

From disruption narratives to visual scenario cards: A design-prior-guided AIGC framework for manufacturing resilience communication

Yuxuan Mao¹, Liyao Xiao^{2*}

¹Hubei College of The Arts, Wuhan, China

²School of Art and Tourism, Guizhou Communications Polytechnic University, Guiyang, China

Abstract

INTRODUCTION: In discrete and process manufacturing and their upstream supply networks, disruptions (natural hazards, cyber incidents, logistics blockages, geopolitical actions, facility failures, and demand shocks) reach decisionmakers as fragmented text narratives, slowing shared understanding across roles, while direct text-to-image generation leaves causal structure visually undifferentiated and design knowledge tacit. The research object is information-design knowledge (an art and design discipline) applied to manufacturing-disruption communication; it is not art manufacturing.

OBJECTIVES: To represent art and information-design knowledge as an explicit, inspectable v0.1 rule table and quantify the effect of a deterministic causal-organisation prompt manipulation under matched generation conditions.

METHODS: The dataset comprises 1,694 RSS-harvested manufacturing-disruption narratives, of which 758 passed deterministic lexical screening. Records were embedded with sentence-transformers/all-MiniLM-L6-v2, and a six-category query-derived taxonomy was audited with HDBSCAN and sensitivity analysis. Eighteen pre-generation design-prior records were stored in SQLite. A deterministic constructor emitted matched direct (D) and causal organisation (G) prompts whose verbatim text and hashes were recorded in a CSV manifest. Stable Diffusion XL produced 96 images for 24 scenarios under two conditions and two fixed seeds; outputs were assessed with CLIP image-text cosine probes, visual-complexity descriptors, and 192 label-blinded paired judgments from two independent raters using Wilson intervals, sign tests, and chance-corrected reliability statistics.

RESULTS: Clustering showed moderate taxonomy correspondence (NMI 0.578–0.613, ARI 0.456–0.592, noise 67.3–72.4%). The grammar condition won 97.4–100% of non-tied judgments across four dimensions (all sign tests $p < .001$; AC1 = 0.663) without improving CLIP alignment, while edge density and luminance entropy increased significantly after Holm correction; these are changes in structural complexity, not quality verdicts.

CONCLUSION: Design knowledge can operate as an auditable computational control layer whose benefit registers in expert-judged communication value rather than automatic similarity scores, offering a reusable template for knowledge-conditioned industrial AIGC. The evidence concerns resilience communication artefacts, not resilience performance or decision quality, and is bounded by one generation model, 24 scenarios, and two raters.

Keywords: manufacturing resilience, AIGC, information design, visual grammar, prompt compilation, human–AI collaboration

Received on 17 September 2026, accepted on 12 October 2026, published on 20 October 2026

Copyright © 2026 Yuxuan Mao *et al.*, licensed to EAI. This is an open access article distributed under the terms of the [CC BY-NC-SA 4.0](#), which permits copying, redistributing, remixing, transformation, and building upon the material in any medium so long as the original work is properly cited.

doi: 10.4108/eetsis.15183

*Corresponding author. Email: gzxiaoliyao@gmail.com

1. Introduction

Modern manufacturing systems fail in networked ways: an earthquake near a supplier cluster, a ransomware intrusion, or a port closure propagates across tiers before its consequences surface on a production floor [1–3]. Resilient response therefore depends not only on prediction and rescheduling but on how quickly heterogeneous teams—engineering, operations, management, and design—converge on a shared understanding of what is happening and what can be done [4–6]. In practice, the raw material for this shared understanding is fragmented: news headlines, incident bulletins, and technical narratives that different roles parse differently and slowly.

This study is situated in discrete and process manufacturing and their upstream supply networks (electronics, automotive, chemicals, and general industrial production as reported in public news feeds) rather than in any single sector, because the six disruption types examined cut across sectors. Its disciplinary home is information design; the branch of art and design concerned with the visual organisation of information for communication. The research object is therefore neither art nor art manufacturing: it is the design knowledge that governs how a manufacturing disruption should be depicted so that a mixed-role team can read it. Manufacturing is the application domain; art and design theory is the source of the prior knowledge.

Text-to-image generation (AIGC) offers a fast route from narrative to picture [7, 8]. Yet direct prompting routinely yields images that look plausible while leaving the causal structure of a disruption—what triggered it, which node is exposed, how effects propagate, and which responses exist—visually undifferentiated: prompt wording strongly shapes what is depicted [9], and diffusion models are known to neglect or mis-bind prompt elements and to struggle with spatial relations and multi-object composition [10, 11]. For resilience communication this is not a cosmetic defect: an attractive but causally flat image can misdirect a recovery meeting more than no image at all.

Semantic-to-image conversion, as used here, means turning a five-slot semantic representation of a disruption (trigger, exposed node, propagation path, manufacturing consequence, response option) into a single-frame image. Its significance is that these five slots are exactly what a resilience meeting must establish before any action is chosen. Its objective is not photorealism but a composition that makes the five slots and their causal order legible at a glance. Its intended usage is as a shared starting artefact in resilience meetings, training sessions, and early what-if discussion, where one common picture shortens the time that engineering, operations, management, and design roles need to reach the same understanding. Generated cards are never used as evidence of what actually happened.

Existing research approaches this gap from two sides that rarely meet. Manufacturing-resilience studies concentrate on prediction, optimisation, and recovery policy, often treating communication artefacts as secondary [3, 12, 13]. AIGC studies concentrate on image quality and text–image similarity [14, 15], while visual hierarchy, figure–ground

organisation, and causal reading order remain largely implicit in prompt practice [9, 16]. What is missing is a transparent mechanism study in which a recorded design-prior table informs a reproducible prompt manipulation and the resulting cards are evaluated by human raters on task-relevant communication criteria.

The problems addressed and objectives pursued can be summarised as follows. Problems: (P1) disruption information arrives as fragmented text; (P2) direct prompting can produce plausible but causally flat images; (P3) design knowledge is often left implicit; and (P4) automatic text–image similarity is often treated as a proxy for communicative usefulness without task-level validation. Objectives: (O1) build a screened corpus and audit an organising taxonomy; (O2) store a bounded set of information–design actions as explicit v0.1 rule records; (O3) operationalise a causal–organisation subset as a deterministic, token-audited prompt manipulation; and (O4) test under matched generation whether this manipulation changes expert-judged communication value and whether automatic proxies detect the change.

This study builds and tests such a framework for manufacturing-disruption scenario cards: single-frame generated visualisations intended for resilience meetings, training, and early what-if discussion. The pipeline couples (i) automatic discovery and lexical screening, (ii) density-based clustering used as an exploratory audit of a six-category organising scheme, (iii) an explicit 18-row design-prior table, (iv) deterministic construction of matched D/G prompts, and (v) a controlled image-generation and paired-rating experiment. The study evaluates communication artefacts; it does not estimate disruption probabilities, reconstruct real events, or automate manufacturing decisions.

Four research questions are addressed. RQ1 asks how strongly unsupervised embedding clusters correspond to the query-derived organising taxonomy. RQ2 asks how a bounded set of art and information–design actions can be represented explicitly and operationalised as a reproducible prompt manipulation. RQ3 asks how the causal–organisation condition affects perceived completeness, causal clarity, visual hierarchy, and meeting usefulness under matched generation. RQ4 asks which category-level effects and failure patterns persist. The contributions are a traceable corpus-to-visualisation workflow; a design–research operationalisation of five causal–organisation directives; and empirical evidence that human-rated communication value can dissociate from generic image–text similarity. Claims concern resilience communication, not resilience performance or decision quality.

2. Related work

2.1 Manufacturing disruption and resilience communication

Supply-chain and manufacturing resilience research characterises disruptions by their triggers, propagation, and

recovery dynamics [1–3], spanning reviews of resilience principles and network viability [12, 13], risk-management and severity-design perspectives [17, 18], and a recent major review of the field [19]. A consistent finding is that disruption effects are systemic: local failures cascade through material, information, and financial flows, so that effective response requires cross-functional situation awareness [4]. Human-centric smart-manufacturing agendas likewise position shared understanding between humans and cyber-physical systems as a first-class requirement [5, 6]. Digital supply-chain twins extend this agenda by giving decisionmakers a continuously updated, model-based view of network state for disruption management [20]; the scenario cards studied here are complementary, lightweight communication artefacts rather than analytic models. However, the artefacts through which that understanding forms—risk matrices, network diagrams, and textual bulletins—either abstract away the concrete scene or overload non-specialist readers. Studies of visual decision support show that representation choices systematically change what viewers extract and decide [21, 22], motivating disruption visuals whose organisation is deliberately designed rather than incidental.

Disruption taxonomies in this literature are typically theory-derived: categories such as natural hazards, equipment failure, cyber incidents, logistics blockage, geopolitical action, and demand shocks recur across reviews [12, 13, 19]. Whether such categories also organise the language in which disruptions are actually reported is rarely tested. Density-based clustering over narrative embeddings offers a direct audit: if theory-derived labels align with latent semantic structure, category-conditioned generation and retrieval inherit a defensible basis; if not, the taxonomy remains a communication convention whose limits should be stated. This study performs exactly that audit before using the taxonomy to anchor scenarios.

2.2 AIGC for industrial visualisation and its evaluation

Diffusion models [23] and successors such as Stable Diffusion [7] and SDXL [8] made high-resolution text-to-image generation broadly accessible [24, 25]. Design practitioners already use these models for concept ideation, illustration, and communication material, with prompt wording as the principal control interface [9, 26]. Structural conditioning methods such as ControlNet [27] use auxiliary images; the present experiment instead changes a short text organisation clause so the manipulation remains inspectable. Prompt-level control has become an object of study through human-authored guidelines and modifier taxonomies [9, 16], learned rewriting [28], language-grounded layouts [29], and attention steering [10]. Evaluation often uses CLIP-space similarity [15, 30], learned preference models [31], or question-answering faithfulness [11], but these remain generic proxies. The present work therefore combines a reproducible text-only manipulation with a task-specific paired-rating protocol.

Recent evaluation research further shows that automatic scores and human judgments answer different questions. CLIP-based alignment asks whether an image resembles a text in a joint embedding space [15, 30]; a meeting participant asks whether the image lets the room see what triggered the disruption and what can be done next. Otani et al. document how loosely human evaluation is often reported in text-to-image work and call for verifiable protocols [14]. The present study responds with a fully paired, position-balanced, label-blinded expert protocol whose reliability is reported as a battery of chance-corrected coefficients rather than a single favourable statistic.

