From News Text to Causal Infographics via Semantic Triplets: Making an Opaque Image Pipeline Inspectable

Georgi Trajkov, Marko Grobelnik, Dunja Mladenić
Jožef Stefan Institute
Department for Artificial Intelligence
Jamova cesta 39, 1000 Ljubljana, Slovenia
geotrajkov0@gmail.com, {marko.grobelnik, dunja.mladenic}@ijs.si

Abstract. *Generative text-to-image models can turn news articles into a causal-network infographic, but the graph lives only inside the model – it cannot be inspected or corrected. We insert a semantic-triplet stage between text and image and compare direct and triplet-mediated renders on ten Slovene news events, under a rubric that separates triplet correctness from image faithfulness. We hypothesized that decomposition would improve both faithfulness and controllability; in our preliminary single-rater evaluation only controllability holds – the direct image is at least as faithful in every event. The triplet stage's value is auditability: extraction is roughly 70 percent clean, while the renderer mis-draws roughly 19 percent of an otherwise-correct graph's edges – an error the black box hides.*

Keywords. generative AI, text-to-image generation, semantic triplets, knowledge graph construction, causal graphs, compositional fidelity, infographic generation, RDF, large language models, explainability

1 Introduction

A news event of any complexity is covered by many articles across many publishers, and its causal structure – who did what, what followed from what – is spread over dozens of paragraphs. An infographic condenses that structure into a single image a reader can grasp at a glance, which is why editorial node – link graphics are a staple of data-driven outlets; generative image models now make them producible automatically, at the scale of whole event streams. Aggregator platforms such as Event Registry (Event Registry, 2025; Leban et al., 2014) and Ground News (Ground News, 2025) cluster news around events and expose bias and sentiment indicators. Building on Event Registry, our production system BetweenTheLines (MedVrsticami, https://medvrsticami.ijs.si) (Trajkov et al., 2025) analyzes how individual claims and entities are portrayed across publishers; its key lesson is that decomposing a hard large language model (LLM) task into staged prompts – extract, then classify – sharply improves accuracy. We carry that lesson to a different output: a *causal-network infographic*, a single editorial-quality image that lays out the actors and events of a story as a node – link diagram. The infographic module we built into BetweenTheLines generates such an image for each analyzed event, directly from the article text by a text-to-image model in a single step – and that direct text→image generation, our own production design, is the baseline we audit here.

In this one-step setup the article text goes in and a picture comes out, but the graph the picture depicts – which entities appear as nodes, which causal relations appear as edges, and in which direction – is never materialized; it exists only implicitly, inside the image model. This makes the artifact impossible to inspect, query, correct, or evaluate except by eye, and any error (a reversed arrow, an invented link, a dropped relation) is indistinguishable from a deliberate editorial choice.

This paper asks what is gained, and what is lost, when an explicit and inspectable knowledge-graph stage is inserted between the text and the image. We add a semantic-triplet extraction step – subject – predicate – object relations with stable entity identifiers, mirroring the Resource Description Framework (RDF) (W3C, 2014) – and feed the resulting graph to the renderer instead of the raw text. In other words, we test whether decomposing the image task the way BetweenTheLines (Trajkov et al., 2025) decomposed claim classification – extract first, then render – yields a similar gain.

We entered the study with a directional hypothesis: a triplet-mediated pipeline would improve both the *faithfulness* of the rendered infographic and the *controllability* of the pipeline, because the relations can be checked and edited before they are drawn. Our results support only the second half. On image quality the explicit graph does not help – the direct baseline is at least as faithful in every event we examined – but the decomposition delivers exactly the controllability and inspectability the hypothesis anticipated, and that is where its value lies.

Our contributions are: (1) a triplet schema and extraction prompt that produce connected, source-grounded causal graphs with confidence and source-

language constraints; (2) a two-condition comparison (direct vs. triplet-mediated image, same image model, same style prompt) under a two-part rubric that separates triplet correctness from rendered-image faithfulness on the same events; and (3) the finding that, in this preliminary single-rater evaluation, the decomposition's value is transparency and fault-localization rather than image quality – it separates the two error sources the black box conflates, localizing the largest avoidable failure to the generative rendering step.

2 Related Work

Cross-source news analysis. Ground News (Ground News, 2025) classifies publishers on a left – right spectrum and aggregates coverage, but does not track individual claims or produce structured graphs. Event Registry (Event Registry, 2025; Leban et al., 2014) clusters articles into events and provides article-level sentiment (Hutto & Gilbert, 2014). BetweenTheLines (Trajkov et al., 2025) builds on Event Registry with multi-prompt LLM analysis of entity and claim portrayal, and showed that decomposing the task into staged prompts – extraction then classification – markedly improved accuracy; the present study tests whether the same decomposition helps for infographic generation.

