

"Did You CoSeCh?": Faithfulness- and Coherence-Aware Auditing of Content Selection in Poster Generation

Aaditya Deshpande
TCS Research
aaditya.deshpande@tcs.com

Savita Bhat
TCS Research
savita.bhat@tcs.com

Shirish Karande
TCS Research
shirish.karande@tcs.com

Abstract

As generative models continue to evolve, the automated generation of scientific posters from research papers has advanced rapidly. However, its evaluation almost exclusively focuses on whether the poster layout is visually well-organized and whether it covers the paper's content. We argue that these approaches overlook a prior question: whether the selection of content for the poster is sound. A poster can be complete and well-formed, yet distort findings, break internal coherence, or spend its limited space on duplicated information. We introduce **CoSeCh** (**C**ontent **S**election **C**heck), a suite of automatic rule-based, LLM-assisted diagnostics that probe the selected content on four distinct dimensions of quality: faithfulness (are the qualifiers that bound a claim preserved?), coherence (do selected units remain interpretable once their dependencies are removed?), recoverability (can the paper's core claims be reconstructed from the selection?), and efficiency (is the budget spent on distinct content?). Applying CoSeCh to 5 poster generators ranging from a random selection baseline to state-of-the-art works like P2P and Paper2Poster, we find that content selection quality is strongly generator-dependent and that generators fail along different axes. Some generators preserve the qualifiers for the claims better than human authors, while others' selection is no more coherent than random choice. Existing coverage-focused evaluation metrics like PaperQuiz cannot detect these failures. The four axes are near-independent, showing that no single coverage measure can capture them.

1 Introduction

Scientific posters serve as a visual representation of research papers and aim to present key information in an engaging way. Research papers usually consist of complex structures and dense details, which require corresponding posters to summarize them intuitively and concisely. A good poster com-

bines visual and textual elements to make complex information visually appealing and easier to understand. However, manual poster generation requires domain knowledge and design expertise along with significant time and effort (Wang et al., 2024; Qiang et al., 2019). Consequently, automatic poster generation presents a critical research problem for disseminating niche knowledge and reducing the manual burden on researchers.

Early works like (Xu and Wan, 2021a; Qiang et al., 2019) extracted text and figures to heuristically arrange them in poster panels. In addition, a relevant body of work focused on sub-tasks such as layout generation (Wang et al., 2024), text summarization (Saxena et al., 2025; Gao et al., 2025), and panel attribute inference (Huang et al., 2022). However, these methods often result in posters with quality issues such as content overflow (Qiang et al., 2019). With better and more powerful generative models, recent poster generation approaches have improved significantly. Recent methods, including P2P (Sun et al., 2026), PosterForest (Choi et al., 2026), and Paper2Poster (Pang et al., 2026) use multi-agent pipelines along with task decomposition to provide focused solutions for sub-problems such as parsing, content summarization, layout planning, and rendering. This progress has centered on the generation part of the task, such as better parsing, planning, layout, and styling. Consequently, the evaluation has also been focused on the rendered poster by measuring visual quality, design, and whether the poster covers the core content of the paper, typically using VLM judges or content-recall quizzes (PaperQuiz (Pang et al., 2026)). In most cases, the evaluation overlooks the step every one of these systems performs internally and silently, i.e., content selection. The choice of which content units of the paper should appear on the poster is the critical step that influences the quality and soundness of the rendered poster.

A poster compresses the rich source content of the paper into a visually coherent single page, requiring the selection of high-quality content. Therefore, content selection is where the most consequential information decisions are made, and current evaluation approaches are blind to whether these decisions are good. Most current work performs content selection as a relatively unexamined step, and current evaluation assesses the output’s coverage and appearance, not whether the selection is distorted, hallucinated, or broken. A poster that scores well on coverage and design can still report a result stripping the bounding condition or retain a figure whose context is dropped. These are failures specific to selection and not to rendering. We argue that there exists no poster metric designed to detect them. As a result, a whole class of content errors passes the current evaluation unmeasured.

To address this gap, we introduce ‘CoSeCh’, a suite of automatic diagnostics that evaluate a generator’s selected content directly, against the source paper. CoSeCh probes four dimensions that a sound selection should have: 1) faithfulness, 2) coherence, 3) recoverability, and 4) Efficiency. We present each of them with a rule-based heuristic whose every judgment is inspectable and interpretable rather than the output of a learned model. Applying CoSeCh to five generators with ≈ 100 papers, we find that selection quality is multi-dimensional and generator-specific. The property that aligns with existing coverage-based evaluation measures barely distinguishes between the generators. Faithfulness and coherence exhibit significant variations. With these four axes near-independent, no existing coverage measure captures them.

