Detecting Semantic Manipulation in Structured Academic Figures: A Benchmark and a Structured-Extraction Baseline

Fengkai Gao
Artificial Intelligence, Faculty of Science and Technology
Beijing Normal University–Hong Kong Baptist University (BNBU)
Zhuhai, China
https://github.com/QCYTSN/DSMAF_Light

***Abstract*—Academic figures increasingly carry the quantitative claims of a paper, yet a chart can be *semantically* manipulated — a bar labelled with the wrong value, a caption that names the wrong best method, a heatmap cell whose colour contradicts its number — while remaining visually plausible and free of any generation artefact. This failure class is distinct from AI-generated-image forensics, and it is not addressed by chart question answering. We introduce SciFigSem-1K, a 1000-sample synthetic benchmark spanning six semantic-manipulation types over four figure types (bar chart, line chart, heatmap, confusion matrix), built with a metadata-preserving generation pipeline and a frozen, leakage-safe evaluation protocol. On the frozen 600-item test split we evaluate three vision-language model (VLM) families — one closed-source and two open-source — under both a direct and a checklist prompt, and contrast them with a structured baseline that extracts observable figure content and applies consistency rules. The test split is class-imbalanced (120 clean / 480 manipulated), so we score with balanced accuracy, macro-F1, and MCC against trivial baselines rather than raw accuracy alone. All three VLM families detect manipulations poorly — balanced accuracy 0.58–0.66, barely above the 0.5 chance level and driven by false-negative rates frequently above 0.5 — and in raw accuracy (0.40–0.59) they fall below the trivial always-manipulated baseline (0.80), whereas the offline structured baseline reaches 0.94 accuracy (0.94 balanced). A temperature-sensitivity analysis confirms the weak VLM result is stable, not a tuning artefact. Our findings indicate that detecting semantic figure manipulation needs explicit structural reasoning rather than end-to-end VLM inference.**

***Index Terms*—scientific figure integrity, semantic manipulation detection, vision-language models, benchmark, structured extraction**

I. Introduction

Academic figures are where a paper's quantitative claims become visible. A bar height encodes an accuracy, a confusion-matrix cell encodes a count, a legend binds a colour to a method, and a caption asserts which system is best. Readers, reviewers, and increasingly automated screening tools take these figures at face value, which makes them a natural target for both honest error and deliberate misrepresentation. As automated figure screening is deployed at scale, understanding what such tools can and cannot catch becomes a practical concern for scientific integrity.

Two research directions touch this space but do not cover it. The first is forensic detection of AI-generated or tampered academic images [1]–[4], which asks whether an image was synthetically produced or pixel-edited. The second is chart understanding and question answering [5]–[9], which asks whether a model can read and reason over the content of a chart. Neither directly asks the question we study: given a figure that is genuine at the pixel level, is its *visual evidence internally consistent with the numeric and textual claims it presents*?

We call a violation of this consistency a *semantic manipulation*. A bar keeps its plotted height but its printed label is changed; a caption names a best method the chart contradicts; a legend's colour-to-method mapping is swapped while the curves are untouched; a heatmap cell's colour no longer matches its number. Such a figure carries no generation artefact and would pass a provenance-oriented detector, yet it misrepresents the result. Figure 1 illustrates the difficulty: on a bar chart whose displayed label (0.729) contradicts the true value (0.829), the closed-source VLM qwen3.7-plus reports "authentic" and remarks that the tallest bar matches the caption — an observation that is true but beside the point.

Studying this failure class requires a controlled benchmark in which every sample's ground-truth manipulation is known, so that detection can be measured exactly rather than adjudicated by hand. Building such a benchmark from real publications is hard: the manipulations are unlabelled, sparse, and legally sensitive. We therefore construct a synthetic benchmark in which figures are rendered programmatically together with their clean values, the injected manipulation, and a natural-language description of the inconsistency. This yields exact, per-sample ground truth while keeping the figure styles representative of AI/CS papers.

Using this benchmark we ask two questions. *(i) Can current VLMs, prompted to inspect a figure and its caption, detect semantic manipulations? (ii) Does an approach that first extracts the observable content of a figure and then checks it for internal consistency do better?* To make the first question robust, we evaluate three model families rather than one — a closed-source VLM and two independent open-source VLMs

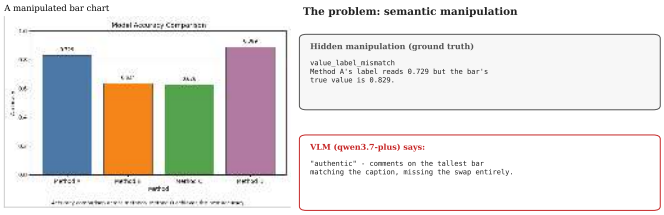

Fig. 1. A manipulated bar chart whose value label (0.729) contradicts the true bar value (0.829). The VLM (qwen3.7-plus) still calls the figure "authentic", illustrating the core problem this work studies.