Relations and graphs from text. Extracting structured relations from natural language is the classic domain of Open Information Extraction (Banko et al., 2007; Fader et al., 2011), which produces subject – relation – object tuples without a fixed schema, and is conventionally evaluated by the precision and recall of the extracted tuples against a reference. More recently, large language models have largely taken over knowledge-graph and triplet construction (Bian, 2025), and have been applied specifically to extracting causal graphs from text in a zero-shot fashion (Antonucci et al., 2024), where LLM-derived priors are found especially useful for orienting edges (Darvariu et al., 2024) – a step our results show the image renderer often gets wrong. Our triplets are an LLM-produced, schema-light variant, but our aim differs: we extract the causal graph in order to render it as an infographic – and to measure how faithfully a downstream image model redraws it – rather than to populate a database. The entities map naturally to RDF subjects/objects and the predicates to RDF properties (W3C, 2014), which is what makes the intermediate queryable and standardizable in principle.

Text-to-image fidelity for diagrams. Text-to-image models render legible editorial graphics but give no guarantee of structural fidelity to a specified graph. General compositional benchmarks show that these models place plausible objects yet fail to bind the *relations* between them: VISOR (Gokhale et al., 2022) measures how often a stated spatial relation is actually realized, T2I-CompBench (Huang et al., 2023) covers attribute and relation binding, and the weakness persists on current frontier models (Huang et al., 2025; Wang et al., 2026). Diagram- and infographic-specific work confirms it directly: IGenBench (Tang et al., 2026) decomposes infographic reliability into atomic yes/no checks and finds even frontier models rarely satisfy every element of a target infographic, and DiagrammerGPT (Zala et al., 2024) shows these models fail on densely connected arrows and labels – which is why it plans an explicit layout and renders structure-critical content deterministically rather than trusting the generative model. Structural fidelity to a supplied graph is precisely what we measure here.

Graph-drawing readability. How readable a node – link diagram is has long been formalized by the graph-drawing community: Purchase (2002) defines seven measurable aesthetic criteria (edge crossings – the dominant factor – bends, symmetry, angular resolution, and so on). Our single relation-clarity rating corresponds to what this literature measures; a richer evaluation could score the criteria separately and, on a deterministic render, compute them objectively.

3 The Current Infographic Pipeline

For an event, our production step in BetweenTheLines concatenates the title and body of every article in the event cluster (capped at the ten most relevant articles) into one text block and sends it to a text-to-image model with a single prompt that asks the model to both *extract* a causal knowledge graph and *render* it as a Bloomberg/Economist-style node – link infographic (Listing 1). The two responsibilities – deciding the graph and drawing it – are fused in one opaque call. The output is a raster image; no node list, edge list, or evidence is emitted. There is therefore no way to ask which relations the picture claims, whether they are supported by the articles, or whether the drawing even matches the model's own intended graph.

Listing 1. Excerpt of the production prompt: a single call both extracts and renders the graph from the concatenated articles.

```
### Content Processing Rules
Extract a Knowledge Graph: Nodes = Entities +
Claims/Events; Links = Causal Relations (Cause
    -> Effect).
### Input Text
{articles}   # title + body of every article in
    the event
```

4 Proposed Method

We insert an explicit triplet stage and render from it (Fig. 1).

Triplet schema. Extraction returns a JSON object with an `entities` list (stable `id`, human-readable `label`, one `central` flag) and a `triplets` list (`subject_id`, `predicate`, `object_id`, `evidence`, `confidence`), where `evidence` is a verbatim support-

ing text span from the articles. Because subjects and objects reference entity `ids`, reuse across triplets is structural, not lexical – the graph is connected by construction. The `entities` map to RDF subjects/objects and the `predicate` to a (currently free-text) property (W3C, 2014).

Extraction prompt. It targets 10 – 15 triplets over 6 – 10 entities – a deliberate cap to keep the rendered graph legible – and enforces: one connected component, no singletons; specific predicates (no generic "affects/relates to"); a *source-language lock* (labels, predicates and evidence stay in the article language); *event reification* (a death or ruling becomes its own node, not collapsed onto a person); and *attribution-in-predicate* for contested claims (e.g. *naj bi*, "allegedly"), so a one-sided allegation is not stated as fact. A validator checks connectivity and id consistency, retrying once. The few-shot examples use a fictional scenario, so none of the ten evaluation events appears in the prompt (no train-on-test leakage); the provenance of every prompt revision is retained.