We summarize our contributions as follows.

- We formulate content-selection quality as a distinct, measurable target for poster generation, separate from layout and rendering;
- We introduce CoSeCh, a suite of four automatic rule-based diagnostics - faithfulness, coherence, recoverability, and efficiency - each auditable and interpretable;
- We show empirically that current generators diverge from human selection distinctly and in specific ways - some exceeding human faithfulness or coherence, others no better than random — and that existing coverage-oriented evaluation cannot detect these differences.

2 Related Work

Paper-to-poster generation Early approaches to automatic poster generation relied on heuristic panel arrangement over extracted text and figures (Xu and Wan, 2021a; Qiang et al., 2019). Subsequent work decomposed the problem into sub-tasks — layout generation (Wang et al., 2024), poster-oriented summarization (Saxena et al., 2025; Gao et al., 2025), and panel attribute inference building on document-layout pretraining (Huang et al., 2022) — but stopped short of an end-to-end system. The current generation of methods instead uses multi-agent LLM pipelines that jointly handle parsing, content summarization, layout planning, and rendering: P2P (Sun et al., 2026), Paper2Poster (Pang et al., 2026), PosterForest (Choi et al., 2026), PosterGen (Zhang et al., 2025), and Any2Poster (Vinaykumar et al., 2026) all follow this template, differing mainly in how planning is decomposed across agents and how aesthetic constraints are enforced during rendering. Across this line of work, content selection is performed as an internal, unexamined step inside the planning agent(s) rather than as a task with its own success criteria.

Evaluation Evaluation has tracked the same generation-centric focus. VLM-judge protocols and layout/aesthetic metrics (Wang et al., 2024) score visual organization and design quality, while content-recall instruments such as PaperQuiz (Pang et al., 2026) and the P2PEval benchmark introduced alongside P2P (Sun et al., 2026) test whether a reader can answer questions about the paper after seeing the poster. These instruments are valuable for measuring coverage — whether the important content made it onto the page in some form — but they are not designed to detect distortion of that content: a poster can preserve enough information to pass a recall quiz while still stripping a claim’s qualifying condition, orphaning a figure from the text that contextualizes it, or duplicating a result across a table and a caption. No existing metric targets the selection step directly, independent of layout and rendering quality.

Content Selection Content selection is a critical step to create coherent and faithful posters from academic papers. Accordingly, early works treat this as a special form of summarization approaches such as probabilistic graphical models (Paramita and Khodra, 2016) and Naive Bayes classification (Qiang et al., 2016, 2019). Later, neural systems

have been proposed to extract representative text and figures from the papers in (Jaisankar et al., 2024; Xu and Wan, 2021b). Recent advances in LLM have enabled LLM-based poster generation pipelines which consist of content selection as an internal and implicit planning stage before rendering. Across all recent works like Paper2Poster (Pang et al., 2026), PosterGen (Zhang et al., 2025), P2P (Sun et al., 2026), PosterForest (Choi et al., 2026), and Any2Poster (Vinaykumar et al., 2026), content selection is optimized for coverage and visual aesthetics, or delegated to an LLM planner, and finally evaluated only through the quality of the rendered poster.

Faithfulness Several works develop automatic and human evaluation protocols for factual consistency in generated summaries. FENICE (Scirè et al., 2024) extracts atomic claims from a summary and aligns them with the source using NLI, with coreference resolution applied across multiple granularities. TRUE (Honovich et al., 2022) brings eleven factual-consistency datasets spanning summarization, dialogue, and paraphrasing into a common binary evaluation setting, and finds that large-scale NLI and QA-based metrics are the strongest and most complementary detectors of inconsistency. FABLES (Kim et al., 2024) and LongEval (Krishna et al., 2023) consider book-length summarization, using human annotations to verify claims against long source documents. These methods evaluate the faithfulness after summarizing. They do not evaluate whether the earlier selection decision of which source content should be included is faithful.

Across poster generation, evaluation, content selection, and faithfulness one gap recurs: generators treat selection as an unexamined internal step; evaluation assesses the rendered poster’s coverage rather than whether the selection itself and not summarization distorts or breaks the paper’s claims. No prior work evaluates the selection itself for faithfulness, coherence, recoverability, and efficiency — the gap CoSeCh fills with an auditable, reference-free diagnostic suite.

3 Proposed Framework

In this section, we formally define four diagnostic dimensions.