2.3 Art and information-design priors

Visual-communication scholarship provides explicit principles for reading paths [32], graphic variables [33], data-ink economy [34], perceptual guidance [35], and cognitive-load management [36]. These principles motivate actions such as trigger salience, directional propagation, figure-ground contrast, and separation of consequence from response. In this project, 18 such actions were recorded as v0.1 design priors with activation conditions, weights, hard flags, and rationales; a five-directive causal-organisation subset was operationalised in the G prompt. The table makes the intended design actions inspectable even though the archived project does not contain a separate rule-admission study or source-level provenance for each row.

Making design actions explicit changes what can be audited. Reviewers can inspect the 18 stored actions, the exact D/G organisation strings, every positive prompt, the negative prompt, and the run hashes. This is narrower than full rule-to-source provenance: the archived schema records rationales but not a separate provenance table or the derivation history of each rule. The contribution is therefore transparency of the experimental manipulation, not a claim that the current v0.1 knowledge base already provides complete lifecycle governance.

2.4 Gap and positioning

To our knowledge, and without claiming the completeness of a systematic review, we have not found work combining a screened manufacturing-disruption corpus, an exploratory clustering audit, explicit design-prior records, matched prompt construction, and paired human evaluation in one mechanism study. This differs from prompt engineering [9, 16, 28], which optimises wording; controllable generation [10, 27, 29], which often intervenes inside the model; design grammar [32, 33], which is descriptive rather than an evaluation protocol; knowledge-guided AIGC, which usually injects domain facts; and interactive human-AI design [26, 37, 38]. The specific empirical novelty is the matched test of a five-directive causal-organisation clause, while the broader contribution is the traceable end-to-end workflow. Table 1 compares the study with eight representative works without claiming that the unimplemented general admission/search formulation was executed in the present experiment.

This study occupies that gap and deliberately bounds its claims: generated images are labelled generated scenario visualisations, and all conclusions concern communication

value under matched conditions rather than predictive validity or decision quality.

Table 1. Comparison of the present work with representative related studies

Study	Control mechanism	Design knowledge	Task anchoring	Evaluation	Difference from this work
Liu & Chilton 2022 [9]	Manual prompt wording guidelines	Implicit in guidelines	General art/photo prompts	Human ratings of quality	Knowledge not formalised, versioned, or gated; no task-level paired protocol
Oppenlaender 2024 [16]	Taxonomy of prompt modifiers	Community practice, descriptive	AI art	Ethnographic	Descriptive taxonomy; no compiler, no controlled experiment
Hao et al. 2023 [28]	Learned prompt rewriting (RL)	Implicit in reward and policy	Generic user prompts	Aesthetic score, human preference	Learned, opaque policy; optimises aesthetics, not causal organisation
Zhang et al. 2023 [27]	Auxiliary-image conditioning inside the model	None	Edges, pose, depth	FID/CLIP, user study	Requires structural input image; control not text-readable
Lian et al. 2024 [29]	LLM-generated layout guides diffusion	None (spatial only)	Compositional benchmarks	Accuracy on layout tasks	Layout grounding without communicative rules or domain evaluation
Chefer et al. 2023 [10]	Cross-attention steering at inference	None	Subject-binding prompts	CLIP similarity, user study	Model-internal; no design theory; no task-level evaluation
Otani et al. 2023 [14]	None (evaluation study)	None	General text-to-image	Verifiable human protocol	Protocol only; adopted and extended here with a reliability battery
Xu et al. 2023 [31]	None (reward model)	None	Generic prompts	Learned human-preference score	Generic proxy; our results show proxies can miss organisational gains
This study	Deterministic matched prompts from an explicit rule table	Explicit v0.1 records with stored rationales	Manufacturing-disruption scenarios anchored in a clustering-checked taxonomy	Paired, label-blinded expert judgment + automatic metrics + Holm correction	—

Note: Citations refer to the reference list; "design knowledge" indicates whether visual-communication knowledge is represented as an explicit, inspectable artefact.

3. Method

3.1 Study design

Figure 1 shows the pipeline as three layers. The data layer discovers, exactly deduplicates, lexically screens, records, and clusters disruption titles. The knowledge and prompt-construction layer anchors 24 scenarios in the taxonomy, stores 18 v0.1 design-prior records, and deterministically constructs paired prompts that differ only in the organisation clause. The generation and evaluation layer produces matched image pairs and evaluates them with automatic metrics and independent expert paired judgment. Two analysis levels are kept separate throughout: the 758-title corpus supports an exploratory check of taxonomy structure (RQ1), while the 24 anchored scenarios support the controlled visual experiment (RQ3, RQ4); the corpus is never treated as a sample of the experiment, nor vice versa.

The three layers are connected by three recorded data hand-offs. (i) The data layer exports the screened corpus as a versioned CSV and stores candidate and screening rows in SQLite; the category field is then used to select four scenarios per category. (ii) Each scenario is represented by a seven-field semantic record (trigger, exposed node, affected product, first-order impact, propagation path, consequence, and response). A deterministic prompt-construction script emits the D/G prompt pair and writes the verbatim prompt, condition, seed, SHA-256 hashes, model, and generation parameters to a CSV run manifest. The SQLite prompt_runs table was defined but contains no rows for this experiment, so the manifest—not that table—is the run-level provenance record. (iii) The generation and evaluation layer writes 96 raw images, 96 uniformly framed cards, per-card metric rows, and blinded expert judgments. No generated image or evaluation result was used to alter the corpus or the v0.1 rule records.

The experiment is deliberately narrow: a single manipulated factor, one base model, two fixed seeds, and a

uniform presentation frame. This mirrors a controlled mechanism test rather than a benchmark: with every other generation choice held constant, any systematic difference between conditions is attributable to the explicit causal-organisation grammar, at the acknowledged cost of generality across models and settings. Analysis code, prompts, and

logged artefacts support replication of every number reported below; the complete rule set, all constructed prompt pairs, generation settings, probe texts, rating forms, and the judgment cross-tabulation are provided as Supplementary Material (Tables S1–S4).

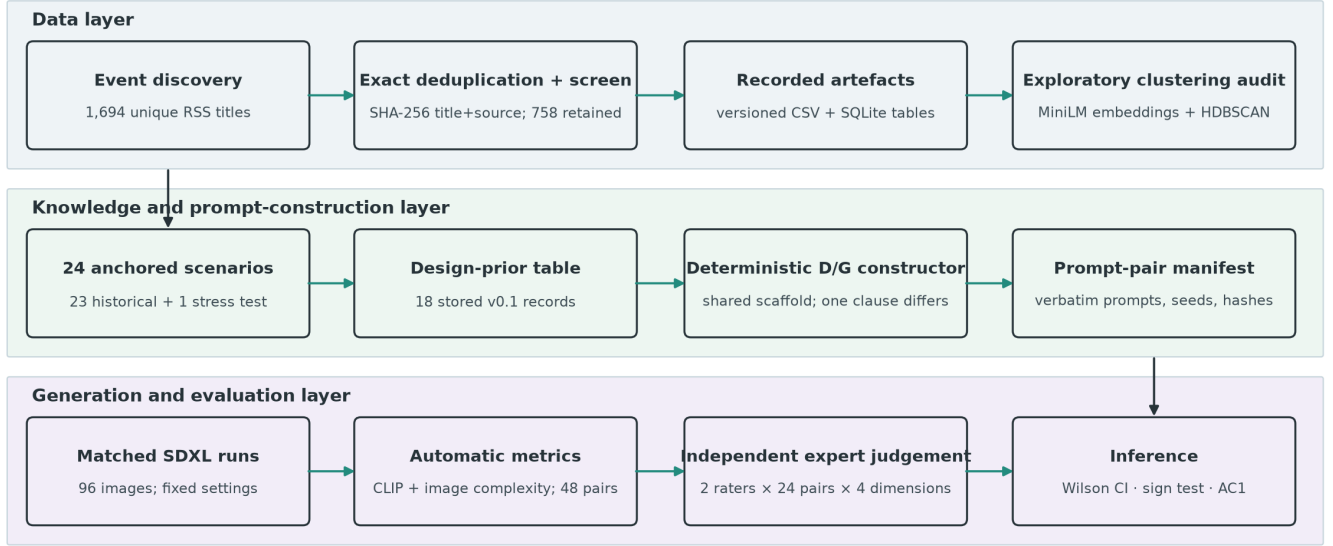


Figure 1. End-to-end computational pipeline

Note: The design-prior database is independent of the narrative audit database; the taxonomy anchors scenario sampling and constructed prompt pairs drive matched generation.

3.2 Corpus discovery, screening, and audited storage

Candidate narratives were harvested from public RSS news feeds. Each candidate d_i is a record (title, summary, date, source, URL). Automatic relevance screening assigns

$$R_i = 0.35 M_i + 0.35 D_i + 0.20 C_i + 0.10 Q_i, \quad \text{retain } d_i \Leftrightarrow R_i \geq \tau_R = 0.55 \quad (1)$$

where M_i , D_i , $C_i \in \{0, 1\}$ indicate the presence of manufacturing-domain, disruption, and causal/action vocabulary, respectively, and Q_i is a source-quality score. Matching is case-insensitive against the lists in Supplementary Table S1. In the executed script, $Q_i = 1.0$ when the stored source-name string ends in .gov or .int (or contains .gov.), $Q_i = 0.8$ when it ends in or contains .org, and $Q_i = 0.6$ otherwise. With weights (0.35, 0.35, 0.20, 0.10) and $\tau_R = 0.55$, both manufacturing and disruption evidence always retains a record ($R_i \geq 0.76$ under the implemented Q range), as does exactly one of those indicators together with causal/action evidence ($R_i \geq 0.61$); a record carrying neither has $R_i \leq 0.30$. Thus Q affects the score but not the inclusion decision bands in this dataset. The 1,694 unique candidate

rows yielded 758 retained rows. Their unit-normalised 384-dimensional embeddings were produced with sentence-transformers/all-MiniLM-L6-v2 [39].

$$z_i = f_\theta(d_i) / \|f_\theta(d_i)\|_2, \quad \delta_{ij} = 1 - z_i^\top z_j, \quad \delta_{ij} < \tau_d = 0.08 \Rightarrow \text{near-duplicate} \quad (2)$$

Equation (2) defines the cosine distance used to describe the embedding geometry. The executed acquisition script did not apply a semantic near-duplicate threshold. Instead, exact collection duplicates were collapsed before screening by SHA-256 hashing lower-case title plus lower-case source name; the pre-collapse item count was not retained, so no semantic duplicate-removal count is reported. The analysed CSV contains 1,694 unique candidate rows and 758 retained rows. Candidate and screening records and the 18 v0.1 design rules are stored in SQLite; the final prompts and generation provenance are stored in the 96-row run manifest. This distinction is stated to avoid implying that the empty prompt runs table was populated.