Confidence filtering. Each triplet carries a low/medium/high confidence tag, and low-confidence triplets are dropped before rendering. In our ten events this filter is barely exercised: 95 of 115 triplets were tagged high, 19 medium, and only 1 low, so a single triplet was dropped (leaving the 114 edges sent to the renderer) – too few low-confidence cases to judge how reliably the flag catches errors. We keep the mechanism because exposing a per-triplet confidence the pipeline can act on is itself a controllability lever the black box lacks.

Two rendering paths (evaluated). (1) *Direct*: the unchanged baseline, text → image. (2) *Triplet → image*: the same image model and the same visual-style prompt, but fed the explicit entity and edge lists instead of raw text and told to draw only those (Listing 2). Both use the same image model, so the only difference is the input.

Listing 2. Excerpt of the triplet→image prompt: the graph is supplied and the model is told to render *only* it – no extraction, no invention (contrast Listing 1).

```
You MUST use ONLY the entities and relations
    supplied below.
Do NOT introduce any node not in the entities
    list, invent
any edge not in the relations list, or translate
     a label.
### Entities (nodes)     {entities}   # id,
    label, central
### Relations (edges)    {relations}  # subject
    -> pred -> object
```

A third, programmatic path. Because the triplets are now explicit, they could also be rendered programmatically (e.g. a deterministic `graphviz` layout (Gansner & North, 2000)). Such a renderer reproduces the supplied graph with perfect structural fidelity by construction – there is no generative binding step to get wrong – so it removes the render-binding error entirely and collapses the pipeline's error onto the extraction stage. Diagram-generation systems already adopt this discipline (Zala et al., 2024). As a proof of the path we did produce a basic deterministic render (a `graphviz`-layouted SVG) of every evaluated event; they are included in the supplementary viewer, and they reproduce their graphs exactly. Producing a result as readable and information-dense as the generative image, however, requires a typed node ontology and an icon/visual-encoding library (flags, currency, institutional symbols) – substantial design work that is out of scope here. We therefore score only the two generative conditions and treat styled programmatic rendering as a direction enabled by, but not delivered in, this study.

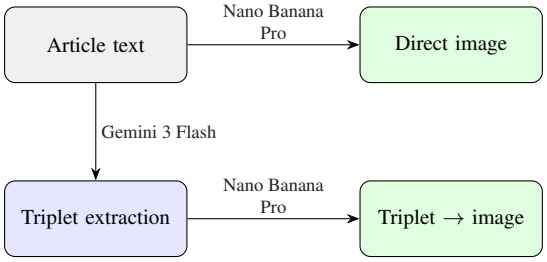


Figure 1. The two evaluated conditions. The direct baseline fuses extraction and rendering in one opaque call; the triplet condition exposes the graph, then renders it with the same image model. Extraction uses Gemini 3 Flash; both image renders use Gemini 3 Pro Image (*Nano Banana Pro*).

5 Evaluation Design

Data. Ten Event Registry events, each with 5 – 10 Slovene-language articles (mean 9.0 articles per event, mean article body ≈1,571 characters); per-event statistics are in Table 1. For triplet extraction, each article body is capped at 3,000 characters – a constraint inherited from BetweenTheLines; every article is always included, only the tails of individual long articles are cut. The cap touches 16 of the 90 articles (11.4% of corpus characters, concentrated in three events). The direct condition is the unchanged production baseline and receives the full text; Section 7 examines this asymmetry and shows the direct image's advantage is not explained by the extra input.

Models. Extraction uses `gemini-3-flash-preview` in JSON mode at temperature 0.2, with article bodies capped at 3,000 characters as above. Both image conditions use the same model, `gemini-3-pro-image-preview` ("Nano Banana Pro"), at 2K resolution and 16:9 with identical settings, so the rendering step itself is held constant. The comparison is deliberately pipeline-level rather than input-controlled: the triplet pipeline caps entities and relations while the direct baseline is free-form, and the extraction stage caps article length while the baseline does not. These asymmetries are part of what is being compared – each

Table 1. Per-event input statistics. Articles are Slovene; chars is the mean article-body length.

Event	Articles	Mean chars
E1	5	1,278
E2	7	920
E3	8	1,151
E4	10	1,088
E5	10	956
E6	10	967
E7	10	1,210
E8	10	1,825
E9	10	3,069
E10	10	3,247
Mean	9.0	1,571

condition runs its production configuration – and their consequences are analyzed in Section 7.

Rubric (two parts). A single rater – one of the authors – scores each event through a purpose-built side-by-side interface that shows the source articles, the extracted triplets, and both renders; scores are exported as per-rater JSON, so independent raters can later be added without changing the tooling. The rubric has two parts because the two pipelines expose two different things to inspect.

