SD). The submitted D1 remains 75.20/30.84/4.00%. This measures repeatability within the 2B backbone, not across backbones or independent test distributions. Non-three-leg and non-holding subsets yield about 11–13% exact legs but no exact procedures, versus no exact legs for the chart-blind prior.

D.4 Selected Cross-Cycle Probes

Cycle 2603 includes 30 selected charts, 10 from VOR families, with 24/30 airports outside the original window. Cycle 2607 includes 28 revised charts and 25 unchanged controls; 49/53 charts and 44/48 airports lie outside the original window. D1 recovers 23/77 exact legs and

1/30 procedures in 2603, and 14/177 legs and 0/53 procedures in 2607. The 2607 atom score is 59.01% versus a 49.60% same-set chart-blind reference. These probes use default strict comparison rather than formal200 field-specific display equivalence, so denominators cannot be pooled. The 2607 references also lack formal200-level independent validation.

E Diagnostic Evidence and Residual Errors

E.1 Source-View Diagnostic

Table 12: D1 V0–V5 source-view diagnostic; visible chart views change under the same scorer and output interface. Accuracies use the corrected formal200 reference.

Variant	Visible input	D1 accuracy
V0	full chart	75.20
V1	MA text only	11.82
V2	full chart minus MA prose	68.39
V3	plan view only	50.49
V4	icon/detail only	35.76
V5	plan/detail without MA prose	55.33

Removing MA prose leaves 68.39% atom accuracy, while the MA-text-only crop yields 11.82%. These view-dependent results do not separate evidence availability from changes in input presentation or attribute correct atoms to specific regions. V1 is an image crop passed to the visual extractor, whereas reviewed MA_TEXT in Table 13 is human-reviewed text passed to separate converters. Thus 11.82% and 67.92% are not paired measurements of one protocol.

E.2 Explicit Evidence Conversion

Table 13: Four selected rows from the strict reviewed-evidence pipeline; B3_TPD and B4_TPD share the same TPD input and differ in converter.

Setting	Input and converter	Accuracy
A3_GoldText_Rules	reviewed MA_TEXT + deterministic rules	67.92
B2a_GoldText_LLM	reviewed MA_TEXT + GPT-5.4	23.94
B3_TPD	reviewed MA_TEXT + PD candidates + GPT-5.4	28.87
B4_TPD	reviewed MA_TEXT + PD candidates + rules	67.08

Clean reviewed MA text carries useful evidence when rules convert it, but the current GPT-5.4 prompt/schema does not convert it reliably into ordered canonical output. B3_TPD and B4_TPD hold the TPD input fixed while changing converter; their gap points to a current conversion bottleneck, not useless TPD evidence. For A3, fix is 89.10%, terminator 90.34%, turn 86.76%, altitude 14.02%, course/radial 55.92%, and hold 65.11%. These differences mix converter behavior, reference normalization, and evidence granularity, rather than ranking intrinsic field difficulty.

E.3 A1+ Residuals and Complementarity

The corrected formal200 reference yields 494 wrong A1+ atoms. The frozen residual taxonomy classified the 493 errors scored before the KALN_L29 radial-direction reference correction: 47.1% as unresolved representation/normalization, 22.5% as OCR character noise, and 30.4% as text-to-leg binding ambiguity. The additional KALN_L29 leg-3 course/radial mismatch is reported separately; these aggregate categories are descriptive rather than per-instance causal attributions.

Under the same corrected reference and scorer, D1 gets 226/494 of A1+'s errors right, while A1+ gets 737/1005 of D1's errors right. A reference-visible diagnostic union yields 3784/4052 atoms (93.4%) and 44/200 procedures only as a diagnostic upper bound, not a deployed fusion system. D1 also gets 114 atoms right that both A1+ and the chart-blind prior miss; this does not identify the causal evidence for each answer.