The category field for the 758 retained records was assigned automatically from the predeclared acquisition-query family (NAT, FAC, LOG, GEO, DEM, or CYB). If an exact title–source pair was returned by more than one query

family, the collector retained all query tags, while the screened export used the first query-family tag. No item-wise human recoding, double-coded subset, or inter-rater agreement study was performed. The category counts therefore describe a query-stratified organising taxonomy rather than adjudicated event labels; the definitions and boundary notes in Supplementary Table S1 support interpretation, and this limitation is carried into Section 5.

3.3 Embedding and density-based clustering of the taxonomy

A six-category taxonomy was specified before clustering: natural hazard (NAT), facility/equipment failure (FAC), cyber incident (CYB), logistics and transport blockage (LOG), geopolitical and trade action (GEO), and demand/labour/market shock (DEM) (Table 2). To test whether these query-derived labels correspond to latent semantic structure (RQ1), the 758 unit-normalised 384-dimensional embeddings were clustered with HDBSCAN [40, 41] (sklearn.cluster.HDBSCAN in scikit-learn 1.9.0) using `metric='euclidean', min_samples=5, min_cluster_size=22,` and the default `cluster_selection_method='eom'`. On unit vectors, squared Euclidean distance is a monotone transform of cosine distance ($\|z_i - z_j\|^2 = 2\delta_{ij}$). HDBSCAN, a hierarchical extension of density-based clustering [42], requires no preset cluster count and returns an explicit noise label. The complete algorithmic workflow is as follows. First, for each point x the core distance is the distance to its k -th nearest neighbour, where k is the `min_samples` smoothing parameter and m denotes the minimum cluster size; `min_samples = 5` was fixed as a low smoothing value, and m was tied to corpus size so that the smallest admissible cluster is 3% of records at the primary setting,

$$d_{\text{core}}(x) = \delta(x, x^{(k)}), \quad k = \min_samples = 5, \quad m = \max\{5, \lfloor 0.03N \rfloor\} = 22 \quad (3)$$

from which the mutual-reachability distance between any two points a and b is

$$d_{\text{mr}}(a, b) = \max\{d_{\text{core}}(a), d_{\text{core}}(b), \delta(a, b)\} \quad (4)$$

Second, a minimum spanning tree is built on the mutual-reachability graph and cut at decreasing density thresholds $\lambda = 1/d_{\text{mr}}$, producing a condensed hierarchy in which components smaller than the minimum cluster size m are treated as fall-out rather than splits. Third, each candidate cluster C receives a stability score

$$\sigma(C) = \sum_{x \in C} (\lambda_x - \lambda_{\text{birth}}(C)) \quad (5)$$

where $\lambda_{\text{birth}}(C)$ is the density at which C appears and λ_x the density at which point x leaves C ; the final flat clustering selects the set of non-overlapping clusters maximising total stability, and all unassigned points receive the noise label -1 . Cluster quality among non-noise points is described by the mean silhouette [43],

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \quad \bar{s} = \frac{1}{N} \sum_i s(i) \quad (6)$$

and correspondence between the clustering C and taxonomy labels Y by normalised mutual information [44] and the adjusted Rand index [45]:

$$\text{NMI}(C, Y) = \frac{2I(C; Y)}{H(C) + H(Y)} \quad (7)$$

$$\text{ARI} = \frac{\sum_{ij} \binom{n_{ij}}{2} - E}{\frac{1}{2}[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}] - E} \quad (8)$$

where n_{ij} are contingency counts, a_i and b_j their margins, and E the expected index under random labelling. Sensitivity of the flat clustering was assessed at `min_cluster_size` $\in \{[0.02N], [0.03N], [0.05N]\}$. Clustering here audits the taxonomy; it does not replace it.

HDBSCAN is chosen over centroid methods for three properties that match short, heterogeneous news narratives: it does not force every record into a cluster, so boilerplate-heavy or ambiguous items fall into an explicit noise class instead of distorting cluster shapes; it accommodates clusters of varying density, as expected when some disruption types share vocabulary (geopolitical and demand narratives) while others are lexically distinctive (cyber incidents); and its principal parameter has a transparent interpretation as the smallest admissible cluster, enabling the predeclared sensitivity grid rather than post-hoc tuning.

A high noise fraction is expected rather than anomalous for this material. Screened news titles are short, originate from heterogeneous outlets, and share boilerplate wording; a general-domain sentence encoder can place such records between dense lexical regions. HDBSCAN reports them as noise rather than forcing them into a cluster. Correspondence indices are therefore computed on the non-noise subset and describe only the separable core of each query category, while the noise fraction is reported as a finding. Because clustering audits rather than creates the organising taxonomy, the objective was not to minimise noise. The predeclared sensitivity grid reported in Section 4.1 is the mitigation applied; UMAP reduction and soft membership are left for future work because neither was used in the executed pipeline.

Table 2. Theory-derived disruption taxonomy, screened-corpus counts, and anchored scenarios

Code	Category	Working definition (inclusion focus)	Corpus n	Scenarios
NAT	Natural hazard	Earthquake, flood, typhoon, or other natural events disrupting production or supply	160	4
CYB	Cyber incident	Ransomware, intrusion, or IT/OT outage affecting manufacturing operations	148	4
DEM	Demand / market shock	Abrupt demand collapse or surge, order cancellation waves, market dislocation	128	4
GEO	Geopolitical / trade action	Export controls, sanctions, tariffs, or conflict-related supply restrictions	128	4
FAC	Facility / equipment failure	Fire, explosion, breakdown, or quality event inside a plant	112	4
LOG	Logistics / transport blockage	Port congestion, route closure, carrier failure, or freight disruption	82	4
Total			758	24 (23 historical + 1 stress test)

3.4 Prior-knowledge import: an independent design-prior database

Art and information-design knowledge enters the pipeline as an explicit 18-row prior table (Figure 2a; Table 3). Equations (9)–(10) describe a general admission specification for future growth of that table: candidate actions should be visually mappable, operationalisable, and non-redundant. The current v0.1 release, however, was entered directly before generation, and the archive contains neither a larger candidate set nor an executed automated admission log. Accordingly, the equations are a framework specification rather than a retrospective claim about how the 18 rows were selected.

$$\text{admit}(u_k) = \mu_k \cdot o_k \cdot \mathbf{1}[\rho_k < \tau_\rho] \quad (9)$$

$$\rho_k = \max_{r \in \mathcal{D}} z_k^\top z_r \quad (10)$$

In the general specification, $\mu_k \in \{0,1\}$ denotes visual mappability, $o_k \in \{0,1\}$ operationalisability, and ρ_k similarity to existing rules. No numeric τ_ρ was executed or recorded for the current 18-row release, so the study does not report candidate rejection counts or a rule-deduplication statistic. A stored rule is represented by the tuple of Eq. (11).

$$r_k = (g_k, c_k, a_k, w_k, h_k, p_k) \quad (11)$$

The executed design_rules table stores rule_id, semantic_field g_k , activation_condition c_k , visual_action a_k , weight $w_k \in [0,1]$, hard flag $h_k \in \{0,1\}$, rationale, and version='0.1'. It does not instantiate separate rule_provenance or rule_versions tables. Although the schema defines prompt_runs, that table contains zero rows for this experiment; run provenance is instead carried by run_manifest.csv. The 18 stored rules cover trigger, exposed node, propagation, consequence, response, hierarchy, palette, style, figure-ground organisation, negative constraints, and complexity. Supplementary Table S1 reports the stored fields and explicitly marks non-encoded prohibition/conflict/provenance fields. Given scenario semantics S_i , Eq. (12) expresses the conceptual activated set used to motivate the organisation directives.

$$\mathcal{A}(S_i) = \{r_k \in \mathcal{D} : c_k(S_i) = 1\} \quad (12)$$

The v0.1 rule records were present before the matched generation run, and no rule was added, removed, or reweighted in response to the generated images or ratings. This temporal separation supports interpretation of the D/G manipulation as pre-specified for the experiment, while the absence of a formal public preregistration and complete rule-development log remains a limitation.

Table 3. Design-prior rule families in the independent database

Rule family	Serves slot g	Visual action a (abridged)
-------------	---------------	----------------------------

Trigger salience	trigger	Concentrate contrast/scale on the initiating hazard element
Exposed-node identification	exposed node	Render the affected facility/asset as the primary figure
Directional propagation	propagation	Encode spread with oriented composition from trigger toward node
Consequence depiction	consequence	Show the manufacturing effect (halt, backlog, damage) as a distinct zone
Response separation	response	Isolate response/recovery cues from the consequence zone
Palette restriction	global	Constrain colours to the shared scenario palette
Figure-ground organisation	global	Suppress background detail competing with causal elements
Compositional hierarchy	global	Order visual weight along the causal reading path
Negative constraints	global	Forbid text overlays, icons, badges, photorealistic-event framing

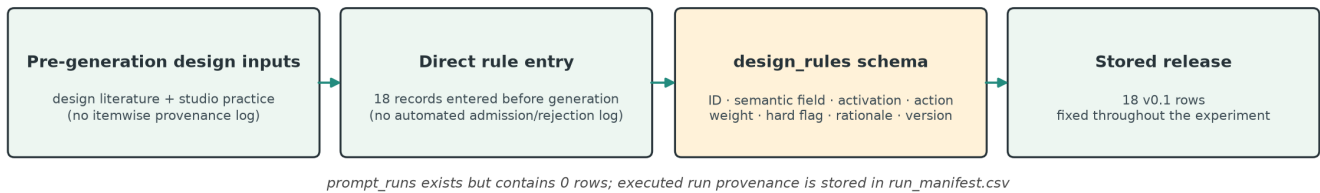
Note: the SQLite design_rules table contains 18 rules, each with semantic field, activation condition, visual action, weight, hard flag, rationale, and version='0.1'.

The stored v0.1 rule table is intentionally simple. It does not contain a pairwise conflict matrix or a separate rule_versions history table, and the prompt_runs table contains no records for the reported experiment. Reproducibility therefore rests on the frozen 18-row rule export, the generation script, and the 96-row run manifest with prompt and negative-prompt hashes. Supplementary Table S1 reports the fields that are actually present and marks fields not encoded in the schema rather than reconstructing them retrospectively.