The first part is *the final image*, which both pipelines produce and which is judged against the articles: faithfulness (1 – 5), relation clarity (1 – 5), and a node count for density. Faithfulness scores how well the image is grounded in the source articles and how much of the story it manages to convey: 5 means the depicted content is fully supported by the articles and all important details of the story are communicated; 1 means hallucinated or wrong information and missing crucial points. An image that is fully grounded but omits crucial details therefore scores lower. Relation clarity scores how readable and digestible the drawn structure is – whether it is clear what happened, what led to what, and who is connected to whom; it corresponds to what the graph-drawing literature formalizes as readability (Purchase, 2002), compressed into a single rating: 5 means the flow of events and relations is immediately clear and the nodes look as they should; 1 means the image is effectively unreadable – overlapping arrows, confusing wording, an unclear flow of events, or icons and visual elements that confuse rather than clarify. The node count is taken by eye from the rendered image in both conditions. For the direct image this single number is all its hidden graph exposes; for the triplet image the accounting is finer-grained – the per-edge render verdicts of the second part record which supplied elements were drawn, omitted, or added – and the supplied graph size (8 in every event) cross-checks the count.

The second part is *the explicit graph*, which exists only in the triplet pipeline; because it is inspectable it can be scored partly objectively. Each triplet receives a per-triplet verdict on a severity scale (correct / *minor*: misleading, awkward, imprecise / *error*: wrong direction, wrong predicate, wrong endpoints, unsupported, redundant), judged against the articles – this measures extraction quality much as precision does for Open Information Extraction (Banko et al., 2007). Event-level coverage (1 – 5) captures the recall side – how much of the story the graph covers: 5 means all main strands are present, 3 means the core is covered but side strands are missing, 1 means major aspects are absent. Finally, each supplied edge gets a per-edge render verdict (drawn correctly, wrong direction, wrong connection, omitted) recording how the image redrew it, plus a count of unsupported additions; this per-element check mirrors the atomic verification used by infographic-reliability benchmarks (Tang et al., 2026). The direct pipeline has no such intermediate, so this second part has nothing to score there – which is exactly the asymmetry the study is about.

6 Results

Ten events, a single rater; 115 triplets (mean 11.5 per event) and 114 rendered edges scored.[1] The evaluation is preliminary – one rater, one render per condition – so we report means, sign consistency, and per-element counts rather than significance tests.

6.1 The triplet image does not win

The direct image is more faithful (4.10 vs. 3.10, mean difference 1.0) and clearer (3.60 vs. 2.80) (Table 2). The gap is not large, but it is consistent: the triplet image ties the direct image on faithfulness in 5 of the 10 events and never exceeds it – direct $\geq$ triplet in 10/10 events. With only ten paired events and five ties, we do not claim statistical significance; the robust, defensible statement is the sign consistency. The one decisive direct advantage is density: 18.4 vs. 8.1 nodes counted by the same by-eye procedure, a $2.3\times$ gap. Every one of the ten supplied graphs contained exactly 8 entities; in nine events the render drew exactly those 8 nodes, and in one it drew 9, duplicating a supplied node into two boxes. The gap is structural – the direct image freely promotes quantities, dates and outcomes to nodes, whereas the triplet schema admits only typed entities.

6.2 Extraction is mostly sound

Of the 115 rated triplets, 69.6% are correct and 25.2% have only minor issues (chiefly *imprecise* or *misleading* phrasing); 5.2% are hard errors, dominated by *unsupported* relations (3.5%) (Table 3). Event-level coverage averages $3.7/5$: the 10 – 15-triplet cap that keeps

[1] All extracted triplets, generated graphs, prompts, source articles, and rubric scores – with an interactive read-only viewer and a full data download – are available at https://causal-infograph.netlify.app.

Table 2. Final-image quality and density, direct vs. triplet→image (10 events, single rater). Faithfulness/clarity 1 – 5 (higher is better); mean ± SD. Nodes drawn is counted by eye from the rendered image in both conditions; each triplet condition's supplied graph had exactly 8 entities.

Metric	Direct	Triplet→image
Faithfulness (1 – 5)	4.10 ± 0.74	3.10 ± 0.74
Relation clarity (1 – 5)	3.60 ± 0.97	2.80 ± 1.03
Nodes drawn	18.4 ± 8.0	8.1 ± 0.3

Table 3. Per-triplet verdicts by severity tier (115 triplets). This scores each *extracted* triplet against the articles – whether the relation is correct – independent of how the image later drew it (Table 4).