— under two prompting strategies, and we probe sampling temperature to rule out a tuning artefact. To answer the second, we build a leakage-safe pipeline that extracts axes, values, legends, and cell contents and applies one consistency rule per manipulation type.

We are explicit about scope. We do not claim to introduce the first forensic figure benchmark, and we do not target universal detection of AI-generated images; our subject is semantic consistency in synthetic, structured figures of four common types. Within that scope, our contributions are:

- **SciFigSem-1K**: a 1000-sample synthetic benchmark covering six semantic-manipulation types across four figure types (bar chart, line chart, heatmap, confusion matrix), built with a metadata-preserving generation pipeline that yields exact per-sample ground truth (Sec. III–IV).
- A **frozen, leakage-safe evaluation protocol** with four baselines, in which prompts, thresholds, and extraction are tuned only on train/validation and the 600-item test split is checksum-frozen before use (Sec. V).
- A **multi-family VLM study** — one closed- and two open-source VLMs, two prompts, and a temperature-sensitivity analysis — showing that VLM detection stays weak (balanced accuracy 0.58–0.66, below the trivial baselines in raw accuracy), while a structured extraction-and-rules baseline reaches 0.94 on the same frozen split (Sec. VI).
- A **robustness and failure analysis** by manipulation type and figure type, including an honest account of where the structured baseline itself fails (Sec. VII–VIII).

Taken together, the results indicate that detecting semantic figure manipulation is better served by explicit extraction and consistency checking than by end-to-end VLM inference — a conclusion that holds across model families, prompts, and sampling temperatures.

II. Related Work

AI-generated-image and figure forensics. A rapidly growing line of work asks whether an academic image was AI-generated or tampered with. AEGIS [1] and SciFigDetect [2] benchmark forensic analysis of AI-generated academic images; Rescind [3] targets image misconduct in biomedical publications with vision-language and state-space modelling; and THEMIS [4] evaluates multimodal LLMs for scientific-paper fraud forensics. General-purpose forgery-detection suites — ForensicHub [10], GIM [11], and Forensics-Bench [12] — localise generative manipulation or splicing at the pixel level. All of these are *provenance- or artefact-oriented*: the signal is that pixels were synthesised or edited. Our setting is orthogonal. Our manipulations leave the plotted geometry intact and introduce no pixel artefact; the only evidence of manipulation is a contradiction between a claim and the visual content, which a provenance detector is not designed to see.

TABLE I
Positioning against neighbouring benchmarks. *Semantic*: targets claim/visual inconsistency rather than generation artefacts. *Structured*: focuses on structured plots (bar/line/heatmap/matrix). *Exact GT*: exact per-sample ground truth for the manipulation.

Benchmark	Semantic	Structured	Exact GT
AEGIS [1]	✗	partial	✓
SciFigDetect [2]	✗	partial	✓
ChartQA [5]	✗	✓	✓
ChartBench [6]	✗	✓	✓
SciFigSem-1K (ours)	✓	✓	✓

Chart understanding and question answering. A second body of work teaches or tests models to read charts. Chart QA benchmarks — ChartQA [5], ChartBench [6], ChartX [7], PlotQA [8], and FigureQA [9] — measure whether a model can answer questions about chart content, and instruction-tuned chart models such as ChartLlama [13] and MMC [14] improve that ability. Chart derendering recovers structured data from a plot: MatCha [15] pretrains for math reasoning and chart derendering, DePlot [16] translates a plot into a table so a language model can reason over it, and SciCap [17] studies figure captioning. Our structured baseline shares the extract-then-reason philosophy of DePlot and MatCha, but applies it to *consistency verification* rather than question answering, and crucially reads only observable fields so that gold labels cannot leak into prediction.

Vision-language models. We evaluate general-purpose VLMs as detectors rather than proposing a new one. The open-source families we test descend from lines such as Qwen2-VL / Qwen2.5-VL [18], [19] and InternVL [20], [21], which report strong perception and reasoning across resolutions. Reasoning-enhanced forensic evidence analysis [22] is complementary: it strengthens the explanation stage of detection, whereas we ask whether the detection decision itself is reachable end-to-end. Table I situates our benchmark against its closest neighbours along the axes that matter for our task.

To our knowledge, semantic-manipulation detection in structured figures — as opposed to generation-artefact forensics or chart question answering — has not previously been benchmarked in this form.

III. The SciFigSem-1K Benchmark

SciFigSem-1K contains 1000 synthetic samples: 200 clean figures and 800 with exactly one semantic manipulation, spanning four figure types (bar chart, line chart, heatmap, confusion matrix). We choose these four because they are

TABLE II
SCIFIGSEM-1K COMPOSITION. 200 CLEAN + 800 MANIPULATED OVER 4 FIGURE TYPES; SPLITS 200/200/600 (TRAIN/VAL/TEST), FROZEN AND SHA-256-CHECKSUMMED.

Property	Value
Total samples	1000
Clean / manipulated	200 / 800
Figure types	4 (bar, line, heatmap, confusion matrix)
Manipulation types	6
Train / val / test	200 / 200 / 600
Ground truth	exact (original + modified value per sample)
Split integrity	family-safe, SHA-256 frozen

the workhorses of quantitative reporting in AI and computer-science papers and because each admits a well-defined notion of internal consistency — a bar's label versus its height, a matrix cell's colour versus its count — that a manipulation can violate.

Generation pipeline. Clean figures are produced by seed-controlled generators, one per figure type, that render an image together with a structured record of everything plotted: axis ranges and ticks, per-series values, legend colour-to-label bindings, cell values, and a caption stating a claim about the data (e.g. which method is best, or the overall accuracy). Each manipulation is then injected by a dedicated operator that edits exactly one facet of a clean sample while leaving the rest visually plausible: it might rewrite a printed label without moving the bar, swap two legend bindings without re-colouring the curves, or recolour a heatmap cell without changing its number. Crucially, the operator records both the original and the modified value and writes a natural-language evidence string, so every manipulated sample carries an exact, machine-checkable ground truth rather than a hand annotation. Clean and manipulated variants that share a common source figure form a *family*.

Freezing and leakage control. We split the data into 200 train / 200 validation / 600 test samples with a fixed seed. The split is *family-safe*: all variants of a source figure are assigned to the same split, so a model cannot see a clean figure in training and its manipulated sibling at test time. Before any test-split access we record a freeze that fixes the sample and split checksums (SHA-256), the model versions, the prompt checksums, the extraction commit, and the rule thresholds; all prompt, threshold, and extraction tuning happens on train/validation only. This protocol makes the test numbers reproducible and forecloses the most common evaluation leakage. Table II summarises the composition.

IV. MANIPULATION TAXONOMY

We define six semantic-manipulation types. Each preserves an overall plausible figure and violates exactly one relationship between the visual content and a numeric or textual claim. For each we note the relationship a detector must check, which motivates the corresponding consistency rule in Sec. V.

- **value-label mismatch**: a displayed numeric label differs from the bar's true plotted value. Detection requires reading both the printed label and the bar geometry and comparing them.
- **axis-scale manipulation**: a truncated or uneven axis exaggerates a visual difference while the plotted values are unchanged. Detection requires recovering the axis range and tick spacing, not just the bar heights.
- **legend-color mismatch**: legend label/colour assignments are swapped while the plotted series keep their values. Detection requires recovering the legend's colour-to-label binding and matching it against the drawn series — the hardest relationship to recover from a single rendered image.
- **caption-figure inconsistency**: the caption asserts a claim — the best method, a trend, or an overall accuracy — that the figure contradicts. Detection requires parsing the caption claim and re-deriving it from the plotted data.
- **heatmap color-value mismatch**: a cell's colour is inconsistent with its numeric value or the colourbar. Detection requires reading both the cell number and its colour and mapping colour back to a value.
- **confusion-matrix inconsistency**: a cell's colour intensity contradicts its displayed count. Detection is analogous to the heatmap case on a matrix layout.

These six span the main ways a structured figure can lie while looking correct, and they exercise complementary detector capabilities — text reading, geometry recovery, colour-to-value inversion, and caption parsing.

V. METHOD

All baselines emit predictions under a shared contract: a binary decision (authentic / manipulated), a manipulation type from the taxonomy of Sec. IV, a natural-language evidence string, a confidence, and an explicit parse status. Fixing this contract lets us score four very different systems on the same footing and makes parse failures first-class rather than silently coerced into a default label.

VLM baselines. *vlm_direct* presents the figure image and its caption and asks the model to return the decision, type, and evidence as JSON in a single pass. *vlm_checklist* prepends a figure-type-aware checklist that walks the model through the axes, plotted values, legend, caption, colourbar, and matrix cells before it judges, in the spirit of chain-of-thought prompting for charts. Both prompts forbid unsupported provenance claims and require schema-compatible output; a response that cannot be parsed into the contract is recorded as a parse failure rather than repaired.

Structured baseline (parser_rules). This baseline never sees the model API. An OCR/CV pipeline first converts the figure into an *observable representation*: detected axes and their ranges, per-bar and per-line values, legend items, heatmap and confusion-matrix cell values, and the caption text. Cell colours are sampled from the background median rather than a single glyph pixel, which removes a large class of spurious colour readings. Over this representation we apply one consistency rule per manipulation type — for example, comparing a bar's displayed label against its geometric value,

or a matrix cell's recovered colour value against its OCR'd count — and emit the first inconsistency found, with an evidence string naming the offending element. Because the representation contains only observable fields, the rules cannot consult the ground truth.

Formally, extraction maps a figure image I and caption c to an observable representation $O = \Phi(I, c)$ containing, for each visual mark i, a displayed value d_i (read from the label or cell text) and a visual value v_i (recovered from the geometry or colour). A value-level rule flags a manipulation when displayed and visual values disagree beyond a tolerance τ calibrated on the training split:

$$\mathrm{manip}(O) \;=\; \bigvee_{i} \; \big[\, |d_i - v_i| > \tau \,\big], \tag{1}$$

where $\bigvee$ is a logical OR over marks. For colour-bearing figures the visual value is obtained by inverting the colourbar map $g(\cdot)$ on the sampled cell colour κ_i, i.e. $v_i = g^{-1}(\kappa_i)$, and the same test (1) applies. The caption rule re-derives the claimed quantity (e.g. the best method $\arg\max_i v_i$) from the plotted values and compares it against the parsed caption claim. The tolerance τ is set on train/validation so that clean figures pass while genuine manipulations are caught; it is frozen before test access. The predictor is leakage-safe by construction: Φ is defined only over observable fields, so gold labels are unreachable in (1).

Orchestrated baseline (figcheck). *figcheck* combines the two: it prefers a specific rule finding when one fires and otherwise falls back to the VLM judgement. The main-results figcheck (Table III) is this orchestrated configuration — it uses the rule decision where a rule fires and the closed-source VLM (qwen3.7-plus, checklist prompt) as the fallback otherwise — which is why its numbers differ from parser_rules (it recovers a few caption cases the rules miss, at the cost of a higher false-positive rate). Only when figcheck is run with *no* VLM available does it degenerate to the pure rule result and coincide with parser_rules; we use that fully-offline degeneration in the robustness study (Sec. VII), where no VLM is invoked.

Leakage safety. A recurring risk in figure-analysis evaluation is that a predictor reads a gold field as a feature. We enforce a hard boundary: the fields `is_manipulated`, `manipulation_type`, `ground_truth`, `manipulation`, and `evidence_text` are evaluator-only and are rejected by the observable-representation schema, so no baseline can consume them. A metadata-based detector that reads gold fields is implemented only as a separately named diagnostic and is never reported as a deployable baseline. Figure 2 summarises the structured pipeline.

VI. Experiments

Setup. We run all four baselines on the frozen 600-item test split. For the VLM baselines we evaluate three families under both prompts: one closed-source model, qwen3.7-plus (accessed via a commercial API), and two open-source models, Qwen3-VL-32B and GLM-4.5V (accessed via a hosted

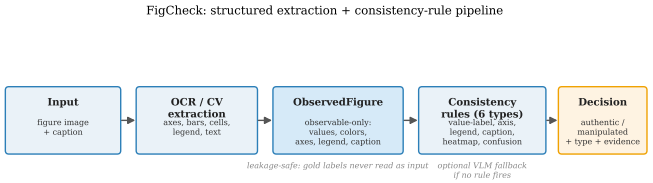

Fig. 2. FigCheck: structured extraction + consistency-rule pipeline. Gold labels are never read as predictor input (leakage-safe); an optional VLM fallback is used only when no rule fires.

OpenAI-compatible API). All VLM runs use the frozen prompts verbatim and the provider-default sampling temperature; each produces exactly one prediction per test ID, with parse failures recorded explicitly rather than silently repaired. The structured baselines re-extract the 600 test images offline (no API tokens). We report binary accuracy, macro-F1, and the false-positive and false-negative rates, together with per-manipulation-type F1 and per-figure-type accuracy. Because the split is frozen and checksummed, every run is scored against an identical reference, and all reported numbers are read programmatically from the stored per-run metrics rather than transcribed by hand.

The three VLM families were chosen to span providers and architectures rather than to find the single best model: a commercial closed-source system and two independently trained open-source systems. This design directly addresses the concern that a weak-detection result might reflect one idiosyncratic model; a consistent result across three families is far harder to dismiss.

Metrics. Let $\mathrm{TP}, \mathrm{TN}, \mathrm{FP}, \mathrm{FN}$ be the binary confusion counts, treating "manipulated" as positive. We report accuracy, the false-positive and false-negative rates, and the binary macro-F1:

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}, \qquad \mathrm{FNR} = \frac{\mathrm{FN}}{\mathrm{FN} + \mathrm{TP}}, \tag{2}$$

$$\text{macro-F1} = \tfrac{1}{2}\big(\mathrm{F1}_{\text{manip}} + \mathrm{F1}_{\text{auth}}\big). \tag{3}$$

For manipulation-type performance we compute a per-type F1 over the six types and report the unweighted mean (type macro-F1). A high FNR is the failure mode of most interest here: it means a detector passes manipulated figures as authentic.

Main results. Table III and Fig. 3 report the metrics. The offline structured baselines dominate: parser_rules reaches 0.940 accuracy (0.941 balanced, MCC 0.83) and figcheck 0.933. The VLMs are weak detectors. Because the test split is 80% manipulated, raw accuracy is misleading — a rule that always predicts "manipulated" scores 0.80 (Table IV) — so we read the VLMs through balanced accuracy and MCC. On that lens every VLM configuration lands only modestly above chance (balanced accuracy 0.58–0.66, MCC 0.13–0.27), and in raw accuracy (0.405–0.585) they fall *below* the trivial always-manipulated baseline. The checklist prompt does not

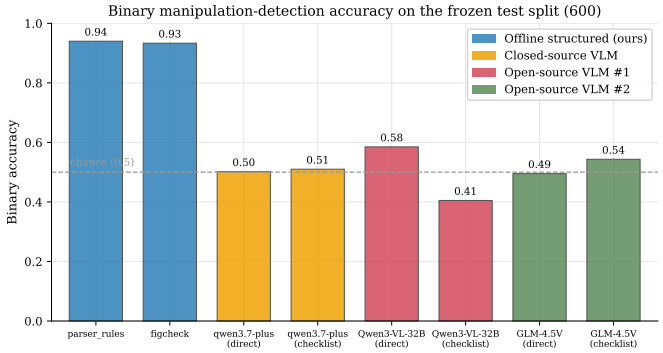

Fig. 3. Binary manipulation-detection accuracy on the frozen test split (600). Offline structured baselines reach 0.94 while all three VLM families are weak detectors (see Table III for balanced accuracy and MCC, and Table IV for trivial reference baselines).

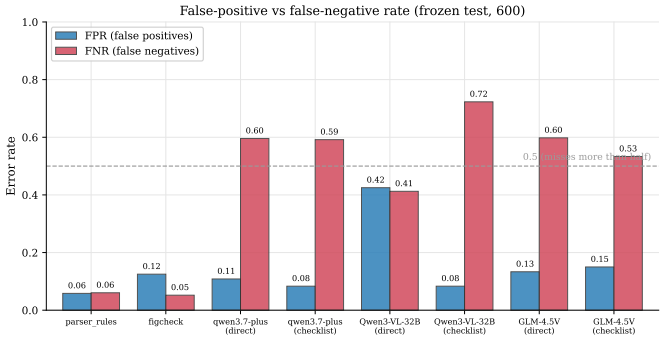

Fig. 4. False-positive vs. false-negative rate. VLMs have high FNR (they miss most manipulations); offline baselines keep both low.

consistently help over the direct prompt on this task. The result is consistent across a closed-source and two independent open-source families, so the weak VLM detection is not an artefact of a single model.

Error decomposition. Figure 4 decomposes error into false-positive and false-negative rate. The VLMs' shortfall is dominated by high false-negative rates (0.53–0.72): they miss most manipulations, typically describing a salient but irrelevant feature of the figure. The offline baselines keep both error rates low (FPR/FNR around 0.06 for parser_rules).

Per-type breakdown. Figure 5 shows per-manipulation-type F1 and Table V gives the per-type support together with parser_rules' detection and type-attribution counts. The per-type F1 in Fig. 5 is a *type-classification* score (did the baseline both flag the figure *and* name the correct manipulation type), which is stricter than binary detection. parser_rules solves value-label, axis-scale, heatmap, and confusion-matrix manipulations almost perfectly (F1 near 1.0), because those inconsistencies are recoverable from the extracted representation: the printed label, the axis geometry, and the cell colour/value are all readable, and the corresponding rule fires reliably.

It scores 0 on the type-F1 for legend-color mismatch, but this does not mean it misses every such figure: of the

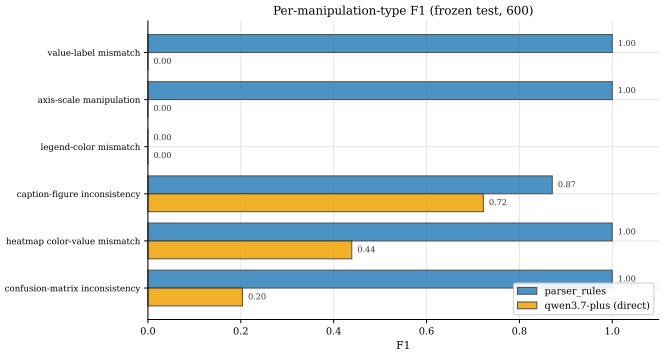

Fig. 5. Per-manipulation-type F1. parser_rules solves most types but not legend-color (F1=0, a single-PNG attribution limit shown honestly).

60 legend-color samples, it *detects* 33 as manipulated (and mislabels their type as caption-inconsistency) and misses 27 (Table V). This is why the type-F1 is 0 (no correctly-typed legend case) yet the overall FNR stays at 0.06: only 27 of the 480 manipulated figures are passed as authentic, all but two of them legend-color. Recovering a legend's colour-to-label binding from a single rendered PNG is generally not identifiable — the same colours recur in the plotted series and the mapping cannot be disambiguated without vector or interaction data — so the correct *type* is unrecoverable even when the figure is flagged. We report this zero honestly rather than suppress the type. The VLMs, by contrast, are low across *all* types, including the ones that are trivial for the structured baseline — a sign that the bottleneck is not a specific hard manipulation but the end-to-end reading-and-checking process itself.

Sensitivity to sampling temperature. A natural objection is that the weak VLM detection reflects an unlucky decoding temperature rather than a genuine capability gap. To test this we re-ran *vlm_direct* (qwen3.7-plus, frozen prompt) on a validation subset at three temperatures (Table VI). Accuracy stays in 0.38–0.45 and the false-negative rate in 0.69–0.75 across the sweep — the behaviour is flat, not a tuning artefact. We report this as an exploratory analysis on validation, kept separate from the frozen test results.

TABLE III

MAIN RESULTS ON THE FROZEN TEST SPLIT (600 SAMPLES; 120 CLEAN / 480 MANIPULATED). "MANIPULATED" IS THE POSITIVE CLASS. BALACC IS BALANCED ACCURACY $1 - \frac{1}{2}(\mathrm{FPR} + \mathrm{FNR})$; PARSEFAIL IS THE NUMBER OF UNPARSEABLE PREDICTIONS. VLM ROWS ARE SINGLE-RUN, OBSERVED; ALL NUMBERS ARE READ FROM THE STORED PER-RUN METRICS. TRIVIAL REFERENCE BASELINES (TABLE IV) REACH UP TO 0.80 RAW ACCURACY ON THIS CLASS PRIOR, SO RAW ACCURACY ALONE IS NOT THE RIGHT LENS — SEE TEXT.

Baseline	Group	Acc.	BalAcc	macro-F1	MCC	FPR	FNR	ParseFail
parser_rules	offline	**0.940**	**0.941**	**0.912**	**0.829**	0.058	0.060	0
figcheck (orch.)	offline+VLM	0.933	0.911	0.899	0.799	0.125	0.052	0
qwen3.7-plus (dir)	closed	0.502	0.648	0.491	0.249	0.108	0.596	0
qwen3.7-plus (chk)	closed	0.510	0.662	0.500	0.274	0.083	0.592	0
Qwen3-VL-32B (dir)	open	0.585	0.581	0.525	0.131	0.425	0.412	8
Qwen3-VL-32B (chk)	open	0.405	0.597	0.404	0.182	0.083	0.723	0
GLM-4.5V (dir)	open	0.495	0.634	0.484	0.226	0.133	0.598	8
GLM-4.5V (chk)	open	0.543	0.658	0.524	0.258	0.150	0.533	5

TABLE IV

TRIVIAL REFERENCE BASELINES ON THE TEST CLASS PRIOR (120 CLEAN / 480 MANIPULATED). BECAUSE THE SPLIT IS 80% MANIPULATED, ALWAYS PREDICTING "MANIPULATED" ALREADY REACHES 0.80 RAW ACCURACY, WHILE EVERY TRIVIAL RULE HAS BALANCED ACCURACY 0.5. BALANCED ACCURACY AND MCC ARE THEREFORE THE HONEST LENS FOR TABLE III.

Reference rule	Acc.	BalAcc
always "manipulated"	0.800	0.500
always "authentic"	0.200	0.500
random 50/50	0.500	0.500
prior-matched random	0.680	0.500

TABLE V

PER-MANIPULATION-TYPE SUPPORT ON THE TEST SPLIT AND PARSER_RULES' BEHAVIOUR. *Detected* = FLAGGED MANIPULATED (BINARY HIT); *Typed* = ALSO ASSIGNED THE CORRECT TYPE. LEGEND-COLOR IS DETECTED IN 33/60 CASES BUT NEVER CORRECTLY TYPED, WHICH IS WHY ITS TYPE-F1 IS 0 WHILE OVERALL DETECTION STAYS HIGH.

Manipulation type	Support	Detected	Typed
value-label mismatch	60	60	60
axis-scale manipulation	60	60	60
legend-color mismatch	60	33	0
caption-figure inconsistency	120	118	118
heatmap color-value mismatch	90	90	90
confusion-matrix inconsistency	90	90	90
total (manipulated)	480	451	418

TABLE VI

TEMPERATURE SENSITIVITY OF *vlm_direct* (QWEN3.7-PLUS) ON A VALIDATION SUBSET. NEAR-CHANCE BEHAVIOUR IS STABLE ACROSS TEMPERATURES (EXPLORATORY; NOT MIXED INTO THE FROZEN TEST RESULTS).