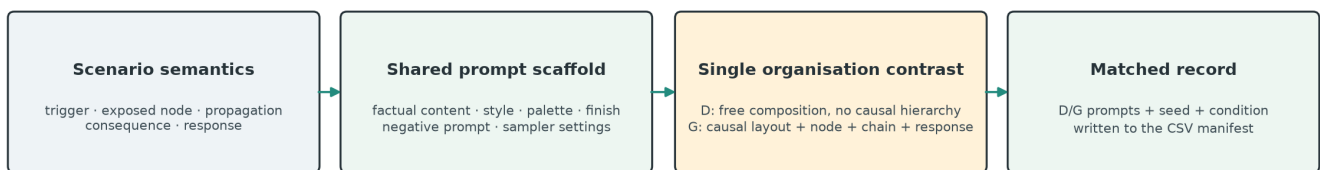
The project record establishes that 18 operational information-design rules (version 0.1) were stored before the

matched generation run. It records each rule's action and rationale but does not retain a larger candidate inventory, rejection counts, a duplicate threshold for candidate rules, per-rule bibliographic provenance, or an independent admission judgment. We therefore do not report such counts or agreement statistics. The rules should be read as a pre-generation design-prior specification informed by the cited visual-communication literature and the project's studio constraints, not as the outcome of a separately validated rule-selection study. No rule was fitted to the expert outcomes.

(a) Design-prior representation as implemented



(b) Deterministic construction of the paired prompt conditions



Executed test: all components were held constant within a pair except the organisation clause.

Figure 2. Design-prior representation and deterministic matched prompt construction

(a) The design-prior representation as implemented, including absent development logs; (b) deterministic construction of the matched D/G prompt conditions used in the experiment

3.5 Design-prior-guided deterministic prompt construction

For generalisation beyond the present experiment, Eqs. (13)–(14) define a feasible-set view in which a constructor

balances semantic coverage, prohibitions, organisation, and token budget. The executed experiment uses the deterministic special case described below rather than a search procedure.

$$\Omega(S_i) = \{P: \text{cover}(P, S_i) = 1 \wedge T(P) \leq B \wedge \text{forbid}(P) = 0 \wedge \text{conflict}(P) = 0\}, \quad B = 77 \quad (13)$$

Equations (13)–(14) state the general design target for a constraint-guided compiler: semantic coverage, a 77-position text-encoder budget, prohibition of unwanted content, and parsimonious use of design actions. In the reported bounded experiment these criteria were instantiated by a deterministic template rather than a general conflict-search engine. The factual fields and shared style/negative clauses were fixed, and the manipulated organisation string was selected by condition. Because the v0.1 schema encodes no pairwise conflicts, no conflict optimisation or tie-breaking was executed for these 24 scenarios.

$$P_l^* = \arg \min_{P \in \Omega(S_l)} L(P), \quad L(P) = \lambda_s L_{\text{sem}} + \lambda_v L_{\text{vis}} + \lambda_h L_{\text{hier}} + \lambda_c L_{\text{conf}} + \lambda_t L_{\text{tok}} \quad (14)$$

The executed construction has five reproducible steps (Algorithm 1; Figure 2b): (1) read the seven scenario fields; (2) compact each field to its first two whitespace-delimited tokens; (3) assemble a shared isometric-industrial style clause, the category accent, the same factual core, and the same finish/negative prompt; (4) insert either the D organisation string ('free composition, no causal hierarchy') or the G string ('causal layout, peripheral trigger, central node, directed chain, separate response'); and (5) generate both conditions with identical seeds and parameters while

logging the verbatim prompt and hashes to run_manifest.csv. This is deterministic and auditable, but it is not a general greedy optimiser and does not claim optimal rule allocation.

Condition D and condition G use the same deterministic constructor and differ only in the organisation string. SDXL was run through diffusers 0.39.0 and torch 2.7.1+cu128 with StableDiffusionXLPipeline and stabilityai/stable-diffusion-xl-base-1.0; no refiner was used. The checkpoint uses CLIP ViT-L/14 and OpenCLIP ViT-bigG/14 text encoders, each with a 77-position window. For audit, T(P) was computed with both archived checkpoint tokenizers including start/end special tokens; the two tokenizers gave identical counts for all 48 unique positive prompts. D ranged from 63–68 positions and G from 71–76, so no positive prompt was truncated. The common negative prompt exceeded the window and the generation log records truncation; because the same negative string was used for every D/G image, this truncation is shared across conditions and does not create a paired-condition difference. The verbatim string is supplied in Supplementary Table S3.

Algorithm 1. Executed deterministic prompt construction (per scenario S_i , condition $q \in \{D, G\}$).

- 1: read seven scenario fields and category accent
- 2: compact each field to its first two whitespace-delimited tokens
- 3: org ← 'free composition, no causal hierarchy' if $q=D$
- 4: org ← 'causal layout, peripheral trigger, central node, directed chain, separate response' if $q=G$
- 5: P ← shared style $\oplus$ org $\oplus$ compact factual core $\oplus$ shared finish
- 6: generate at seeds 271828 and 314159 with identical SDXL settings; append prompt and hashes to CSV manifest
- 7: return P

3.6 Matched image-generation experiment

The experiment manipulates one factor. Condition D (direct) uses a free-composition/no-causal-hierarchy organisation string; condition G (grammar) uses five comma-separated causal-organisation directives (causal layout, peripheral trigger, central node, directed chain, and separate response). Everything else is fixed within each scenario (Table 4): event facts, stabilityai/stable-diffusion-xl-base-1.0, EulerDiscreteScheduler, 30 denoising steps, guidance scale 6.5, 1024 × 768 output, category-specific accent, shared negative prompt, random seeds (271828, 314159), and a uniform unlabelled card frame. Twenty-four scenarios were used (four per category); twenty-three are historical anchors and DEM_004 is a constructed stress test. The full sample is

24 × 2 × 2 = 96 raw images and 96 uniform cards. An earlier presentation scheme containing condition-specific identifiers was withdrawn before formal rating; the final cards carry no condition labels, badges, or strips.

Condition D is the baseline. It contains the same compact factual core, art direction, category accent, negative prompt, sampler settings, and compositing script as G; only the organisation string differs. Supplementary Table S2 lists all 24 verbatim D/G pairs, and Figure 3 shows two illustrative style-matched pairs. Including the two special tokens, D prompts contained 65.38 ± 1.31 positions and G prompts 73.38 ± 1.31 (n = 24 each). G replaced the one D organisation clause with one causal-organisation clause containing five directives and was exactly eight tokenizer positions longer for every scenario; no G prompt reached 77 (maximum 76). Prompt length therefore co-varies with the manipulation, and a length-matched filler control was not run.

Table 4. Matched generation protocol

Controlled factor	Condition D (direct)	Condition G (grammar)
Event facts	identical	identical

Base model (SDXL base 1.0; no refiner)	identical	identical
Art direction, palette, negative prompt	identical	identical
Random seeds	271828, 314159	271828, 314159
Presentation frame	uniform unlabelled card	uniform unlabelled card
Causal-organisation clauses (sole manipulation)	absent	present
Sampler, steps, guidance, resolution	EulerDiscreteScheduler; 30 steps; guidance 6.5; 1024 × 768	EulerDiscreteScheduler; 30 steps; guidance 6.5; 1024 × 768
Prompt length, CLIP tokens (mean ± SD)	65.38 ± 1.31	73.38 ± 1.31
Baseline role	baseline (D organisation string)	treatment

Cards were composited by an identical script for both conditions: the raw generation is placed on a neutral frame with no captions, icons, or condition markers. The withdrawn earlier scheme had carried condition-specific corner badges and directional arrows; it was discarded precisely because

such marks both cue raters and contaminate the visual-complexity metrics. Withdrawal occurred before any formal rating, and no analysis in this article uses images or ratings from that scheme.

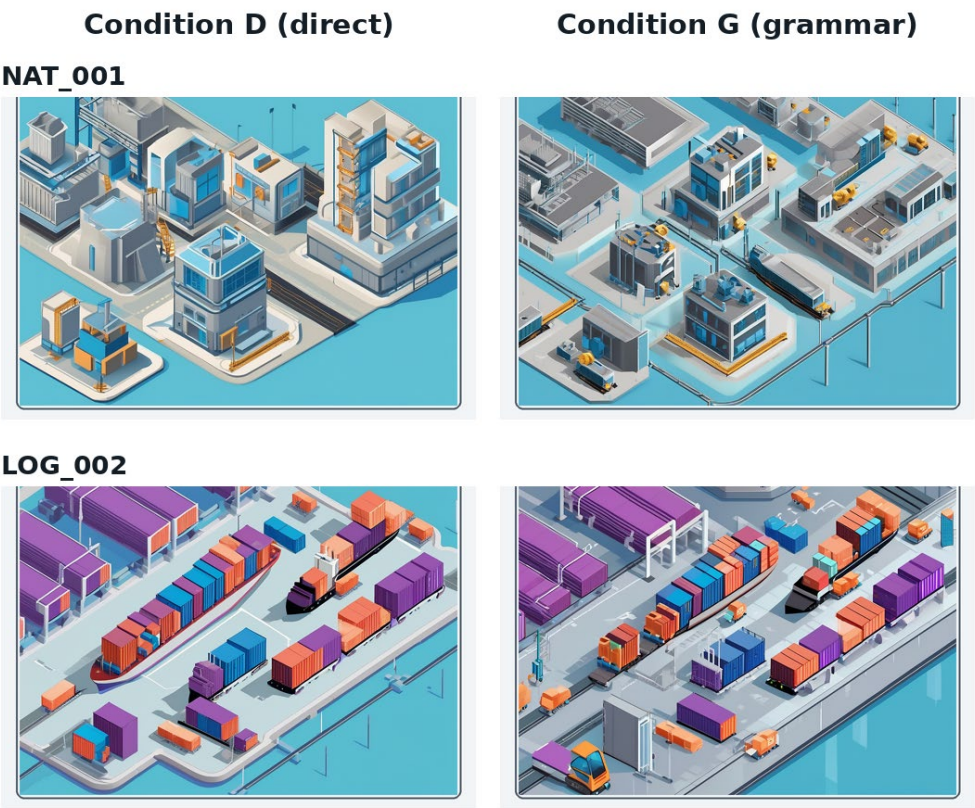


Figure 3. Two illustrative style-matched D/G pairs (generated scenario visualisations, seed 271828)

Note: The conditions share model, facts, style, palette, negatives, and framing; only the causal-organisation clauses differ. Pairs selected by automatic-metric extremes are analysed separately in the failure-atlas analysis of Section 4.4.

3.7 Human paired evaluation

Two experts independently judged the 24 D/G pairs generated with seed 271828 using separate offline Excel workbooks and a shared Word instruction sheet. Rater 1 holds a doctorate in new-media communication and specialises in

scene design and AIGC-based visualisation generation, with more than ten years of study and professional practice in the field; Rater 2 holds a doctorate in digital media and specialises in industrial and architectural design and AIGC production workflows, with more than fifteen years of study and practice. Neither rater participated in rule authoring or prompt construction. Neither rater was shown the design-prior rule table, the condition assignment, or the study

hypotheses before rating; they received only the instruction sheet describing the four judgment dimensions and rated from their own professional experience. Pair order was independently randomised for each rater (order seeds 202608082 and 202608092; maximum same-category run = 2). A/B placement was balanced 12/12 per rater and complementary between raters. The Excel medium technically permitted revisiting earlier pairs; actual revisits and session duration/breaks were not logged. Instructions required judgments from visible content only, prohibited guessing the generation method, and offered A, B, or T (no clear difference) on content completeness, causal-chain clarity, visual-hierarchy clarity, and meeting-communication usefulness. The scenario ($n = 24$) is the inferential unit for consensus sign tests; 192 judgments are used descriptively, and the 96 scenario–dimension pairs are the unit for agreement coefficients. Ties are excluded from preference rates/sign tests but retained in agreement statistics.