3.1 Problem Formulation

We consider a paper as a set of content units $U = \{u_1, u_2, \dots, u_n\}$, where each unit is a text block, figure, table, equation, or graph/chart, with its own text or caption, its section context, and its modality. When a poster generator performs content selection, it chooses a subset $S \subseteq U$ to place on the poster under an implicit space budget B where $\sum_{u \in S} b(u) \leq B$ and $b(u)$ is a unit’s space cost, discarding the remainder of the units. Given a source paper with units U , and a set of selected content units S , CoSeCh evaluates S along four axes that probe the soundness of a selection: 1) faithfulness - it does not distort the paper’s claims, 2) coherence - its units remain interpretable within S , 3) recoverability - its core content can be reconstructed, and 4) efficiency - the limited space budget is spent on distinct content. All probe instruments are rule-based with minimal LLM-assistance and deterministic by design. We use a language model for perception tasks such as extracting claim-qualifier pairs, mapping a generator’s selected text onto its source paragraph as a fallback-measure, answering recoverability questions, etc but never for scoring itself. This design keeps all scoring auditable and interpretable (More in Section 5.6). Each axis probe is a rule-based function $\Phi(S, U) \in [0, 1]$ and is reference-free (it is computed from the selection and the source paper alone). When a human-authored poster $H \subseteq U$ is available, we report $\Phi(H, U)$ as an oracle or upper anchor. None of the CoSeCh metrics requires it, so the suite applies to any paper, including those with no gold poster.

3.2 Faithfulness

Faithfulness measures whether the selection S preserves the qualifiers that bound a paper’s claim, or presents results without the limiting conditions. For example, if a finding "improves X under condition Y" is presented as unconditional "improves X", it is categorized as unfaithful to the source paper. To investigate this, we extract claim-qualifier pairs from the source. The extraction of qualifiers is restricted to the scope, conditional, statistical, magnitude, and assumption qualifiers. The Author-retained qualifiers in the abstract serve as a reference for faithful compression.

Formally, the faithfulness is quantified as below: Let $C(U)$ be the set of claim-qualifier pairs (c_i, q_i) where claim unit u_c and qualifier unit u_q are distinct

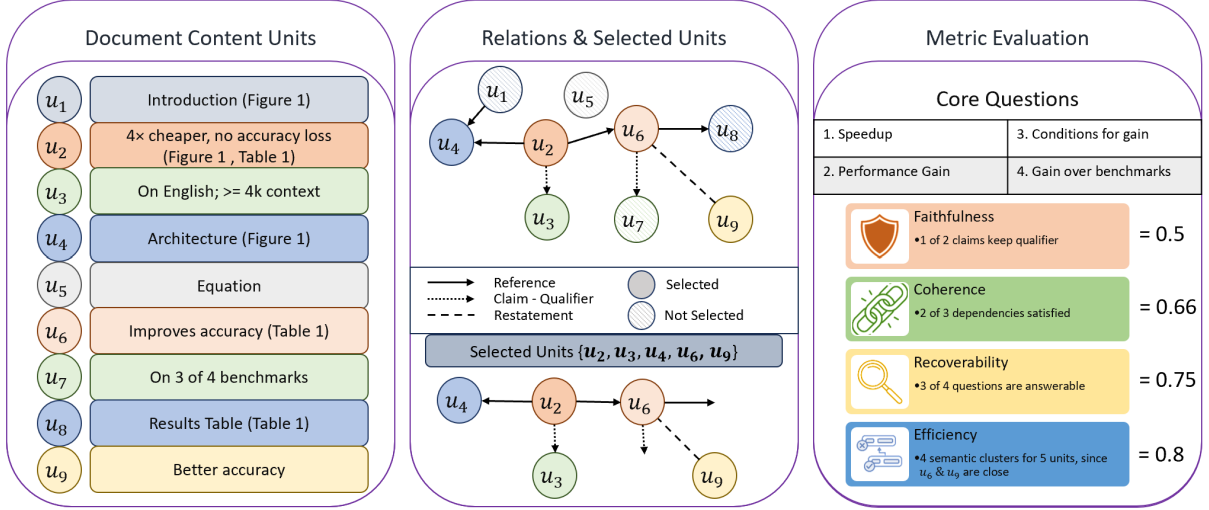


Figure 1: A worked example of the four-axis diagnostic on a toy paper. *Left*: The paper’s nine units with their type and content. *Center*: the relation graph - reference edges (solid), claim–qualifier links (dotted), restatement (dashed) — with the selected content units. *Right*: each axis scored on the selection. The scores diverge, illustrating that the axes capture distinct, non-redundant properties. Values are illustrative, not experimental results.