Verdict	Count	Share
Correct	80	69.6%
Minor	29	25.2%
imprecise	15	13.0%
misleading	10	8.7%
awkward wording	4	3.5%
Error	6	5.2%
unsupported	4	3.5%
wrong endpoints	1	0.9%
redundant	1	0.9%

graphs legible also forces real omissions, the recall cost of the same budget that drives the density gap.

6.3 The renderer is the largest avoidable error

Given the explicit graph, the image model draws 80.7% of edges correctly, but 14.9% are *wrong connections* – an edge attached to the wrong node, the single most frequent failure across the whole pipeline – with a further 2.6% direction-flipped, 1.8% omitted, and 2 hallucinated edges plus 1 duplicated node added across the ten events (Table 4).

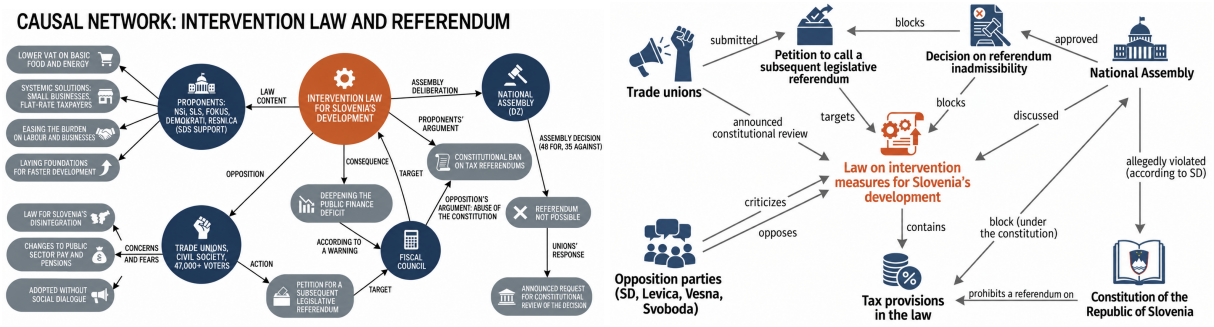
Fig. 2 is the illustrative case. Its triplets were rated "very good" and its triplet→image tied the direct image on *faithfulness* (4/5 each); what was poor was its *render fidelity* – only 8 of 12 supplied edges were drawn to the right nodes. Concretely, the four mis-connected edges were: Tax provisions → block (per constitution) → Referendum petition, whose arrow instead terminated at the National Assembly; two edges that should both end on the Inadmissibility-decision node – Opposition → opposes and Unions → announce constitutional review – which were attached to the Intervention-law node (the graph's central, most prominent node) instead; and a fourth, running *from* the decision to the Constitution, which was re-rooted at the Assembly. The pattern is that the model attaches an arrow to the wrong – typically nearer or more prominent – node, drawing a structurally different graph from the one supplied. This is the relational/spatial-binding weakness that text-to-image models are independently documented to exhibit on general compositional scenes (Gokhale et al., 2022; Huang et al., 2023), that persists on current frontier models (Huang et al., 2025; Wang et al., 2026), and that diagram-specific work confirms for node – link rendering in particular (Tang et al., 2026; Zala et al., 2024). A programmatic render would not suffer this class of error, the practical motivation for treating it as future work.

Table 4. Triplet→image render fidelity: how the model drew each of 114 supplied edges, plus 2 hallucinated additions.

Rendered as	Count	Share
correctly drawn	92	80.7%
wrong connection	17	14.9%
wrong direction	3	2.6%
omitted	2	1.8%

6.4 Drawing style: uniform triplet renders vs. tailored direct renders

Both conditions receive the identical visual-style prompt, so any difference in appearance comes from the input, not the specification. Alongside the uniform quality gap (direct $\geq$ triplet on faithfulness in 10/10 events and on clarity in 9/10), the two paths differ in visual character in a way that mirrors the density gap. The triplet renders are uniform and skeletal: every event hands the model the same kind of input – a fixed list of 6 – 10 entities and 10 – 15 edges – so it simply lays out the supplied skeleton, and the ten renders come out structurally alike. The direct render instead takes liberty and fits a richer graph to each event. Fig. 3 is typical: fed the raw articles, the direct image promotes event-specific detail to nodes – passenger counts, a timestamp, a casualty breakdown, separate emergency-service and media clusters – in a bespoke multi-flow layout, whereas the triplet image of the same event is a plain eight-node chain. This is the same input freedom that produces the density gap of Section 6.1, now visible as composition: fed raw articles the model has both the most to work with and the most latitude, and the result is a looser, less rigid layout, whereas the triplet render is locked to the supplied skeleton. Narrowing the gap is therefore mainly a matter of giving the triplet renderer comparable latitude – a broader node ontology, or licence to add supporting detail.