Expert judgment is subjective by nature; the protocol reduces the influence of that subjectivity on the *comparison* without pretending to remove it from the ratings. Six safeguards were built in: (i) the design is paired, so each rater compares two cards of the same scenario and rater-specific taste cancels within the pair; (ii) left–right placement was balanced and tested for position bias; (iii) the sheets and cards carried no event identifiers, categories, condition labels, or hypotheses, and the raters had not seen the rule table; (iv) an explicit tie option removed the pressure to force a preference; (v) consensus was computed under two predeclared definitions (lenient and strict) and both are reported; and (vi) rater agreement is quantified by a battery of chance-corrected coefficients and by the directional-agreement rate, and every disagreement is retained in the tables. What the protocol cannot remove is shared taste between two raters; that residual is the reason the conclusions are framed as expert-judged communication value rather than as a property of the images.

Judgments are coded $y_{i,r,d} \in \{-1, 0, +1\}$ for direct, tie, grammar. For each dimension the grammar preference rate among non-ties is $p_G = n_G/(n_G + n_D)$ with a Wilson 95% interval [46], chosen because the proportions lie near 1 with non-tie counts of 32–41, a regime in which the Wald interval is inaccurate. Event-level consensus is computed under two predeclared definitions: lenient—an event counts toward a condition when at least one rater prefers it and the other does not prefer the opposite; strict—both raters prefer the same condition. Non-tied consensus directions are tested with an exact two-sided sign test on the 24 events. Because the preference distribution is highly unbalanced, chance-corrected agreement is reported as a battery rather than a single coefficient: Cohen's κ [47] is the conventional two-rater statistic and is reported for comparability; Krippendorff's α generalises to more raters and missing data and is reported because the protocol is designed to scale; and Gwet's AC1 is reported because κ and α are depressed by the prevalence paradox when almost all judgments fall in one category [48]. With observed three-category agreement p_o and rater marginals π_1, π_2 , Cohen's κ corrects by the product-marginal chance rate,

$$\kappa = \frac{p_o - \sum_c \pi_1(c) \pi_2(c)}{1 - \sum_c \pi_1(c) \pi_2(c)} \quad (15)$$

nominal Krippendorff's α [49] compares observed to expected disagreement, $\alpha = 1 - D_o / D_e$, and Gwet's AC1 [50] replaces the chance term with one robust to unbalanced prevalence,

$$AC1 = \frac{p_o - p_e^G}{1 - p_e^G}, \quad p_e^G = \frac{1}{K-1} \sum_c \bar{\pi}(c) (1 - \bar{\pi}(c)) \quad (16)$$

with $K = 3$ categories and $\bar{\pi}(c)$ the pooled marginal. The directional agreement rate conditions on both raters choosing a direction. κ and α are known to be depressed under extreme prevalence even when raters almost always point the same way [48]; reporting the battery bounds the reliability claim from both sides.

3.8 Automatic metrics and failure analysis

For all 48 scenario–seed pairs, CLIP image–text similarity is computed on the uniformly composited cards with open_clip 3.3.0, model ViT-B-32, and weights laion2b_s34b_b79k. The model's standard preprocessing resizes by the shortest side with bicubic interpolation, centre-crops to 224×224 , converts to RGB, and normalises with the checkpoint mean and standard deviation. The four probe strings are used verbatim (trigger, exposed node, manufacturing consequence, response option), with no 'a photo of' wrapper. Image and text embeddings are L2-normalised; their dot product is cosine similarity and is reported $\times 100$. It is not CLIPScore. Scores are computed per card; G–D differences are formed within scenario–seed pairs and the two seeds remain separate ($n = 48$). Two descriptive complexity measures are also computed: edge density is the fraction of pixels above 32 after Pillow FIND_EDGES on an 8-bit greyscale card, and luminance entropy is Shannon entropy of a 32-bin greyscale histogram. These are complexity descriptors, not quality scores.

$$e(I) = \frac{| \{ p: \text{Edge}(I)_p=1 \} |}{|I|}, \quad H(I) = - \sum_b q_b \log q_b \quad (17)$$

with $\text{Edge}(I)$ the binary FIND_EDGES map thresholded at 32 and q_b the normalised frequency of the 32 greyscale bins. Each metric is summarised as the grammar-minus-direct paired difference with a bootstrap 95% interval (10,000 resamples), a Wilcoxon signed-rank p-value, and the paired effect size $d_z = \text{mean}(\Delta)/\text{SD}(\Delta)$. Seven paired tests are reported in Section 4.2. The mean-field similarity was the single predeclared primary automatic outcome; the four field-level probes and the two complexity descriptors are secondary. To address the multiple-comparison problem, all seven p-values are reported both raw and after Holm's step-down correction [51] applied across the family of seven, and only Holm-adjusted values are interpreted as significant. Edge density and luminance entropy are treated as descriptors of visual/structural complexity: a change in either is reported as a change in complexity and is not interpreted as an

improvement or a deterioration in visual quality unless the expert judgments point the same way. CLIP scores are semantic-alignment proxies, not measures of factual or causal correctness; distributional scores such as FID [52] require reference image sets and are not applicable to matched within-pair comparison. A failure atlas is assembled from pairs with the largest negative alignment differences and from expert-flagged pairs, characterising semantic substitution, insufficient local emphasis, decorative complexity, and weakened response cues.

4. Results

4.1 Corpus and clustering (RQ1)

The discovery stage yielded 1,694 unique candidate narratives, of which 758 met the predeclared automatic relevance threshold of Eq. (1); category counts are given in Table 2. At the primary setting ($m = 22$), HDBSCAN produced four non-noise clusters with 72.4% of records labelled noise (Figure 4). Among non-noise records the mean silhouette was 0.168, NMI with the six taxonomy labels was 0.578, and ARI was 0.456. Across the sensitivity settings (Table 5; Figure 5), cluster counts ranged from three to six, noise rates from 67.3% to 72.4%, NMI from 0.578 to 0.613, and ARI from 0.456 to 0.592. The category-cluster contingency shows interpretable concentrations—cyber narratives dominate clusters 2 and 3, geopolitical and demand narratives share cluster 0—but the high noise rate and parameter-sensitive cluster count indicate moderate semantic correspondence only: the six labels are a defensible organising taxonomy, not six stable natural clusters. "Moderate" is used in the descriptive sense of the NMI and ARI scales, on which 0 denotes chance agreement between

partitions and 1 an identical partition, so that values of 0.46–0.61 indicate that roughly half of the label structure is recovered by the density structure of the embedding space; there is no accepted universal cut-off, and the term is a reading of the scale, not a pass/fail verdict.

The contingency structure (Figure 5b) is informative beyond the aggregate indices. Cyber narratives are the most lexically separable—clusters 2 and 3 are almost purely CYB—consistent with a distinctive ransomware/IT vocabulary. Geopolitical and demand narratives concentrate jointly in cluster 0, suggesting a shared trade-and-market register that the taxonomy separates on causal grounds rather than lexical ones. Natural-hazard records, despite being the largest category, remain mostly in noise, reflecting the heterogeneity of hazard reporting. These patterns justify retaining the six-way taxonomy as the anchoring scheme for scenario sampling while cautioning against any category-frequency claims from this corpus. Four causes of the moderate correspondence can be named: the general-domain sentence encoder was not adapted to disruption reporting; news items are short and boilerplate-heavy, which compresses their embeddings towards a common region; GEO and DEM share a trade-and-market register that the taxonomy separates on causal rather than lexical grounds; and the taxonomy is defined by the *initiating event* whereas reporting is often organised around the *consequence*. Approaches that would raise correspondence—fine-tuning the encoder on labelled disruption text, seeding a semi-supervised clustering with the taxonomy, or reducing the space with UMAP before clustering—were deliberately not applied, because each would move the audit towards the taxonomy it is meant to check; they are listed as future work in Section 5. A non-optimal clustering is therefore not a deficiency of the framework: the framework uses the taxonomy, which the audit supports as an organising scheme, and never uses the clusters.

Table 5. HDBSCAN parameters and sensitivity of the flat clustering ($N = 758$; $\text{min_samples} = 5$)

min_cluster_size	Clusters k	Noise rate	Silhouette (non-noise)	NMI	ARI
[0.02N] = 15	6	71.8%	0.196	0.613	0.514
[0.03N] = 22 (primary)	4	72.4%	0.168	0.578	0.456
[0.05N] = 37	3	67.3%	0.142	0.587	0.592

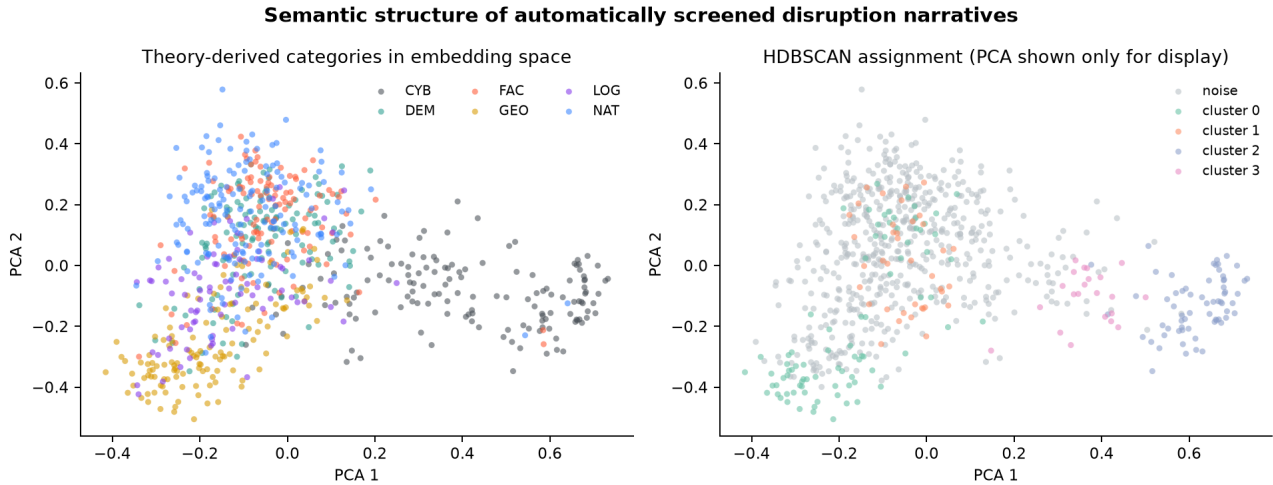


Figure 4. Semantic structure of the automatically screened narratives: theory-derived categories (left) and HDBSCAN assignment (right) in a PCA projection used for display only