($u_c \neq u_q$). A qualifier is dropped when its claim is selected but the qualifier is not. We quantify the qualifier-drop rate over the selected, caveated claims:

$$D(S) = \{(c, q) \in C(U) : u_c \in S \wedge u_q \notin S\} \quad (1)$$

and the *qualifier drop_rate* is

$$QDR(S) = \frac{|D(S)|}{|\{(c, q) \in C(U) : u_c \in S\}|} \quad (2)$$

$$Faithfulness(S) = F(S) = 1 - QDR(S) \quad (3)$$

The qualifier drop rate is quantified as the ratio of selected claims without the respective qualifier not selected to the total claims in the selected content units. The denominator is the number of selected claims that have a qualifier — the claims actually at risk of being stripped — rather than all selected claims, so uncaveated claims neither inflate nor deflate the score. This is validated by manually checking extracted claim–qualifier pairs on a sample of papers for precision and separability. We report faithfulness against human reference S_h whenever available, since human posters also compress, making the comparison relative rather than absolute.

3.3 Coherence

Coherence measures whether a selected unit remains interpretable given the rest of the selection

or is left orphaned by the removal of units it depends on. For example, a figure whose defining unit is dropped or a claim citing a table that is not selected. We build a dependency graph over units using explicit cross-references where a unit mentioning a figure, table, equation, or section induces an edge to that referent.

The coherence is quantified using the dangling dependency rate, the fraction of dependencies of selected units whose referent was not selected to the total dependencies in the selected set. The coherence score is defined as 1 - dangling rate.

Formally, let $E \subseteq U \times U$ be explicit reference edges, $(u, v) \in E$ meaning u references v . Dangling dependency rate is defined as

$$DDR(S) = \frac{|\{(u, v) \in E : u \in S \wedge v \notin S\}|}{|\{(u, v) \in E : u \in S\}|}, \quad (4)$$

and coherence score is $1 - DDR(S)$. $DDR(S)$ becomes undefined when S contains no unit with an explicit outgoing dependency edge; such selections are excluded from corpus-level aggregation and their frequency is reported as a separate statistic rather than assigned a default value - assigning 0 or 1 would assert coherence or incoherence where there is in fact no evidence either way.

Edges are directed by construction — dropping u ’s prerequisite orphans u , but dropping a unit that merely refers to a selected unit does not — so only the numerator’s asymmetric condition is penalized.

We manually validate this by measuring the precision of the reference edges and the corresponding referents. We report coherence against S_h if available and the random floor.

3.4 Recoverability

Recoverability measures whether the paper’s core claims can be reconstructed from the selected content units. We use a set of core-claim questions extracted using a protocol similar to PaperQuiz. The questionnaire is answered by a reader model that is provided with only the selected content units. Recoverability is quantified as the accuracy of responses derived from the selection, and information loss is the drop from full-paper accuracy to selection accuracy. The quiz protocol also allows for a systematic categorization of the information loss based on the type of question and the corresponding units of content.

Formally, for a question set $Q(U)$ on the main claims of the paper, each question with the gold answer a_q^* , and a reader R answering question q only from the selected content, i.e. $R(q|S)$,

$$Rec(S) = \frac{1}{|Q|} \sum_{q \in Q} 1[R(q|S) = a_q^*], \quad (5)$$

with information loss $\Delta(S) = Rec(U) - Rec(S)$ against the full paper ceiling.

3.5 Efficiency

Efficiency measures how much of the space budget is allocated to distinct content. It is the opposite of *Redundancy*, which measures how much of the space budget is spent on repetitious content instead of distinct information. For example, the same result is stated in a table and is restated in text, with both the table and the text in selected units. We use pairwise semantic similarity over selected units, along with the detection of repeated quantities and restated claims. We cluster the selected units by semantic similarity (a threshold τ on pairwise similarity, plus detection of repeated quantities and restated claims), producing k clusters among the selected units $|S|$. We quantify the wasted budget as the fraction of selected units beyond one per cluster:

$$Redundancy(S) = \frac{|S| - k}{|S|}, \quad (6)$$

$$Efficiency(S) = \frac{k}{|S|} \quad (7)$$

Because some repetition is a deliberate choice for emphasis rather than a defect, we frame this as budget efficiency rather than an error.

3.6 A Worked Example

We demonstrate the formulations with a worked example presented in Figure 1. The four scores differ, and none predicts the others. For this example, this selection is moderately faithful, moderately coherent, mostly recoverable, and mostly efficient, illustrating that the axes capture distinct properties.

3.7 Independence

As mentioned above, the four diagnostics are computed independently. Further, we investigate whether they measure distinct properties or collapse to a single "*selection quality*" factor. For each pair of (generator, paper), we obtain a vector of four axis scores and compute the pairwise correlation matrix $\rho(\Phi_i, \Phi_j)$ across all instances. Low off-diagonal correlation indicates that a selection’s behavior on one axis does not predict its behavior on another. In other words, a faithful selection may still be incoherent, or a completely recoverable selection can be redundant. In such cases, the suite is not reducible to a single number. