INFOAGENT: TRACEABLE GENERATION AND REPAIR OF EVIDENCE-GROUNDED INFOGRAPHICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Reliable infographic generation requires facts, symbols, and visual relations to remain consistent through rendering and revision. Correcting one element also requires tracking its supporting evidence and the dependencies affected by the change. We present **InfoAgent**, a training-free framework for *evidence-bound visual-symbolic program synthesis*. Its Infographic Visual Description (IVD) records factual payloads, evidence provenance, execution routes, and verification obligations in a typed dependency graph. Retrieved design priors guide compilation, and layered execution combines raster synthesis with editable symbolic and binding objects while retaining their traces. Dependency-aware repair localizes corrections, rechecks affected dependencies, and requires protected obligations to remain satisfied under the declared checkers. Unresolved obligations remain explicit. On IGenBench, InfoAgent achieves 93.0 Q-ACC and 59.0 I-ACC. We also introduce InfoGraphicBench-Evidence, where complete-checklist pass rates on 200 test requests increase from 21.5% for Same-IVD Prompt to 23.5% for the initial layered output and 28.5% after repair, using the same evidence and initial IVD. On 120 audited repair cases, localized repair edits 12.4% of the canvas on average, compared with 67.3% for global regeneration.

1 INTRODUCTION

Infographics combine factual claims, numerical data, and visual explanations on one canvas. In journalism, education, and public communication, their usefulness depends on readable design and accurate information Feng et al. (2026a); Cui et al. (2019); Tyagi et al. (2022). A polished composition can still mislead when a value is unsupported, a source note is missing, or a label refers to the wrong object Wang et al. (2019); Vu et al. (2025); Tang et al. (2026). Reliable generation requires preserving relationships among evidence, text, and visual elements, not just their appearance.

Infographic research now spans document-conditioned generation, authoring tools, chart resources, and reliability evaluation Li et al. (2026b); Ghosh et al. (2025); Tang et al. (2026). Image models offer increasingly capable synthesis and text rendering Esser et al. (2024); Labs (2024); Betker et al. (2023); Wu et al. (2025); Cai et al. (2025a;b); Team et al. (2025). Research agents acquire and organize external knowledge Jin et al. (2025); Zheng et al. (2025b); Yang et al. (2026b); Chng et al. (2025), and image-generation agents use search, references, planning, and feedback to improve contextual and compositional fidelity Zhang et al. (2026); Feng et al. (2026b); Chen et al. (2026b;a); Li et al. (2026a). Structured execution and local correction are also established approaches: InfoGen converts document-derived metadata into infographic code Ghosh et al. (2025), GenClaw uses executable visual sketches before raster synthesis Ye et al. (2026), and M3 checks constraints and validates targeted image edits Yang et al. (2026a).

We study how to preserve evidence and execution dependencies as an infographic is revised. Consider a percentage callout whose text is correct but whose arrow points to the wrong chart segment. Correcting the arrow requires locating its intended target; moving the callout may obscure a neighboring value or source note. An editable object provides a location for the change, but does not by itself specify the supporting evidence or the dependent requirements that need rechecking. A reliable repair needs these relationships to remain explicit across planning, rendering, and verification.

We introduce **InfoAgent**, a training-free framework for *evidence-bound visual-symbolic program synthesis*. Its intermediate representation, **Infographic Visual Description** (IVD), records content,

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.

evidence spans, spatial regions, execution routes, bindings, and verification obligations in a typed dependency graph. The same element identifier connects a factual claim to its supporting span, rendered object, and post-composition checks. This evidence-linked form of *element addressability* allows a detected failure to be traced to its content, geometry, or local relation. Retrieved layout and style priors guide the design, while factual content comes from supplied or retrieved evidence. All foundation models remain frozen.

As shown in Figure 1, InfoAgent routes open-ended imagery to a visual generator, exact text and data graphics to symbolic rendering, and arrows, callouts, and legend associations to binding execution. Composition retains rendering-tree handles for symbolic objects and region-level or grounded associations for raster elements. Checkers use these traces to associate each PASS, FAIL, or UNKNOWN decision with an element and a diagnostic witness. Repair follows the dependency graph to identify affected elements and relevant checks. Candidate patches must preserve previously passed critical obligations and all passed obligations outside the target's dependency closure under the declared checkers. The resulting certificate records these checks and unresolved cases; semantic verification remains model-assisted.

We evaluate InfoAgent on the complete IGenBench and introduce **InfoGraphicBench-Evidence**, a 260-request benchmark with frozen evidence bundles and held-out annotations, including 200 test requests. InfoAgent achieves 93.0 Q-ACC and 59.0 I-ACC on IGenBench. On InfoGraphicBench-Evidence, the complete-checklist pass rate (Full) rises from 21.5% for Same-IVD Prompt to 23.5% for the initial layered output and 28.5% after repair. These comparisons use the same evidence and initial IVD. On the audited repair subset, localized repair changes 12.4% of the canvas on average, compared with 67.3% for global regeneration. The results support the value of trace-preserving execution and scoped repair, while the remaining failures show that evidence quality and semantic verification continue to limit reliability.

We introduce an evidence-bound IVD that connects source spans, rendered elements, local relations, and verification obligations. Building on this representation, InfoAgent combines layered execution with dependency-aware repair to localize corrections, recheck affected content, and record unresolved requirements. We also construct InfoGraphicBench-Evidence and evaluate whole-infographic reliability, output quality, and repair locality alongside IGenBench.

2 RELATED WORK

Agentic image generation. Modern text-to-image models offer strong visual synthesis and increasingly capable instruction following and text rendering Esser et al. (2024); Labs (2024); Betker et al. (2023); Wu et al. (2025); Cai et al. (2025a;b). Recent agentic systems extend these models with external knowledge, tool use, and iterative control. Gen-Searcher retrieves textual evidence and visual references before generation Feng et al. (2026b), while Qwen-Image-Agent constructs missing generation context through planning, reasoning, search, memory, and feedback Zhang et al. (2026). M3 decomposes complex requests into checkable constraints and revises failed components Yang et al. (2026a); related approaches use grounded recaptioning, tool trajectories, specialized agents, or multi-round reasoning to improve factual and compositional generation Chen et al. (2026a;b); Li et al. (2026a); Bian et al. (2026). GenClaw further introduces executable SVG, HTML, and code-based canvases before raster synthesis Ye et al. (2026). InfoAgent complements these directions by binding each information-bearing element to evidence, an execution route, local relations, and checkable obligations. The resulting traces support element-indexed verification and scoped re-execution after composition, rather than treating the generated artifact only as a holistic image.

Structured visual communication. Deep-research and presentation systems organize evidence into multimodal reports or slides using visualization descriptions, reference layouts, editable code, and iterative refinement Zheng et al. (2025b); Yang et al. (2026b); Zheng et al. (2025a); Tang et al. (2025); Zheng et al. (2026); Zeng et al. (2026); Xu et al. (2025). Infographic research has progressed from constrained linguistic or tabular inputs to automatic data stories, visualization code, and AI-assisted authoring Cui et al. (2019); Wang et al. (2019); Shi et al. (2020); Tyagi et al. (2022); Dibia (2023); Vu et al. (2025). Recent work contributes large-scale infographic-chart resources Li et al. (2026b), document-conditioned statistical infographic generation Ghosh et al. (2025), reliability evaluation Tang et al. (2026), and narrative-centered co-creation Feng et al. (2026a). These

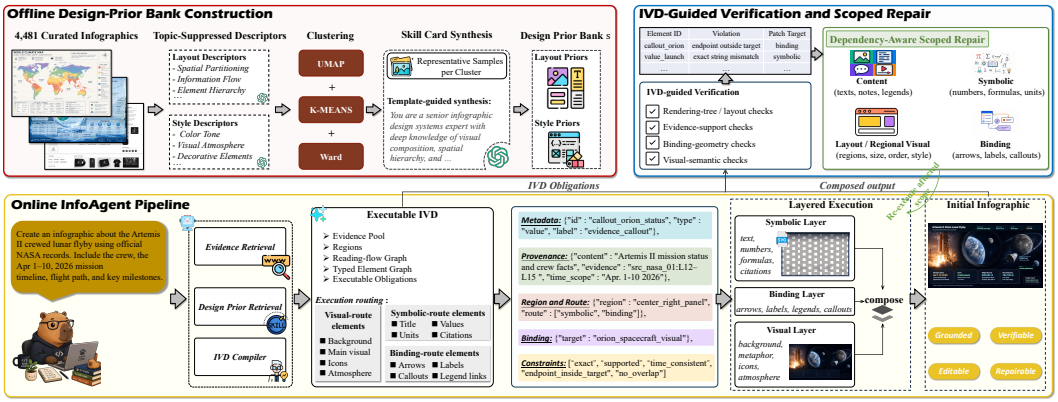

Figure 1: Overview of InfoAgent. A user request, retrieved evidence, and design priors are compiled into a typed IVD. Each information-bearing element retains evidence provenance, a spatial region, an execution route, binding dependencies, and checkable obligations. Visual, symbolic, and binding layers are composed with execution traces. IVD-guided verification then maps element-indexed violations to scoped content, symbolic, binding, layout, or visual patches.

systems provide complementary mechanisms for content selection, page organization, visual encoding, rendering, and assessment. InfoAgent focuses on open-ended, evidence-grounded single-canvas infographics, where factual claims, exact symbols, visual objects, and local bindings must remain jointly traceable. IVD connects these stages through an executable element graph, combining open-ended visual synthesis with editable symbolic and binding execution.

3 METHOD

We propose **InfoAgent**, a training-free framework that formulates knowledge-intensive infographic generation as *evidence-bound visual-symbolic program synthesis*. Given a user request x, an evidence pool $\mathcal{E}$, and a retrieved design prior s, InfoAgent compiles them into an executable IVD and produces the composed infographic together with its execution trace:

$$z = F_{\text{ivd}}(x, \mathcal{E}, s), \qquad (y, T) = \text{Exec}(z), \tag{1}$$

where z is the compiled IVD, y is the composed infographic, and T records correspondences between IVD elements and rendered objects or regions. All foundation models and rendering tools remain frozen. Figure 1 summarizes the pipeline from IVD compilation to scoped repair.

Rather than serving as a longer prompt, IVD makes each information-bearing element addressable by recording its payload, evidence provenance, spatial role, execution route, binding and dependency relations, and post-composition obligations. InfoAgent executes these elements through visual, symbolic, and binding layers, preserving traces for checking and repair. This separation allows an incorrect value, source note, or binding relation to be re-executed without necessarily regenerating the complete artifact.

Verification is defined relative to the obligations declared by IVD. An output is marked as certified relative to z only when every critical obligation returns PASS; a critical UNKNOWN remains unresolved. The certificate records checker statuses, confidence values, and witnesses for the declared content, layout, and binding obligations. It provides an operational account of these declared requirements rather than a guarantee of open-world factual correctness.

3.1 EVIDENCE AND DESIGN PRIORS

InfoAgent retrieves two complementary forms of context. The *evidence retriever* collects passages, numerical values, dates, entity names, and source metadata. Retrieved documents are deduplicated and segmented into addressable spans. Every non-decorative factual unit must be linked to at least one span; otherwise, it is marked `blocked` and is excluded from execution.

The *design retriever* supplies layout and style constraints without changing the factual payload. We construct a non-parametric prior bank from 4,481 curated infographics. Topic-suppressed layout and style descriptions are embedded and clustered using UMAP, K-Means, and Ward merging McInnes et al. (2018); McQueen (1967); Ward Jr (1963). Representative samples are summarized into skill cards that describe information structures, region patterns, density budgets, binding patterns, and style constraints. At inference time, retrieved cards guide IVD compilation, while evidence coverage and symbolic readability take precedence over aesthetic preferences. Bank construction and retrieval details are provided in the supplementary material.

3.2 IVD AS AN EXECUTABLE CONTRACT

Typed element graph. IVD is represented as $z = (\mathcal{E}, \mathcal{U}, \mathcal{R}, G, \Omega)$, where $\mathcal{E}$ is the evidence pool, $\mathcal{U}$ defines canvas regions, $\mathcal{R}$ specifies reading order, Ω contains executable obligations, and $G = (\mathcal{A}, \mathcal{D})$ is a typed element dependency graph. Each element $a_i \in \mathcal{A}$ follows the compact schema $a_i = (m_i, p_i, r_i, b_i, c_i)$, where m_i stores its identifier, type, role, and priority; p_i contains an exact or paraphrasable payload together with its supporting evidence spans; r_i specifies its region and execution route; b_i records local binding relations; and $c_i \subseteq \Omega$ indexes the obligations applied to the element. Typed edges in $\mathcal{D}$ encode `contains`, `precedes`, `supports`, `binds-to`, and `depends-on` relations, allowing changes to an element to trigger the re-execution or re-checking of its affected dependents.

Executable obligations. Each obligation is represented as $\omega_j = (S_j, \phi_j, \nu_j, \gamma_j)$, where S_j contains the scoped object–field identifiers to which the obligation applies, ϕ_j is the required predicate, ν_j is the checker, and γ_j denotes its severity. For example, a value element may require exact string equality, support from a cited evidence span, minimum contrast, and an arrow endpoint within a designated target region. Terms such as `exact`, `support`, and `no-overlap` therefore refer to checkable predicates with explicit scopes and repair targets, rather than descriptive tags.

Compilation and capability-constrained routing. The compiler decomposes the requested message into claims, values, dates, formulas, labels, citations, and visual intents. Exact strings are copied from evidence into protected payload fields, while paraphrasable claims retain their evidence spans. It then instantiates regions, reading order, element dependencies, local bindings, and executable obligations.

Each element is assigned to the smallest combination of layers capable of satisfying its obligations. Let $\mathcal{L} = \{\mathrm{vis}, \mathrm{sym}, \mathrm{bind}\}$, let $\mathrm{req}(a_i)$ denote the capabilities required by element a_i, and define $\mathrm{cap}(L) = \bigcup_{\ell \in L} \mathrm{cap}(\ell)$. The execution route is selected by

$$\begin{aligned} r_i^\star = \arg \min_{\varnothing \neq L \subseteq \mathcal{L}} \big(|L|, \mathrm{cost}(L)\big) \\ \text{s.t.} \quad \mathrm{req}(a_i) \subseteq \mathrm{cap}(L), \end{aligned} \tag{2}$$

where the objective is minimized lexicographically, first by the number of layers and then by execution cost. Since $\mathcal{L}$ contains only three layers, the implementation enumerates its seven non-empty subsets. Exact strings, numerical values, formulas, and vector geometry require symbolic execution; arrows, callouts, and object–label relations require binding execution; and open-ended appearance requires visual synthesis. Multiple layers are assigned only when no single layer covers all obligations. The capability matrix and execution-cost order are provided in the supplementary material.

3.3 TRACE-PRESERVING LAYERED EXECUTION

Each execution layer returns both a rendered artifact and its trace, while composition preserves their element-level or region-level associations:

$$\begin{aligned} (I_\ell, T_\ell) &= R_\ell(z_\ell), \\ (y, T) &= \mathrm{Compose}(\{(I_\ell, T_\ell)\}_{\ell \in \mathcal{L}}), \end{aligned} \tag{3}$$

where z_ℓ contains the IVD elements routed to layer ℓ, and T is the merged execution trace. Symbolic and binding elements retain exact rendering-tree handles, including their payloads, bounding boxes,

z-order, and anchor geometry. Raster elements retain region-level traces from their assigned generation regions. When a binding or checker requires an object-level target, an independent grounding parser associates the corresponding IVD element with a box or mask and returns a confidence score; low-confidence associations are marked UNKNOWN. Traces are therefore exact for editable vector objects and region- or grounding-based for raster content.

The *visual layer* synthesizes backgrounds, visual metaphors, textures, and non-critical icons while preserving IVD-reserved symbolic regions. The *symbolic layer* renders exact text, values, formulas, charts, legends, and source notes as editable SVG objects. The *binding layer* renders arrows, leader lines, callouts, local labels, and legend–object mappings over resolved anchors. Its geometry is deterministic once the source and target anchors are established, while anchor resolution for raster objects may remain model-assisted. Further details on trace construction and anchor resolution are provided in the supplementary material.

3.4 VERIFICATION AND SCOPED REPAIR

Three-valued verification. For each obligation ω_j, the corresponding checker returns $\nu_j(y, T, \mathcal{E}) = (s_j, \kappa_j, w_j)$, where $s_j \in \{\text{PASS}, \text{FAIL}, \text{UNKNOWN}\}$ is its status, κ_j is the confidence, and w_j is a diagnostic witness. Rendering-tree checkers evaluate exact strings, values, formulas, clipping, overlap, contrast, z-order, and endpoint geometry. Evidence checkers compare claims with their cited spans, while visual-semantic checkers inspect only the object or crop identified by T. Model-assisted checkers use fixed prompts and deterministic decoding, with pass and fail thresholds calibrated on a held-out human-labeled set; scores between the thresholds yield UNKNOWN.