Temperature	Accuracy	FPR	FNR
0.0	0.383	0.083	0.750
0.5	0.417	0.083	0.708
1.0	0.450	0.000	0.688

VII. ROBUSTNESS

We test how the structured baseline degrades under four image corruptions applied to a 120-sample stratified test subset: JPEG compression, downsampling, a screenshot-style resize, and slight Gaussian blur (Fig. 6; parser_rules and figcheck coincide offline, since figcheck degenerates to the rule result without a VLM). Accuracy is robust to JPEG (0.94) and downsampling (0.93) relative to the clean-subset reference (0.958), but drops under screenshot-resize (0.82) and slight blur (0.77). The mechanism is OCR degradation: blur shifts recognised digits, which perturbs the recomputed values that the consistency rules depend on. We keep this study offline (no VLM robustness sweep), since the VLMs are already weak detectors on clean images.

VIII. FAILURE ANALYSIS

Figure 7 shows representative cases with evidence quoted verbatim from the committed records; we separate what is observed in the record from the inferred explanation throughout.

When the structured baseline succeeds. In the top row, parser_rules reads a confusion-matrix cell displaying the count 15 whose background colour corresponds to a value of 76, and emits a specific evidence string naming the offending cell. This is the common case for the four types the baseline solves: the relevant quantities are legible, so the corresponding rule fires with a localised, checkable justification — exactly the kind of evidence a human screener can verify.

Why the VLM misses. In the middle row, a bar's printed label (0.729) contradicts its true value (0.829); qwen3.7-plus declares the figure authentic and comments that the tallest bar matches the caption. The observation is factually correct but irrelevant to the manipulation. This pattern recurs across the VLM's false negatives: the model attends to a globally salient feature and does not perform the local cross-check — label against geometry, colour against value — that the task requires. The high false-negative rates in Fig. 4 are the aggregate of this behaviour.

Where the structured baseline itself fails. The bottom row is a legend-color mismatch: the extractor cannot recover the legend's colour-to-label binding from a single PNG, no rule fires, and the sample becomes a false negative. This is

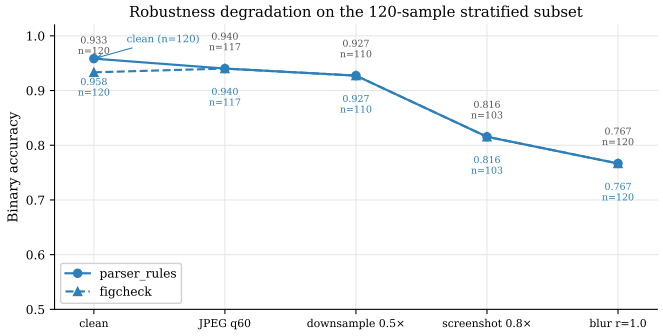

Fig. 6. Robustness degradation on the 120-sample stratified subset. Robust to JPEG/downsampling; degrades under screenshot-resize and blur.

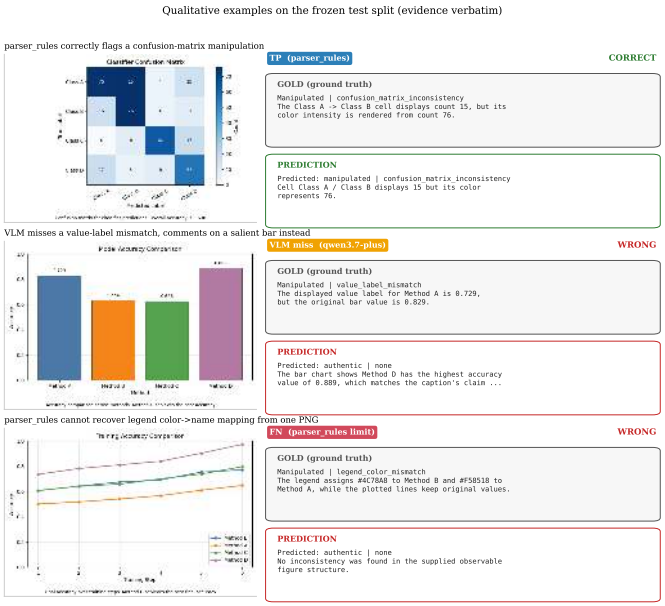

Fig. 7. Qualitative examples on the frozen test split (evidence quoted verbatim from committed records): a structured-baseline success (top), a VLM miss (middle), and a structured-baseline limitation on legend-color (bottom).

the dominant parser_rules error and the direct cause of the F1=0 on that type in Fig. 5. It is a limitation of single-image extraction, not of the consistency-checking idea: with vector or interaction data the mapping would be recoverable.

IX. Discussion and Limitations

Why do VLMs struggle here? The task is not perception in isolation — the VLMs can describe the figures — but the conjunction of reading a specific quantity, re-deriving what it should be, and reporting the mismatch. Our per-type and qualitative evidence suggests the models default to a global, caption-consistent reading and rarely perform the local verification the task rewards, which is consistent with the flat temperature sweep in Table VI: turning the sampling temperature up or down does not induce the missing verification step. The structured baseline succeeds precisely because it makes that step explicit — extract the quantity, recompute, compare — which is also why it fails cleanly and diagnosably when extraction is impossible (legend-color).

Implications. For automated figure screening, this argues for a decompose-then-check design over end-to-end VLM judgement: extraction makes both the decision and its evidence auditable, and its failures are localised to identifiable extraction gaps rather than diffuse model behaviour. A VLM is still useful as a fallback when structured extraction is not identifiable, which is what our figcheck orchestration does.

Limitations. (i) The VLM main results are single-run at default parameters and one prompt version per style; the temperature sweep mitigates but does not exhaust this. (ii) Data are synthetic; real published figures are more varied, so our conclusion that structured extraction is preferable is established *on synthetic structured figures* and not yet on real publications. (iii) Robustness is studied offline on a subset. (iv) legend-color detection is capped by single-PNG attribution. (v) One candidate open-source model (a reasoning model) was excluded on efficiency grounds, and one open-source model produced $\sim$1% empty outputs, recorded as explicit parse failures (reported per run in Table III).

X. Conclusion

We introduced SciFigSem-1K, a benchmark for semantic-manipulation detection in structured academic figures, and used it to compare three VLM families with a structured extraction-and-rules baseline. VLMs are weak detectors — balanced accuracy 0.58–0.66 and below the trivial always-manipulated baseline in raw accuracy — while the structured baseline reaches 0.94, and the gap holds across families, prompts, and sampling temperatures. On these synthetic structured figures, the result suggests that detecting semantic figure manipulation is better served by explicit extraction and consistency checking than by end-to-end VLM inference. Extending the extraction to recover legend mappings, and moving from synthetic to real figures, are natural next steps.

References

[1] "Aegis: A holistic benchmark for evaluating forensic analysis of ai-generated academic images," 2026. [Online]. Available: https://arxiv.org/abs/2604.28177

[2] "Scifigdetect: A benchmark for ai-generated scientific figure detection," 2026. [Online]. Available: https://arxiv.org/abs/2604.08211

[3] "Rescind: Countering image misconduct in biomedical publications with vision-language and state-space modeling," 2026. [Online]. Available: https://arxiv.org/abs/2601.08040

[4] "Themis: Towards holistic evaluation of mllms for scientific paper fraud forensics," 2026. [Online]. Available: https://arxiv.org/abs/2603.25089

[5] "Chartqa: A benchmark for question answering about charts with visual and logical reasoning," 2022. [Online]. Available: https://arxiv.org/abs/2203.10244

[6] "Chartbench: A benchmark for complex visual reasoning in charts," 2023. [Online]. Available: https://arxiv.org/abs/2312.15915

[7] "Chartx & chartvlm: A versatile benchmark and foundation model for complicated chart reasoning," 2024. [Online]. Available: https://arxiv.org/abs/2402.12185

[8] "Plotqa: Reasoning over scientific plots," 2019. [Online]. Available: https://arxiv.org/abs/1909.00997

[9] "Figureqa: An annotated figure dataset for visual reasoning," 2017. [Online]. Available: https://arxiv.org/abs/1710.07300

[10] "Forensichub: A unified benchmark & codebase for all-domain fake image detection and localization," 2025. [Online]. Available: https://arxiv.org/abs/2505.11003
[11] "Gim: A million-scale benchmark for generative image manipulation detection and localization," 2024. [Online]. Available: https://arxiv.org/abs/2406.16531
[12] "Forensics-bench: A comprehensive forgery detection benchmark suite for large vision language models," 2025. [Online]. Available: https://arxiv.org/abs/2503.15024
[13] "Chartllama: A multimodal llm for chart understanding and generation," 2023. [Online]. Available: https://arxiv.org/abs/2311.16483
[14] "Mmc: Advancing multimodal chart understanding with large-scale instruction tuning," in *Proceedings of NAACL 2024*, 2024. [Online]. Available: https://aclanthology.org/2024.naacl-long.70
[15] "Matcha: Enhancing visual language pretraining with math reasoning and chart derendering," 2022. [Online]. Available: https://arxiv.org/abs/2212.09662
[16] "Deplot: One-shot visual language reasoning by plot-to-table translation," 2022. [Online]. Available: https://arxiv.org/abs/2212.10505
[17] "Scicap: Generating captions for scientific figures," 2021. [Online]. Available: https://arxiv.org/abs/2110.11624
[18] "Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution," 2024. [Online]. Available: https://arxiv.org/abs/2409.12191
[19] "Qwen2.5-vl technical report," 2025. [Online]. Available: https://arxiv.org/abs/2502.13923
[20] "How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites," 2024. [Online]. Available: https://arxiv.org/abs/2404.16821
[21] "Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models," 2025. [Online]. Available: https://arxiv.org/abs/2504.10479
[22] "Reveal: Reasoning-enhanced forensic evidence analysis for explainable ai-generated image detection," 2025. [Online]. Available: https://arxiv.org/abs/2511.23158