(a) Direct (text→image) **(b)** Triplet→image

Figure 2. Hero event (the intervention-law referendum). The triplets are sound and (b) tied (a) on faithfulness, but the model mis-connected four of its twelve edges – "good triplets, wrong renders." The direct image (a) is denser and more polished. Labels are translated into English for publication; the depicted connections, including rendering errors, are retained.

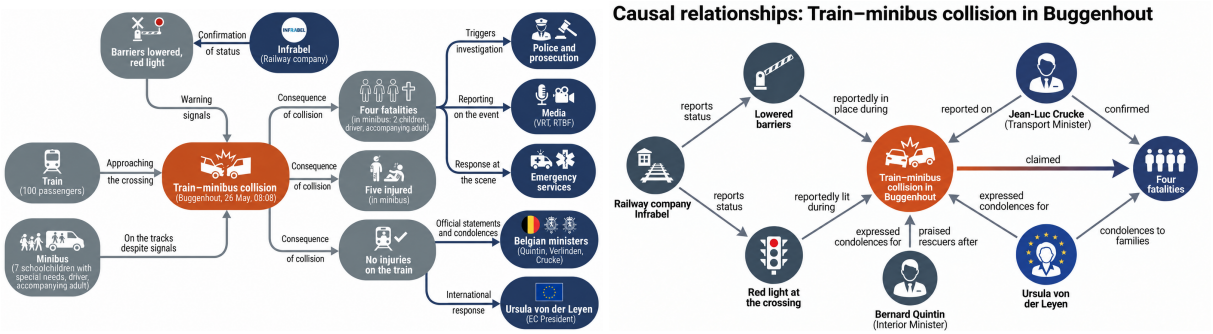
7 Discussion

The result reframes the contribution. Decomposition did not produce a better infographic – on a black-box basis the direct baseline wins – but it turns a single unverifiable image into an *auditable chain*: extraction is $\sim$70% clean ($\sim$5% hard errors), while the rendering stage mis-draws $\sim$19% of edges, predominantly wrong connections. The black box conflates these two error sources; the explicit graph pulls them apart, localizing the largest avoidable failure to generative rendering, not extraction. This is the Semantic-Web payoff in miniature: a standardizable intermediate enables verification that an end-to-end model forecloses.

The density gap follows from the triplet budget – but closing it is not free. The direct prompt asks for "key Entities *and* Key Claims/Events" and caps neither node nor edge count, so the model freely promotes quantities, dates and outcomes to nodes (mean 18.4); the triplet pipeline caps both axes (6 – 10 typed entities, 10 – 15 triplets) and every one of the ten events converged on exactly 8 entities. The same cap that limits its density also limits coverage (the $3.7/5$ above). Is the lower node count also the reason for the triplet image's lower relation clarity? The per-event scores suggest not: within the triplet condition every graph has the same 8 nodes, yet clarity varies from 2 to 5 and tracks render fidelity – the correlation between clarity and the share of correctly drawn edges is $r = 0.75$ – while in the direct condition node count and clarity are essentially unrelated ($r = -0.18$; the densest render, 40 nodes, scores only 3). Low clarity comes from misdrawn edges, not from sparsity. Could the extractor simply be encouraged to produce more nodes? Technically yes – the cap is a prompt parameter – but our render results argue against treating density as a dial one can simply turn up: the model already mis-binds $\sim$19% of edges on these sparse eight-node graphs, and adding nodes means more – and longer – edges, which would most likely make the binding worse, not better. Raising density dependably would therefore require more than a larger target count: extraction rules that yield graphs whose connections render reliably (e.g. locally clustered, layout-friendly structure), or a constrained/programmatic layout that removes the binding error outright. Density and render fidelity trade off, so the rendering failure must be addressed before added density would help rather than hurt.

Combining the two approaches. The pipelines need not be alternatives. The explicit graph can serve the better-looking direct image rather than replace it: (i) *triplets as a QA layer* – render the direct image for density and polish, and use the extracted graph to audit it for missing, unsupported or reversed relations; (ii) *hybrid input* – feed the model both the text (for style and density) and the triplets (as a checklist of relations that must appear, drawn in the stated direction); (iii) *structure-locked styling* – fix the layout with a deterministic graph renderer and let the image model only style and iconify it, removing the binding error while keeping the editorial look; and (iv) *wider triplet latitude* – broaden the node ontology or let the extractor add supporting detail, so the triplet render can tailor a richer graph per event the way the direct one does (Section 6.4). Options (i) and (iii) are the most promising: the first keeps today's better image and adds inspectability for free, the second attacks the dominant error directly.