$\mathcal{C}(y, z) = \{(\text{id}(\omega_j), s_j, \kappa_j, w_j, \text{ver}(\nu_j))\}_{\omega_j \in \Omega}$ is the resulting certificate. It records the status of each obligation together with the checker version and its witness. Witnesses may include cited evidence spans, expected and rendered strings, conflicting boxes, or unresolved target anchors. Obligations returning FAIL, together with critical obligations returning UNKNOWN, are emitted as element-indexed violations for subsequent scoped repair.

Transactional repair. A violation $q = (\omega, s, \kappa, w)$ identifies a field and patch class. The dependency graph induces an impact closure $H(q)$ containing elements that may need to be changed or rechecked. Candidate patches are drawn from a typed operator library, including evidence-backed payload replacement, symbolic re-rendering, constrained layout reflow, anchor reassignment, masked visual editing, and regional regeneration. The repair scope expands from field to element, region, and, when necessary, global recompilation.

$\text{Scope}(H) = \bigcup_{a_i \in H} \text{scope}(a_i)$ collects the object–field identifiers affected by a repair. The protected set is $P(H) = \{\omega : s(\omega) = \text{PASS}, S_\omega \cap \text{Scope}(H) = \varnothing\}$, containing previously passed obligations outside this scope.

Let F^{crit} and U^{crit} denote the numbers of critical failures and critical unknowns, and let F and U denote their totals over all obligations. Candidate certificates are ranked by

$$Q^\star(\mathcal{C}) = \big(F^{\text{crit}} + U^{\text{crit}}, F + U, F^{\text{crit}}, F\big). \tag{4}$$

A candidate is admissible only if the targeted critical obligation becomes PASS, or the total number of critical non-pass obligations strictly decreases. No other critical obligation may change from PASS to a non-pass state, and every obligation in $P(H)$ must remain passed. Among admissible candidates, InfoAgent minimizes $Q^\star$, followed by the number of changed elements, edited canvas area, and tool cost. If no candidate satisfies these conditions, the current candidate is rolled back and the repair scope is expanded.

This acceptance rule provides checker-relative preservation for previously passed obligations outside the affected dependency scope. It does not exclude changes within that scope or errors not captured by the declared checkers. Repair stops when all critical obligations pass, no admissible patch is found, or the maximum of K rounds is reached; unresolved obligations remain explicit in the returned certificate. The patch library, dependency propagation rules, and full repair algorithm are provided in the supplementary material.

Table 1: Reliability on the complete IGenBench benchmark.

Method	Q-ACC ↑	I-ACC ↑	Out ↑
Benchmark-reported references Tang et al. (2026)			
NanoBanana-Pro Google (2025)	90.0	49.0	–
Seedream-4.5 ByteDance (2025)	61.0	6.0	–
GPT-Image-1.5 OpenAI Team (2025)	55.0	12.0	–
Controlled baselines			
Direct T2I Wu et al. (2025)	43.0	2.0	100.0
Same-IVD Prompt	50.0	5.0	100.0
Gen-Searcher Feng et al. (2026b)	76.0	20.0	100.0
GenClaw Ye et al. (2026)	79.0	27.0	98.2
LLM-to-SVG/HTML	82.0	32.0	96.3
InfoAgent w/o repair	88.0	45.0	100.0
InfoAgent	**93.0**	**59.0**	100.0

4 Experiments

4.1 Experimental Setup

We evaluate atomic and whole-infographic reliability, executable IVD beyond prompt serialization, the effectiveness and locality of scoped repair, and the trade-offs among reliability, visual quality, output coverage, and cost.

Benchmarks. **IGenBench** Tang et al. (2026) contains 600 prompts, 30 infographic types, and 5,259 atomic questions from ten reliability categories. We use the benchmark under its official one-output protocol; the generation pipeline receives only the prompt and never accesses benchmark questions, category labels, evaluator prompts, or outputs. **InfoGraphicBench-Evidence** contains 260 requests across six knowledge domains, split into 30 development, 30 calibration, and 200 topic-disjoint test instances. Each instance provides a request and evidence bundle, with held-out annotations for factual and symbolic requirements and text–visual bindings. Test instances contain 8.4 requirements, 4.7 required text units, 2.3 numerical or source-note units, and 2.6 bindings on average. Construction, deduplication, and annotation details appear in the supplementary material.

Baselines. **Direct T2I** uses Qwen-Image Wu et al. (2025) to render the original request without external evidence, planning, verification, or repair. **RAG Prompt**, used only on InfoGraphicBench-Evidence, serializes the shared evidence into a generation prompt; **Same-IVD Prompt** instead serializes the automatically compiled IVD before verification or repair and delegates complete rendering to the same image generator. **Gen-Searcher** Feng et al. (2026b) and **GenClaw** Ye et al. (2026) are rerun under our controlled protocol while retaining their native search/planning and code-driven execution strategies, respectively. **LLM-to-SVG/HTML** renders the complete infographic as executable vector or web code using the same language-model backbone. **InfoAgent w/o repair** is an internal $K = 0$ variant that retains IVD compilation and layered execution but applies no verifier-triggered patch. Benchmark-reported IGenBench rows are shown only as external reference points; adaptation and execution details are provided in the supplementary material.

Metrics. On IGenBench, **Q-ACC** averages atomic-question accuracy, while **I-ACC** requires every question for an infographic to be correct. On InfoGraphicBench-Evidence, **Out** measures valid-output coverage, **ReqCov** averages requirement satisfaction, and **Full** requires the external checklist to pass without additional critical errors. **NonSup.** is the percentage of factual claims unsupported or contradicted by evidence; **ReqText-F1** matches required and recognized strings character-wise; **Bind-Q** rates local text–visual grounding on a 1–5 scale; and **Q-Align** Wu et al. (2023) measures perceptual quality. Table 3 retains the latter three metrics and adds mechanism-sensitive endpoints: **CritPass** requires all critical factual, symbolic, and binding obligations to pass, **Fact** counts omitted required facts as failures, and **Layout-Q** rates region organization, reading flow, hierarchy, overlap, and crowding. Formal definitions are provided in the supplementary material.

Table 2: Controlled results on 200 InfoGraphicBench-Evidence test requests. Evidence-conditioned methods use the same evidence bundle; Direct T2I is evidence-free.

Method	Out ↑	ReqCov ↑	Full ↑	NonSup. (%) ↓	ReqText-F1 ↑	Bind-Q ↑	Q-Align ↑
Direct T2I Wu et al. (2025)	100.0	76.74	17.5	1.63	82.7	4.53	4.508
RAG Prompt	100.0	79.40	19.5	1.18	83.3	4.56	4.515
Same-IVD Prompt	100.0	81.00	21.5	1.02	86.4	4.58	4.522
Gen-Searcher Feng et al. (2026b)	100.0	82.10	22.5	0.87	85.4	4.61	4.556
GenClaw Ye et al. (2026)	98.0	81.50	25.0	0.76	92.7	4.60	4.048
LLM-to-SVG/HTML	96.5	79.61	23.0	0.80	94.7	4.58	3.920
InfoAgent w/o repair	100.0	81.92	23.5	0.89	92.4	4.47	4.502
InfoAgent	100.0	**83.54**	**28.5**	**0.57**	**96.5**	**4.64**	**4.621**

Table 3: Ablations on InfoGraphicBench-Evidence. ReqText-F1, Bind-Q, and Q-Align are shared with Table 2; CritPass, Fact, and Layout-Q isolate critical reliability, required-fact preservation, and layout quality.

Variant	CritPass ↑	Fact ↑	Layout-Q ↑	ReqText-F1 ↑	Bind-Q ↑	Q-Align ↑
Full InfoAgent	**61.5**	**91.6**	**4.45**	**96.5**	**4.64**	**4.621**
InfoAgent w/o repair	46.0	88.0	4.30	92.4	4.47	4.502
Flat plan	42.5	86.8	4.13	88.1	4.24	4.389
Symbolic content to raster	38.5	90.1	4.43	85.2	4.55	4.529
w/o evidence binding	49.0	84.2	4.44	92.9	4.58	4.515
w/o binding route	45.0	89.9	4.42	93.0	4.19	4.511
w/o design-prior retrieval	54.0	89.8	4.18	93.1	4.53	4.407
Whole-image feedback	48.5	88.7	4.29	91.4	4.41	4.487
w/o execution trace	44.0	88.9	4.25	89.8	4.27	4.476

Protocol and implementation. All evidence-conditioned methods receive the same request and evidence bundle, while Direct T2I receives only the request. Where applicable, methods use the same Qwen-Image renderer Wu et al. (2025), canvas resolution, and visual-reference budget; native generation and code-rendering pipelines retain their original renderers. Our primary configuration uses Gemini 3.1 Pro Team (2026) for evidence organization, design-prior selection, IVD compilation, and patch proposal. Claude Opus 4.6 cla (2026) is evaluated as an alternative proprietary planner backbone in the supplementary material. Rendering-tree checks are deterministic; model-assisted online verification uses a separate scoped verifier; and the offline evaluator receives final outputs and held-out annotations but no IVD fields, execution traces, repair logs, or checker witnesses.

All evaluation annotations remain hidden from generation and repair. Main results use one output per request and at most $K = 3$ repair rounds. Checker and repair analysis uses 600 stratified obligation decisions from 120 fully audited infographics, while certificate calibration uses all 200 test outputs. We report paired-bootstrap 95% confidence intervals from 10,000 resamples and Holm-adjusted p-values for pre-specified primary tests. Invalid outputs receive zero for ReqCov, Full, ReqText-F1, and end-to-end Bind-Q; NonSup. and Q-Align use valid outputs, with common-valid Q-Align, model snapshots, prompts, hardware, robustness, retrieval, and cost analyses reported in the supplementary material.

4.2 Reliability on IGenBench

Table 1 separates benchmark-reported references, controlled baselines, and InfoAgent variants. InfoAgent achieves 93.0 Q-ACC and 59.0 I-ACC, exceeding the strongest benchmark-reported reference by 10.0 I-ACC points. Among controlled runs, search- and code-oriented baselines improve over Direct T2I, while Same-IVD Prompt remains well below InfoAgent w/o repair, showing that prompt serialization does not reproduce executable layered rendering. Scoped repair further raises I-ACC from 45.0 to 59.0, a paired gain of 14.0 points (95% CI [10.4, 17.6], Holm-adjusted $p < 0.001$).

Figure 2(a) shows gains across the ten IGenBench categories, with the largest improvements in completeness, encoding, marks, axes, annotations, and legend mapping. Figure 2(b) shows that most detected critical violations are resolved within the first two rounds, while the third round yields smaller gains with limited additional edited area.

Table 4: Repair effectiveness and locality on 120 audited repair-eligible outputs.