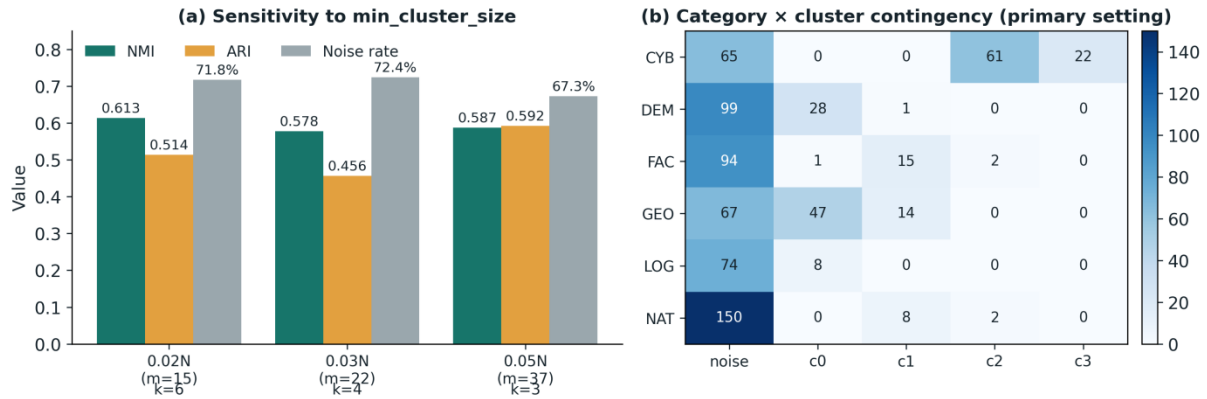


Figure 5. Clustering audit:

(a) Sensitivity of cluster count, NMI, ARI, and noise rate to min_cluster_size; (b) category x cluster contingency at the primary setting

4.2 Generation completion and automatic metrics

All 96 raw images and 96 uniform cards completed under both conditions. The causal-grammar condition did not improve aggregate CLIP alignment (mean field difference -0.095 , 95% CI $[-0.473, 0.283]$, Wilcoxon $p = .696$, $d_z = -0.07$), and no field-level alignment difference reached significance; the largest, for the manufacturing-consequence probe, was negative and non-significant (-0.439 , 95% CI $[-0.908, 0.009]$, $p = .100$). By contrast, grammar cards showed higher edge density ($\Delta = 0.0181$, 95% CI $[0.0141, 0.0228]$, $p < .001$, $d_z = 1.13$) and higher luminance entropy

($\Delta = 0.100$, 95% CI $[0.024, 0.172]$, $p = .004$, $d_z = 0.37$) (Table 6, Figure 6). After Holm correction across the seven tests, the two complexity effects remain significant (adjusted $p < .01$ and $p = .024$) and no similarity difference approaches significance (all adjusted $p \geq .50$). Explicit causal organisation therefore changed measurable image structure without shifting semantic-similarity proxies; the increases in edge density and luminance entropy are reported as changes in visual/structural complexity, a plausible consequence of adding organisational elements to a frame, and are not interpreted as improved or degraded visual quality. The direction in which that complexity is experienced is a question for the expert judgments of Section 4.3, not for these descriptors.

Table 6. Automatic metrics for the 48 matched scenario-seed pairs (bootstrap 95% intervals; Holm-adjusted p across the seven tests)

Metric (grammar – direct, n = 48)	Δ	95% CI	Wilcoxon p	d_z	Holm p
Mean field CLIP	-0.095	[-0.473, 0.283]	.696	-0.07	1.00
CLIP: hazard / trigger	0.117	[-0.394, 0.610]	.718	0.07	1.00
CLIP: exposed node	0.025	[-0.552, 0.609]	1.000	0.01	1.00
CLIP: manufacturing consequence	-0.439	[-0.908, 0.009]	.100	-0.26	.50
CLIP: response option	-0.081	[-0.684, 0.498]	.867	-0.04	1.00
Edge density	0.0181	[0.0141, 0.0228]	<.001	1.13	<.01
Luminance entropy	0.100	[0.024, 0.172]	.004	0.37	.024

Category-level inspection (n = 8 pairs per category, reported descriptively only) shows the edge-density increase to be consistent in direction across all six categories, whereas field-level CLIP differences vary in sign—positive on average for FAC, negative for the CYB consequence probe—with

standard deviations comparable to the means. The structural effect of the grammar is thus uniform, while its semantic-proxy footprint is category-dependent and unstable at this sample size.

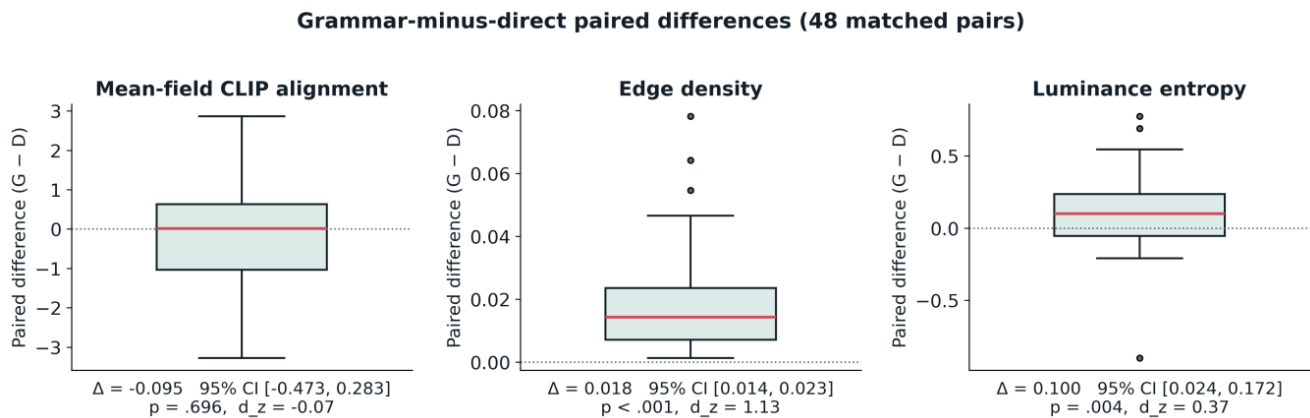


Figure 6. Automatic metric comparison across the 48 matched pairs: grammar-minus-direct differences with bootstrap 95% intervals

4.3 Expert paired evaluation (RQ3)

Table 7 summarises the 192 double-rated judgments. Among non-tied judgments the grammar condition was preferred in 97.6% (40/41) for content completeness, 100% (33/33) for causal-chain clarity, 100% (32/32) for visual-hierarchy clarity, and 97.4% (37/38) for meeting usefulness. Under the lenient consensus definition, 21, 20, 20, and 21 of 24 events favoured grammar with zero events favouring direct (all exact sign tests $p < .001$); under the strict definition the counts are 18, 13, 12, and 15 grammar versus 0 direct (all $p < .001$), with the remainder ties or disagreements. Across all 96 double-rated items, exact three-category agreement was 72.9%; Cohen's κ was 0.307 and nominal Krippendorff's α was 0.310, while Gwet's AC1 was 0.663; when both raters made a directional choice, they agreed on the preferred condition in 96.7% (58/60) of cases. The divergence between α and AC1 is the expected signature of a highly unbalanced preference distribution [48], and all indices are therefore reported

together. Left-right position preference was non-significant for both raters (binomial $p = .556$ and $p = .125$). Figure 7 visualises the preference and reliability pattern. For reconciliation, the counts are stated at every level. Per rater: R1 gave 70 grammar, 2 direct, and 24 tie judgments and R2 gave 72, 0, and 24, so the pooled totals are 142 grammar, 2 direct, and 48 tie judgments ($192 = 24 \text{ scenarios} \times 4 \text{ dimensions} \times 2 \text{ raters}$). Per dimension (Table 7): $40 + 33 + 32 + 37 = 142$ grammar, $1 + 0 + 0 + 1 = 2$ direct, and $7 + 15 + 16 + 10 = 48$ tie. Per category (Section 4.4): $32 + 26 + 22 + 20 + 25 + 17 = 142$ grammar, $0 + 0 + 0 + 0 + 0 + 2 = 2$ direct, and $0 + 6 + 10 + 12 + 7 + 13 = 48$ tie. All three decompositions sum to the same 192 judgments; the full category $\times$ dimension $\times$ rater cross-tabulation from which they are drawn is given in Supplementary Table S4 (pooled) and Tables S4a–S4b (rater-specific); every marginal was checked against the source judgment table. The 96 items used for agreement are the 24×4 scenario-dimension pairs; exact agreement on 70 of them (72.9%) decomposes into 58 pairs on which both raters chose the same direction and 12 on which both chose

tie, consistent with each rater recording 24 ties and 12 of those ties coinciding.

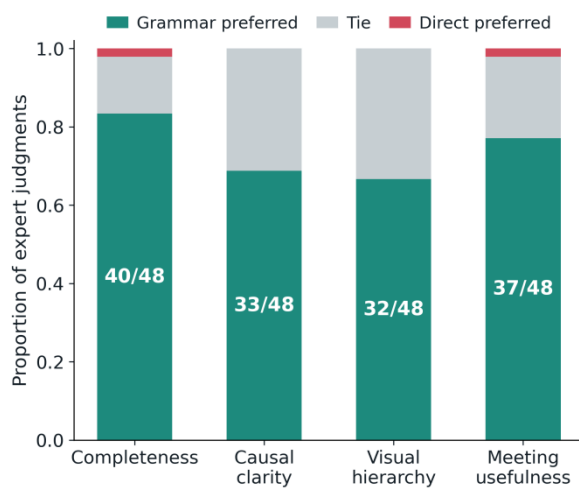
Table 7. Expert paired evaluation

Dimension	G / D / Tie	G among non-ties	Consensus lenient (strict)	Sign p
Content completeness	40 / 1 / 7	97.6%	21 G, 0 D (18 G, 0 D)	<.001
Causal-chain clarity	33 / 0 / 15	100%	20 G, 0 D (13 G, 0 D)	<.001
Visual-hierarchy clarity	32 / 0 / 16	100%	20 G, 0 D (12 G, 0 D)	<.001
Meeting usefulness	37 / 1 / 10	97.4%	21 G, 0 D (15 G, 0 D)	<.001

Rater-level patterns were symmetrical: both experts registered ties in exactly 24 of 96 judgments (25%), with R1 preferring grammar in 70 and direct in 2 of the remaining

judgments and R2 preferring grammar in all 72. The two direct preferences both came from R1 within the DEM category, anticipating the category pattern in Section 4.4.

(a) Paired expert preferences (48 judgments per dimension)



(b) Inter-rater agreement battery

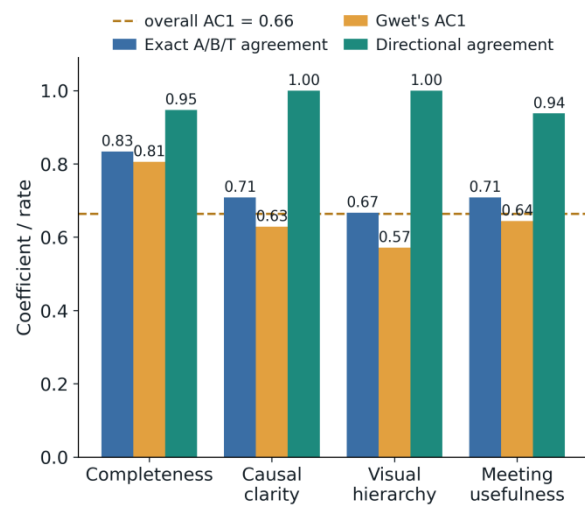


Figure 7. Expert paired preference by dimension and agreement battery (exact agreement, κ , α , AC1, and directional agreement)

4.4 Category variation and failure patterns (RQ4)

Grammar preference was robust across categories: CYB 32 grammar / 0 direct / 0 tie, FAC 26/0/6, GEO 22/0/10, LOG 20/0/12, NAT 25/0/7, and DEM 17/2/13 (each category comprises 4 scenarios $\times$ 4 dimensions $\times$ 2 raters = 32 judgments; the six rows sum to 142/2/48, matching Table 7 and Section 4.3). DEM is the boundary case: it holds the only

two direct preferences and the highest tie rate and includes the constructed stress-test scenario DEM_004, whose judgments were predominantly ties. Demand-side shocks lack a concrete physical scene, which plausibly weakens the leverage of scene-organising rules. The failure atlas (Figure 8) shows the four pairs with the largest negative alignment differences (GEO_002, CYB_001, LOG_003, DEM_004 at seed 271828), illustrating recurring modes: semantic substitution of the exposed node, insufficient local emphasis on the trigger, decorative complexity without causal information, and weakened response cues.

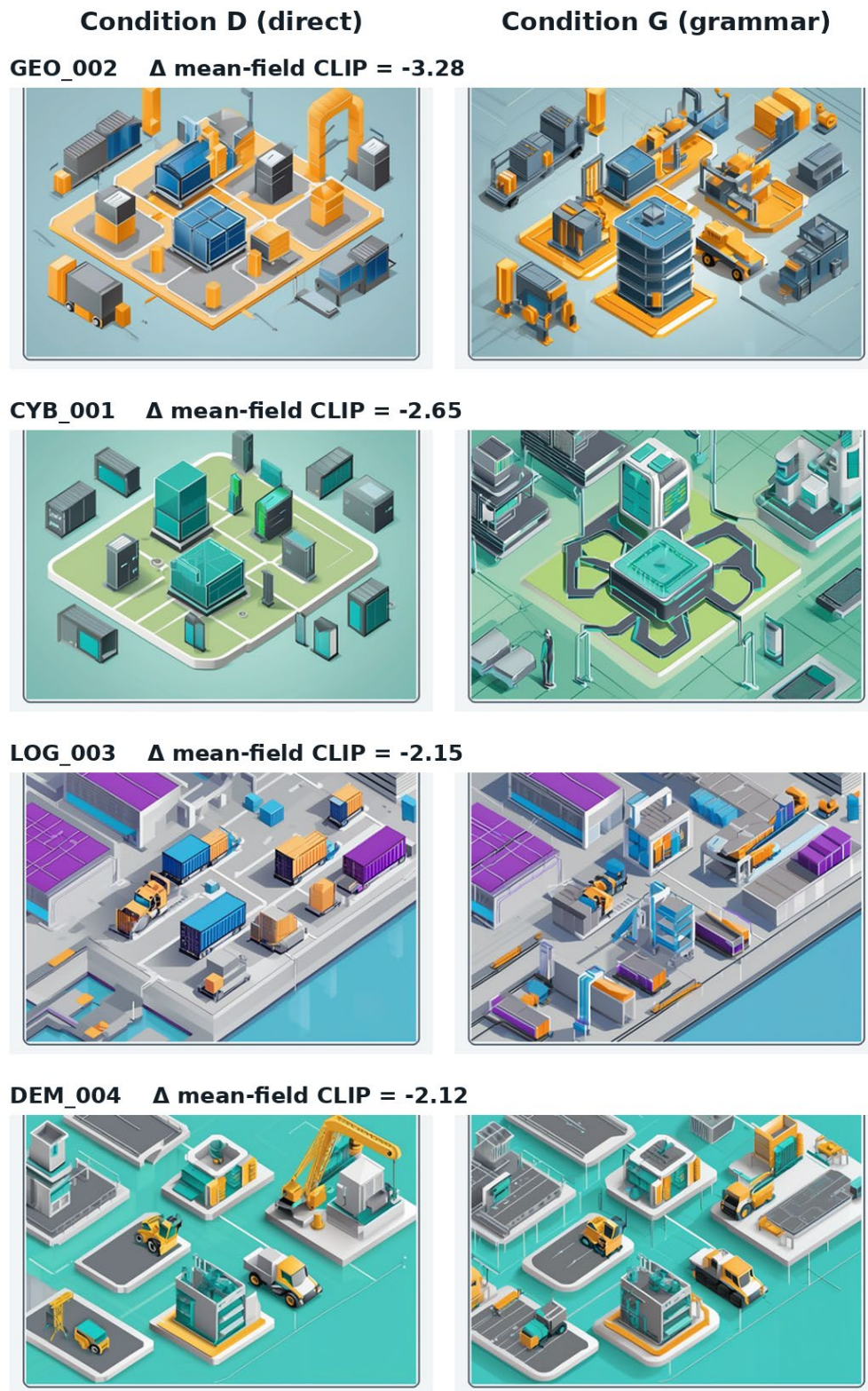


Figure 8. Representative failure candidates (generated scenario visualisations): matched pairs with the largest negative grammar-minus-direct alignment differences

5. Discussion

Design knowledge computes, but its effect is human-readable organisation rather than machine-readable similarity. The grammar condition changed image structure (edge density, luminance entropy) and was almost unanimously preferred by experts yet moved no CLIP-based alignment score. For AIGC evaluation practice this is a cautionary dissociation: task-level communication value cannot be delegated to a single text–image proxy, and pipelines tuned solely on such proxies may discard exactly the organisational qualities practitioners value [14, 22].

The reliability pattern requires honest reading. Very high directional agreement (96.7%) and a moderate-to-substantial AC1 (0.663) support a robust preference signal, while $\kappa \approx \alpha \approx 0.31$ reflects the prevalence paradox under an extreme preference distribution rather than rater randomness [48]. Two raters nevertheless cannot represent the full range of manufacturing roles; the protocol is designed to scale to larger and more heterogeneous panels.

The complexity cost is real. Explicit causal organisation added structural edges and widened luminance distributions; the negative (non-significant) consequence-field alignment suggests that organisational clauses compete with content clauses inside a fixed 77-token budget. Future compiler versions should treat structure as a budgeted resource—optimising the allocation in Eq. (14) between organisational and content clauses—rather than simply adding visual instructions. A related confound must be stated plainly: grammar prompts are longer than direct prompts by construction, so the present design cannot separate the effect of organisational *content* from the effect of prompt *length* or token allocation. The strongest available argument against a pure length explanation is that the CLIP probes did not move with length while the expert judgments did, which is the opposite of what added but uninformative tokens would be expected to produce; nevertheless, a length-matched control condition is the first experiment that should follow this one.

The application boundary is shared visual material for resilience meetings, training, and early scenario discussion. The cards are generated scenario visualisations: they may depict non-existent equipment, sites, or markings, must not be presented as photographs or forecasts of real events, and do not constitute risk monitoring, factual evidence, or automated decision-making. Within that boundary, an explicit design-prior layer offers manufacturing teams a controllable and inspectable alternative to ad-hoc prompting.

Within that boundary the practical value is concrete but limited. A resilience team can use the cards as common discussion stimuli, and trainers can reuse the taxonomy-anchored set as controlled visual cases. The run manifest makes the exact prompts, hashes, seeds, and settings inspectable. The current release does not yet provide full rule-to-source provenance or a populated prompt_runs database; those governance features are requirements for a production deployment rather than capabilities demonstrated here.

From a scalable information-systems perspective, the study demonstrates a modular file-and-database pipeline: acquisition and screening records, an embedding/clustering stage, a frozen rule table, deterministic prompt construction, a generation manifest, metric tables, and blinded judgments. These interfaces can accept additional feeds, scenarios, raters, or generation models without changing the paired comparison logic. Transfer has nevertheless been demonstrated only for SDXL, and the 77-position budget is specific to its CLIP-family text encoders. A production version should add populated provenance/version tables and transactional run logging rather than inferring scalability from the current prototype alone.

Limitations bound the claims. The corpus is automatically screened but not item-wise human-verified, so clustering is an exploratory structure check; the visual experiment covers 24 scenarios, one base model, and two seeds; the expert panel comprises two raters, and the strong preference skew depresses κ and α as discussed; CLIP alignment and complexity metrics are proxies rather than measures of factual correctness; and generated cards can depict non-existent equipment or sites and are therefore labelled generated scenario visualisations throughout. Two further limitations should be noted: the taxonomy labelling of the corpus was inherited automatically from acquisition-query families without independent double coding, and prompt length co-varies with the grammar manipulation as discussed above. The length confound qualifies the attribution of the G effect to organisational content rather than to token allocation; the remaining limitations do not alter the matched D/G comparison itself but restrict generalisation, and together they motivate the extensions listed below.

Research prospects follow directly from the limitations. (1) Length-matched prompt controls, in which the D prompt is padded to the G token count with semantically neutral filler, should separate organisational content from prompt length, and the matched protocol should be replicated across several generative models with different text encoders. (2) A larger, role-diverse rater panel drawn from engineering, operations, management, and design should test whether the preference pattern holds beyond two design-oriented experts. (3) The corpus should be independently double-coded, with inter-coder agreement reported, and the query-derived taxonomy compared with supervised alternatives. (4) The general admission/search formulation of Eqs. (9)–(14) should be implemented and tested together with explicit rule-conflict tables, rule-version control, rule-to-source provenance, data-source traceability, and populated run-log tables; the current experiment used a deterministic template. (5) Task-anchored automatic faithfulness probes should complement generic CLIP similarity. (6) Workshop studies in which the cards mediate authentic recovery discussions should evaluate task-oriented faithfulness and whether the cards shorten the time a mixed-role team needs to reach a shared understanding.

6. Conclusion

This study represented art and information-design actions as an explicit 18-row v0.1 prior table and tested a deterministic causal-organisation prompt manipulation under matched SDXL generation. RQ1 showed moderate correspondence between query-derived categories and unsupervised semantic structure under high HDBSCAN noise, supporting the taxonomy only as an organising scheme. RQ2 showed how five design actions can be operationalised transparently as one organisation clause while keeping the factual core and generation settings matched; it did not validate a general admission or optimisation engine. RQ3 found grammar preferences of 97.4–100% among non-tied expert judgments, no aggregate CLIP-similarity gain, and higher visual complexity. RQ4 identified demand/labour/market shocks as the main boundary case. These findings concern communication artefacts from one model, 24 scenarios, and two raters; they do not show improved resilience, prediction, or operational decisions. Length-matched prompt controls, multi-model replication, larger role-diverse rater panels, independently double-coded corpus labels, and workshop-based faithfulness evaluation are the immediate next steps.

Data availability

De-identified derived data (screened corpus records, clustering outputs, run manifest, expert judgment tables, and automatic metrics) and analysis scripts are available from the corresponding author upon reasonable request. A public repository was not created for this article because the complete project package contains public-news text/links and generated images with heterogeneous third-party reuse conditions. The Supplementary Material nevertheless provides the complete 18-rule table, all 48 positive prompts with token counts, generation settings and seeds, CLIP probes, rating instrument, and pooled/rater-specific judgment cross-tabulations. Generated images are labelled as generated scenario visualisations.

References

- [1] Sheffi Y, Rice JB. A supply chain view of the resilient enterprise. *MIT Sloan Management Review*. 2005; 47(1): 41-48.
- [2] Ponomarev SY, Holcomb MC. Understanding the concept of supply chain resilience. *The International Journal of Logistics Management*. 2009; 20(1): 124-143.
- [3] Ivanov D. Predicting the impacts of epidemic outbreaks on global supply chains: a simulation-based analysis on the coronavirus outbreak (COVID-19/SARS-CoV-2) case. *Transportation Research Part E: Logistics and Transportation Review*. 2020; 136: 101922.
- [4] Endsley MR. Toward a theory of situation awareness in dynamic systems. *Human Factors*. 1995; 37(1): 32-64.
- [5] Wang B, Zheng P, Yin Y, Shih A, Wang L. Toward human-centric smart manufacturing: a human-cyber-physical systems (HCPS) perspective. *Journal of Manufacturing Systems*. 2022; 63: 471-490.
- [6] Leng J, Sha W, Wang B, Zheng P, Zhuang C, Liu Q, et al. Industry 5.0: prospect and retrospect. *Journal of Manufacturing Systems*. 2022; 65: 279-295.
- [7] Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B. High-resolution image synthesis with latent diffusion models. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; 2022. p. 10684-10695.
- [8] Podell D, English Z, Lacey K, Blattmann A, Dockhorn T, Müller J, et al. SDXL: improving latent diffusion models for high-resolution image synthesis. In: *Proceedings of the International Conference on Learning Representations (ICLR)*; 2024.
- [9] Liu V, Chilton LB. Design guidelines for prompt engineering text-to-image generative models. In: *Proceedings of the CHI Conference on Human Factors in Computing Systems*; 2022. p. 1-23.
- [10] Chefer H, Alaluf Y, Vinker Y, Wolf L, Cohen-Or D. Attend-and-Excite: attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics*. 2023; 42(4): 148.
- [11] Hu Y, Liu B, Kasai J, Wang Y, Ostendorf M, Krishna R, et al. TIFA: accurate and interpretable text-to-image faithfulness evaluation with question answering. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*; 2023. p. 20406-20417.
- [12] Kamalahmadi M, Parast MM. A review of the literature on the principles of enterprise and supply chain resilience: major findings and directions for future research. *International Journal of Production Economics*. 2016; 171: 116-133.
- [13] Ivanov D, Dolgui A. Viability of intertwined supply networks: extending the supply chain resilience angles towards survivability. *International Journal of Production Research*. 2020; 58(10): 2904-2915.
- [14] Otani M, Togashi R, Sawai Y, Ishigami R, Nakashima Y, Rahtu E, et al. Toward verifiable and reproducible human evaluation for text-to-image generation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; 2023. p. 14277-14286.
- [15] Hessel J, Holtzman A, Forbes M, Le Bras R, Choi Y. CLIPScore: a reference-free evaluation metric for image captioning. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*; 2021. p. 7514-7528.
- [16] Openlaender J. A taxonomy of prompt modifiers for text-to-image generation. *Behaviour & Information Technology*. 2024; 43(15): 3763-3776.
- [17] Tang CS. Perspectives in supply chain risk management. *International Journal of Production Economics*. 2006; 103(2): 451-488.
- [18] Craighead CW, Blackhurst J, Rungtusanatham MJ, Handfield RB. The severity of supply chain disruptions: design characteristics and mitigation capabilities. *Decision Sciences*. 2007; 38(1): 131-156.
- [19] Katsaliaki K, Galetsi P, Kumar S. Supply chain disruptions and resilience: a major review and future research agenda. *Annals of Operations Research*. 2022; 319(1): 965-1002.
- [20] Ivanov D, Dolgui A. A digital supply chain twin for managing the disruption risks and resilience in the era of Industry 4.0. *Production Planning & Control*. 2021; 32(9): 775-788.
- [21] Padilla LM, Creem-Regehr SH, Hegarty M, Stefanucci JK. Decision making with visualizations: a cognitive framework across disciplines. *Cognitive Research: Principles and Implications*. 2018; 3: 29.

- [22] Franconeri SL, Padilla LM, Shah P, Zacks JM, Hullman J. The science of visual data communication: what works. *Psychological Science in the Public Interest*. 2021; 22(3): 110-161.
- [23] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. In: *Advances in Neural Information Processing Systems (NeurIPS)*; 2020. p. 6840-6851.
- [24] Ramesh A, Pavlov M, Goh G, Gray S, Voss C, Radford A, et al. Zero-shot text-to-image generation. In: *Proceedings of the International Conference on Machine Learning (ICML)*; 2021. p. 8821-8831.
- [25] Saharia C, Chan W, Saxena S, Li L, Whang J, Denton E, et al. Photorealistic text-to-image diffusion models with deep language understanding. In: *Advances in Neural Information Processing Systems (NeurIPS)*; 2022. p. 36479-36494.
- [26] Ko HK, Park G, Jeon H, Jo J, Kim J, Seo J. Large-scale text-to-image generation models for visual artists' creative works. In: *Proceedings of the 28th International Conference on Intelligent User Interfaces (IUI)*; 2023. p. 919-933.
- [27] Zhang L, Rao A, Agrawala M. Adding conditional control to text-to-image diffusion models. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*; 2023. p. 3836-3847.
- [28] Hao Y, Chi Z, Dong L, Wei F. Optimizing prompts for text-to-image generation. In: *Advances in Neural Information Processing Systems (NeurIPS)*; 2023.
- [29] Lian L, Li B, Yala A, Darrell T. LLM-grounded diffusion: enhancing prompt understanding of text-to-image diffusion models with large language models. *Transactions on Machine Learning Research*. 2024.
- [30] Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. In: *Proceedings of the International Conference on Machine Learning (ICML)*; 2021. p. 8748-8763.
- [31] Xu J, Liu X, Wu Y, Tong Y, Li Q, Ding M, et al. ImageReward: learning and evaluating human preferences for text-to-image generation. In: *Advances in Neural Information Processing Systems (NeurIPS)*; 2023.
- [32] Kress G, van Leeuwen T. *Reading images: the grammar of visual design*. 3rd ed. London: Routledge; 2021.
- [33] Bertin J. *Semiology of graphics: diagrams, networks, maps*. Redlands: Esri Press; 2010.
- [34] Tufte ER. *The visual display of quantitative information*. 2nd ed. Cheshire: Graphics Press; 2001.
- [35] Ware C. *Information visualization: perception for design*. 4th ed. Cambridge: Morgan Kaufmann; 2020.
- [36] Sweller J. Cognitive load during problem solving: effects on learning. *Cognitive Science*. 1988; 12(2): 257-285.
- [37] Brade S, Wang B, Sousa M, Oore S, Grossman T. Promptify: text-to-image generation through interactive prompt exploration with large language models. In: *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST)*; 2023. Article 96.
- [38] Wadinambiarachchi S, Kelly RM, Pareek S, Zhou Q, Velloso E. The effects of generative AI on design fixation and divergent thinking. In: *Proceedings of the CHI Conference on Human Factors in Computing Systems*; 2024.
- [39] Reimers N, Gurevych I. Sentence-BERT: sentence embeddings using Siamese BERT-networks. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*; 2019. p. 3982-3992.
- [40] Campello RJGB, Moulavi D, Sander J. Density-based clustering based on hierarchical density estimates. In: *Advances in Knowledge Discovery and Data Mining (PAKDD)*; 2013. p. 160-172.
- [41] McInnes L, Healy J, Astels S. hdbscan: hierarchical density based clustering. *Journal of Open Source Software*. 2017; 2(11): 205.
- [42] Ester M, Kriegel HP, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD)*; 1996. p. 226-231.
- [43] Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*. 1987; 20: 53-65.
- [44] Strehl A, Ghosh J. Cluster ensembles — a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*. 2002; 3: 583-617.
- [45] Hubert L, Arabie P. Comparing partitions. *Journal of Classification*. 1985; 2(1): 193-218.
- [46] Wilson EB. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*. 1927; 22(158): 209-212.
- [47] Cohen J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*. 1960; 20(1): 37-46.
- [48] Feinstein AR, Cicchetti DV. High agreement but low kappa: I. the problems of two paradoxes. *Journal of Clinical Epidemiology*. 1990; 43(6): 543-549.
- [49] Krippendorff K. *Content analysis: an introduction to its methodology*. 4th ed. Thousand Oaks: Sage; 2018.
- [50] Gwet KL. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*. 2008; 61(1): 29-48.
- [51] Holm S. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*. 1979; 6(2): 65-70.
- [52] Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In: *Advances in Neural Information Processing Systems (NeurIPS)*; 2017. p. 6626-6637.