An asymmetry in the evaluation itself. For the triplet image the rater knows the intended graph, so every mis-binding is findable; for the direct image there is no ground truth, so its structural errors are inherently harder to spot. The direct image's faithfulness (4.10) is therefore plausibly *over*-estimated relative to the triplet image, and the measured gap is in part an artifact of exactly the asymmetric inspectability this paper argues for. Far from undermining the result, this sharpens it: the black box's errors are invisible even to a careful human evaluator, which is the strongest possible case for materializing the graph.



(a) Direct – rich, event-tailored **(b)** Triplet→image – uniform, skeletal

Figure 3. Same style prompt, different composition (a fatal train – minibus collision). Fed the raw articles, the direct render (a) tailors a rich, event-specific graph (passenger counts, timestamp, casualty breakdown, emergency-service and media clusters); fed the fixed triplet list, the triplet render (b) is a uniform, skeletal layout. Bound to the supplied skeleton, the triplet render is also the more rigid of the two. Labels are translated into English for publication; the depicted connections are retained.

An input asymmetry. The extraction stage capped each article body at 3,000 characters while the direct baseline received the full text (Section 5). Only three events were affected at all, and the asymmetry runs the wrong way to explain the headline result: the direct image's faithfulness advantage is larger on the seven untruncated events (mean gap 1.29) than on the three truncated ones (0.33), and direct $\geq$ triplet holds in both subsets. The extra input therefore cannot account for the direct image's win; the asymmetry's plausible cost falls on the triplet pipeline's coverage, which we already report as a limitation.

What the explicit intermediate unlocks. Materializing the graph enables what the black box cannot: *selective rendering* (over-extract, then draw only a chosen subgraph rather than let the model silently omit); *cross-event bridging* via shared entities, surfacing links invisible across separate images; *sentiment linking*, joining nodes to the system's per-publisher sentiment; and *interactive images*, where OCR'd nodes let a reader click through to a node's cross-source coverage.

8 Conclusion, Limitations and Future Work

Making the hidden knowledge graph of a news infographic explicit does not, by itself, make the picture better – but it makes the pipeline honest. In our ten-event comparison the direct render was at least as faithful in every event (mean faithfulness 4.10 vs. 3.10) and substantially denser (18.4 vs. 8.1 nodes), yet the explicit stage turned a single unverifiable image into an auditable chain: extraction was roughly 70% clean with only ∼5% hard errors, while the generative renderer mis-drew ∼19% of the supplied edges, predominantly by attaching them to the wrong node – the pipeline's largest avoidable failure, and one the black box hides. We can see where the pipeline succeeds (extraction) and where it fails (generative rendering), and can substitute a faithful renderer where structure matters. For a Semantic-Web view of generative pipelines, that inspectability is the point.

These conclusions carry the limitations of a preliminary study. All ratings come from a single rater who is also an author, so the subjective scores are a self-assessment; with ten paired events and five ties we report sign consistency (10/10) rather than a significance test; and each condition was sampled once per event although image generation is stochastic, so per-event sampling effects cannot be ruled out. Only a handful of prompt iterations preceded freezing, so the results characterize this pipeline instance rather than an upper bound; the narrow node ontology explains part of the density and coverage gaps; and the dominant render error – edge mis-binding – is a documented weakness of current text-to-image models (Gokhale et al., 2022), whose mitigation through a constrained or programmatic layout is a prerequisite to scaling density.

An extended version will accordingly add independent second and third raters with Cohen's κ agreement on the categorical verdicts (the scoring tool already exports per-rater scores), multiple renders per condition with reported variance, and gold-standard triplets; further directions are RDF-star reification (Hartig et al., 2021) for attribution, controlled-vocabulary predicates with ontology alignment, e.g. Wikidata (Chepurova et al., 2025), and a richer deterministic renderer building on the `graphviz` proof-of-path included in the supplementary material.

Acknowledgments

This research was financially supported by the Slovenian Research and Innovation Agency (ARIS) via: the Research Program Knowledge Technologies (grant P2-

0103); the Gravity project AI for Science (grant GC-0001); and EC via the projects TWON (101095095) and DataPact (grant 101189771).

References

Antonucci, A., Piqué, G., & Zaffalon, M. (2024). Zero-shot causal graph extrapolation from text via LLMs [arXiv:2312.14670]. *XAI4Sci Workshop, AAAI Conference on Artificial Intelligence*. https://arxiv.org/abs/2312.14670

Banko, M., Cafarella, M. J., Soderland, S., Broadhead, M., & Etzioni, O. (2007). Open information extraction from the web. *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI)*, 2670–2676. https://dl.acm.org/doi/10.5555/1625275.1625705

Bian, H. (2025). LLM-empowered knowledge graph construction: A survey. *arXiv preprint arXiv:2510.20345*. https://arxiv.org/abs/2510.20345

Chepurova, A., Bulatov, A., Burtsev, M., & Kuratov, Y. (2025). Wikontic: Constructing wikidata-aligned, ontology-aware knowledge graphs with large language models. *arXiv preprint arXiv:2512.00590*. https://arxiv.org/abs/2512.00590

Darvariu, V.-A., Hailes, S., & Musolesi, M. (2024). Large language models are effective priors for causal graph discovery [arXiv:2405.13551]. https://arxiv.org/abs/2405.13551

Event Registry. (2025). *Event registry api documentation*. Retrieved September 7, 2025, from https://eventregistry.org/documentation

Fader, A., Soderland, S., & Etzioni, O. (2011). Identifying relations for open information extraction. *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1535–1545. https://aclanthology.org/D11-1142

Gansner, E. R., & North, S. C. (2000). An open graph visualization system and its applications to software engineering. *Software: Practice and Experience*, *30*(11), 1203–1233.

Gokhale, T., Palangi, H., Nushi, B., Vineet, V., Horvitz, E., Kamar, E., Baral, C., & Yang, Y. (2022). Benchmarking spatial relationships in text-to-image generation [arXiv:2212.10015]. https://arxiv.org/abs/2212.10015

Ground News. (2025). *Ground news — breaking news headlines and media bias*. Retrieved September 7, 2025, from https://ground.news/

Hartig, O., Champin, P.-A., Kellogg, G., & Seaborne, A. (2021). *RDF-star and SPARQL-star (w3c community group final report)*. Retrieved June 1, 2026, from https://www.w3.org/2021/12/rdf-star.html

Huang, K., Duan, C., Sun, K., Xie, E., Li, Z., & Liu, X. (2025). T2I-CompBench++: An enhanced and comprehensive benchmark for compositional text-to-image generation [arXiv:2307.06350]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. https://doi.org/10.1109/TPAMI.2025.3531907

Huang, K., Sun, K., Xie, E., Li, Z., & Liu, X. (2023). T2I-CompBench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*. https://proceedings.neurips.cc/paper_files/paper/2023/hash/f8ad010cdd9143dbb0e9308c093aff24-Abstract-Datasets_and_Benchmarks.html

Hutto, C., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*, 216–225. https://doi.org/10.1609/icwsm.v8i1.14550

Leban, G., Fortuna, B., Brank, J., & Grobelnik, M. (2014). Event registry: Learning about world events from news. *Proceedings of the 23rd International Conference on World Wide Web*, 107–110. https://doi.org/10.1145/2567948.2577024

Purchase, H. C. (2002). Metrics for graph drawing aesthetics. *Journal of Visual Languages & Computing*, *13*(5), 501–516. https://doi.org/10.1006/jvlc.2002.0232

Tang, Y., Liu, X., Zhang, B., Lan, T., Xie, Y., Lao, J., Wang, Y., Li, H., Gao, T., Pan, B., Weng, L., Huang, X., Zhu, M., Feng, Y., Luo, Y., & Chen, W. (2026). IGenBench: Benchmarking the reliability of text-to-infographic generation. *arXiv preprint arXiv:2601.04498*. https://arxiv.org/abs/2601.04498

Trajkov, G., Grobelnik, M., & Mladenić Grobelnik, A. (2025). BetweenTheLines - cross source news analysis. *Proceedings of the Slovenian KDD Conference 2025*. https://doi.org/10.70314/is.2025.sikdd.26

W3C. (2014). *RDF 1.1 concepts and abstract syntax*. World Wide Web Consortium. Retrieved June 1, 2026, from https://www.w3.org/TR/rdf11-concepts/

Wang, Z., Hu, X., Wang, Y., Xiong, F., Zhang, M., & Chu, X. (2026). Everything in its place: Benchmarking spatial intelligence of text-to-image models. *arXiv preprint arXiv:2601.20354*. https://arxiv.org/abs/2601.20354

Zala, A., Lin, H., Cho, J., & Bansal, M. (2024). DiagrammerGPT: Generating open-domain, open-platform diagrams via LLM planning [arXiv:2310.12128]. *Conference on Language Modeling (COLM)*. https://arxiv.org/abs/2310.12128