Strategy	Fix@ Det	Repair Rec.	New Img.	Collat. Reg.	Area Edit
Localized repair	**85.9**	**77.8**	**3.3**	**0.8**	12.4
w/o dependency closure	79.1	71.7	7.5	4.9	**7.8**
Global regeneration	74.8	67.8	10.0	6.8	67.3

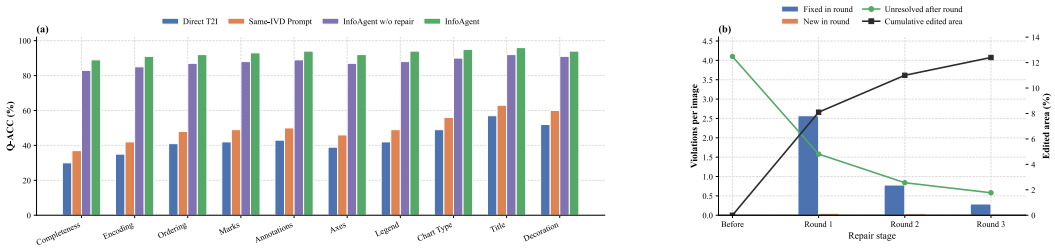

Figure 2: Reliability breakdown and scoped-repair dynamics. (a) Category-wise Q-ACC on IGen-Bench. (b) Mean detected critical violations and cumulative edited area over three repair rounds on 120 audited repair-eligible outputs.

4.3 Evidence-Controlled Evaluation

Table 2 shows a clear progression from evidence conditioning to structured planning, executable rendering, and repair. RAG Prompt improves over evidence-free Direct T2I, while Same-IVD Prompt provides further gains. InfoAgent w/o repair raises Full from 21.5% to 23.5% and ReqText-F1 from 86.4 to 92.4 over Same-IVD Prompt, showing that layered execution contributes beyond prompt serialization; scoped repair further increases them to 28.5% and 96.5. Code-oriented baselines achieve strong text fidelity but lower perceptual quality. Full InfoAgent retains 100% output coverage and obtains the best point estimates on all reported endpoints. Its Full gain over GenClaw is 3.5 points (95% CI [0.6, 6.4], Holm-adjusted $p = 0.041$), while its ReqText-F1 gain over LLM-to-SVG/HTML is 1.8 points (95% CI [0.8, 2.8], adjusted $p = 0.006$).

4.4 Mechanism and Repair Analysis

We isolate addressability through targeted interventions. **Flat plan** retains compiled payloads and coarse layout but removes typed identifiers, dependency edges, executable obligations, and traces. **Whole-image feedback** replaces element-indexed violations with holistic critique and global refinement, while **w/o execution trace** requires post-hoc relocation of rendered elements. Full definitions are provided in the supplementary material.

Table 3 shows that structured planning is insufficient. Flat plan reduces CritPass from 61.5% to 42.5%. Routing symbolic content through raster generation causes the largest ReqText-F1 drop, from 96.5 to 85.2; removing the binding route reduces Bind-Q, from 4.64 to 4.19; and removing design-prior retrieval most strongly affects Layout-Q and Q-Align. Without repair, CritPass is 46.0%. Dependency-aware repair raises it to 61.5%, compared with 54.5% without dependency closure and 50.5% under global regeneration. As shown in Table 4, the verifier detects 90.6% of confirmed critical violations, and localized repair fixes 85.9% of detected violations, yielding 77.8% end-to-end recall. It edits 12.4% of the canvas on average, versus 67.3% for global regeneration, while producing fewer new and collateral errors.

Verifier and certificate audit. Checker performance is evaluated on 600 obligation decisions from 120 audited infographics, with thresholds fixed on the separate calibration split. Rendering-tree, evidence-support, binding, and visual-semantic checks obtain 96.3, 88.1, 84.9, and 82.7 F1, respectively; human agreement reaches Fleiss' $\kappa = 0.82$. On all 200 test outputs, internal certification achieves 92.2% precision and 86.2% recall against external CritPass labels. Detailed counts, UNKNOWN rates, and confusion matrices are provided in the supplementary material.

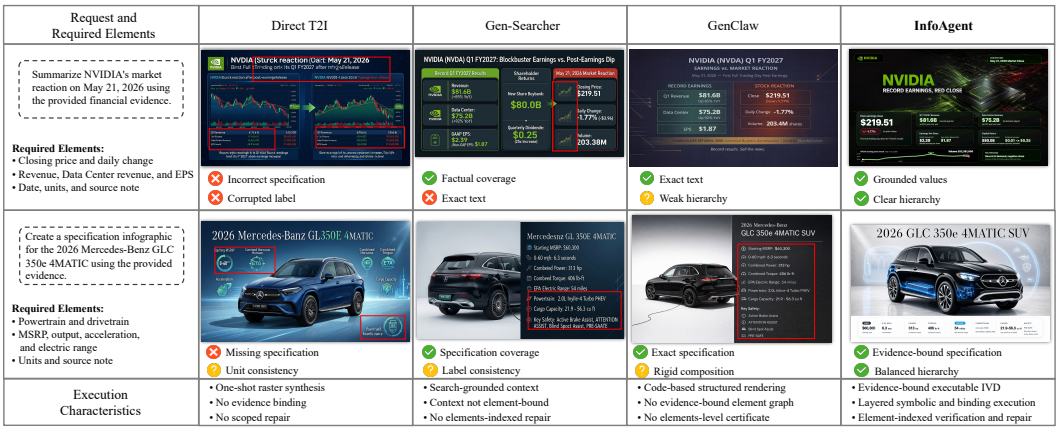

Figure 3: Qualitative comparison on financial-event and product-specification requests. Boxes mark factual, symbolic, binding, and hierarchy differences; ✓, ×, and ? denote satisfied, violated, and partially satisfied requirements. All outputs are shown without manual correction.

4.5 Qualitative Comparison

Figure 3 compares a time-sensitive financial request with a specification-intensive product request. Direct T2I produces visually plausible compositions, but may omit required specifications, corrupt labels, or render unsupported values. Gen-Searcher improves factual and specification coverage through external context, while exact text and label consistency remain less stable because the retrieved evidence is used primarily at the prompt level. GenClaw provides more reliable symbolic content through code-based structured rendering, but its outputs exhibit weaker visual hierarchy and more rigid compositions. InfoAgent instead compiles evidence into addressable IVD elements and executes factual values, specifications, and local relations through coordinated visual, symbolic, and binding layers. The annotated boxes and status markers highlight differences in factual accuracy, symbolic fidelity, label consistency, and visual hierarchy. Additional randomly sampled outputs, failure cases, and repair trajectories are provided in the supplementary material.

Limitations. InfoAgent improves strict reliability but does not make every output correct. On IGenBench, 41.0% of outputs still fail at least one atomic question. On InfoGraphicBench-Evidence, 28.5% pass the complete external checklist, while 61.5% pass all critical factual, symbolic, and binding requirements. Common remaining failures include incomplete or conflicting evidence, uncertain grounding of raster objects, and global layout errors that often require broader replanning. Model-assisted evidence and binding checks also remain less reliable than rendering-tree checks; uncertain obligations are therefore retained as UNKNOWN rather than treated as certified.

5 Conclusion

We presented **InfoAgent**, a training-free framework that links evidence, rendered elements, and verification obligations through IVD. Layered execution retains traces for locating errors, while dependency-aware repair localizes corrections and rechecks affected dependencies and protected obligations. With evidence and the initial IVD held fixed, the complete-checklist pass rate on InfoGraphicBench-Evidence increases from 21.5% for Same-IVD Prompt to 28.5% for InfoAgent. Together with the IGenBench results and repair-locality analysis, these findings support preserving evidence links and execution dependencies throughout generation and revision.

Incomplete or conflicting evidence, uncertain raster-object grounding, and imperfect semantic checks remain sources of failure. The certificate records outcomes under the declared obligations and checkers, retaining unresolved requirements rather than treating them as satisfied. Improving evidence reconciliation and visual grounding, and extending IVD to interactive and multi-page artifacts, are directions for future work.

6 AI USE STATEMENT

We used Gemini (Google) only for grammar correction and language polishing during the writing of this paper. No AI tools were involved in the research ideation, methodology, experiments, or analysis. The authors take full responsibility for all content in this manuscript.

REFERENCES

Claude Opus 4.6. https://www.anthropic.com/news/claude-opus-4-6, 2026. AI assistant.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023.

Feifei Bian, Zhimin Zheng, Wei Deng, Daiguo Zhou, and Jian Luan. Rs-gen: A multi-stage agentic framework for reasoning and search-augmented image generation. *arXiv preprint arXiv:2606.23221*, 2026.

ByteDance. Seedream 4.5. https://seed.bytedance.com/en/seedream4_5, 2025.

Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Steven Hoi, Zhaohui Hou, Shijie Huang, Dengyang Jiang, Xin Jin, Liangchen Li, et al. Z-image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025a.

Qi Cai, Jingwen Chen, Yang Chen, Yehao Li, Fuchen Long, Yingwei Pan, Zhaofan Qiu, Yiheng Zhang, Fengbin Gao, Peihan Xu, et al. Hidream-i1: A high-efficient image generative foundation model with sparse diffusion transformer. *arXiv preprint arXiv:2505.22705*, 2025b.

Shuang Chen, Quanxin Shou, Hangting Chen, Yucheng Zhou, Kaituo Feng, Wenbo Hu, Yi-Fan Zhang, Yunlong Lin, Wenxuan Huang, Mingyang Song, et al. Unify-agent: A unified multimodal agent for world-grounded image synthesis. *arXiv preprint arXiv:2603.29620*, 2026a.

Sixiang Chen, Zhaohu Xing, Tian Ye, Xinyu Geng, Yunlong Lin, Jianyu Lai, Xuanhua He, Fuxiang Zhai, Jialin Gao, and Lei Zhu. Genevolve: Self-evolving image generation agents via tool-orchestrated visual experience distillation. *arXiv preprint arXiv:2605.21605*, 2026b.

Yong Xien Chng, Tao Hu, Wenwen Tong, Xueheng Li, Jiandong Chen, Haojia Yu, Jiefan Lu, Hewei Guo, Hanming Deng, Chengjun Xie, et al. Sensenova-mars: Empowering multimodal agentic reasoning and search via reinforcement learning. *arXiv preprint arXiv:2512.24330*, 2025.

Weiwei Cui, Xiaoyu Zhang, Yun Wang, He Huang, Bei Chen, Lei Fang, Haidong Zhang, Jian-Guan Lou, and Dongmei Zhang. Text-to-viz: Automatic generation of infographics from proportion-related natural language statements. *IEEE transactions on visualization and computer graphics*, 26(1):906–916, 2019.

Victor Dibia. Lida: A tool for automatic generation of grammar-agnostic visualizations and infographics using large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pp. 113–126, 2023.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Jielin Feng, Xinwu Ye, Qianhui Li, Verena Prantl, Jun-Hsiang Yao, Yuheng Zhao, Yun Wang, and Siming Chen. Infoalign: A human–ai co-creation system for storytelling with infographics. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, pp. 1–20, 2026a.

Kaituo Feng, Manyuan Zhang, Shuang Chen, Yunlong Lin, Kaixuan Fan, Yilei Jiang, Hongyu Li, Dian Zheng, Chenyang Wang, and Xiangyu Yue. Gen-searcher: Reinforcing agentic search for image generation. *arXiv preprint arXiv:2603.28767*, 2026b.

Akash Ghosh, Aparna Garimella, Pritika Ramu, Sambaran Bandyopadhyay, and Sriparna Saha. Infogen: Generating complex statistical infographics from documents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 20552–20570, 2025.

Google. Nanobanana-pro. https://blog.google/technology/ai/nano-banana-pro/, 2025.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*, 2025.

Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024.

Chunhan Li, Qifeng Wu, Jia-Hui Pan, Ka-Hei Hui, Jingyu Hu, Yuming Jiang, Bin Sheng, Xihui Liu, Wenjuan Gong, and Zhengzhe Liu. codrawagents: A multi-agent dialogue framework for compositional image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9802–9812, 2026a.

Zhen Li, Duan Li, Yukai Guo, Xinyuan Guo, Bowen Li, Lanxi Xiao, Shenyu Qiao, Jiashu Chen, Zijian Wu, Hui Zhang, Xinhuan Shu, and Shixia Liu. Chartgalaxy: A dataset for infographic chart understanding and generation. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=P4lFbvZ4HH.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

James B McQueen. Some methods of classification and analysis of multivariate observations. In *Proc. of 5th Berkeley Symposium on Math. Stat. and Prob.*, pp. 281–297, 1967.

OpenAI Team. New chatgpt images is here. https://openai.com/index/new-chatgpt-images-is-here/, 2025.

Danqing Shi, Xinyue Xu, Fuling Sun, Yang Shi, and Nan Cao. Calliope: Automatic visual data story generation from a spreadsheet. *IEEE Transactions on Visualization and Computer Graphics*, 27 (2):453–463, 2020.

Wenxin Tang, Jingyu Xiao, Wenxuan Jiang, Xi Xiao, Yuhang Wang, Xuxin Tang, Qing Li, Yuehe Ma, Junliang Liu, Shisong Tang, et al. Slidecoder: Layout-aware rag-enhanced hierarchical slide generation from design. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 9026–9050, 2025.

Yinghao Tang, Xueding Liu, Boyuan Zhang, Tingfeng Lan, Yupeng Xie, Jiale Lao, Yiyao Wang, Haoxuan Li, Tingting Gao, Bo Pan, Luoxuan Weng, Xiuqi Huang, Minfeng Zhu, Yingchaojie Feng, Yuyu Luo, and Wei Chen. Igenbench: Benchmarking the reliability of text-to-infographic generation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (ACL 2026)*, 2026. URL https://arxiv.org/abs/2601.04498.

Gemini Team. Gemini 3.1 pro: A smarter model for your most complex tasks. *Google Blog*, 2026.

Meituan LongCat Team, Hanghang Ma, Haoxian Tan, Jiale Huang, Junqiang Wu, Jun-Yan He, Lishuai Gao, Songlin Xiao, Xiaoming Wei, Xiaoqi Ma, et al. Longcat-image technical report. *arXiv preprint arXiv:2512.07584*, 2025.

Anjul Tyagi, Jian Zhao, Pushkar Patel, Swasti Khurana, and Klaus Mueller. Infographics wizard: Flexible infographics authoring and design exploration. In *Computer Graphics Forum*, volume 41, pp. 121–132. Wiley Online Library, 2022.

Minh Duc Vu, Jieshan Chen, Zhenchang Xing, Qinghua Lu, Xiwei Xu, and Qian Fu. Factflow: Automatic fact sheet generation and customization from tabular dataset via ai chain design & implementation. *arXiv preprint arXiv:2502.17909*, 2025.

Yun Wang, Zhida Sun, Haidong Zhang, Weiwei Cui, Ke Xu, Xiaojuan Ma, and Dongmei Zhang. Datashot: Automatic generation of fact sheets from tabular data. *IEEE transactions on visualization and computer graphics*, 26(1):895–905, 2019.

Joe H Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244, 1963.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023.

Xiaojie Xu, Xinli Xu, Sirui Chen, Haoyu Chen, Fan Zhang, and Ying-Cong Chen. Pregenie: An agentic framework for high-quality visual presentation generation. *arXiv preprint arXiv:2505.21660*, 2025.

Bangji Yang, Ruihan Guo, Jiajun Fan, Chaoran Cheng, and Ge Liu. M3: High-fidelity text-to-image generation via multi-modal, multi-agent and multi-round visual reasoning. *arXiv preprint arXiv:2602.06166*, 2026a.

Zhaorui Yang, Bo Pan, Han Wang, Yiyao Wang, Xingyu Liu, Luoxuan Weng, Yingchaojie Feng, Haozhe Feng, Minfeng Zhu, Bo Zhang, et al. Multimodal deepresearcher: Generating text-chart interleaved reports from scratch with agentic framework. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 34368–34377, 2026b.

Junyan Ye, Jun He, Zilong Huang, Dongzhi Jiang, Xuan Yang, Rui Chen, and Weijia Li. Genclaw: Code-driven agentic image generation. *arXiv preprint arXiv:2605.30248*, 2026.

Wenzheng Zeng, Mingyu Ouyang, Langyuan Cui, and Hwee Tou Ng. Slidetailor: Personalized presentation slide generation for scientific papers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 34584–34592, 2026.

Zekai Zhang, Jiahao Li, Jie Zhang, Kaiyuan Gao, Kun Yan, Lihan Jiang, Ningyuan Tang, Shengming Yin, Tianhe Wu, Xiaoyue Chen, et al. Qwen-image-agent: Bridging the context gap in real-world image generation. *arXiv preprint arXiv:2606.26907*, 2026.

Hao Zheng, Xinyan Guan, Hao Kong, Wenkai Zhang, Jia Zheng, Weixiang Zhou, Hongyu Lin, Yaojie Lu, Xianpei Han, and Le Sun. Pptagent: Generating and evaluating presentations beyond text-to-slides. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 14413–14429, 2025a.

Hao Zheng, Guozhao Mo, Xinru Yan, Qianhao Yuan, Wenkai Zhang, Xuanang Chen, Yaojie Lu, Hongyu Lin, Xianpei Han, and Le Sun. Deeppresenter: Environment-grounded reflection for agentic presentation generation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 31545–31558, 2026.

Yuxiang Zheng, Dayuan Fu, Xiangkun Hu, Xiaojie Cai, Lyumanshan Ye, Pengrui Lu, and Pengfei Liu. Deepresearcher: Scaling deep research via reinforcement learning in real-world environments. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 414–431, 2025b.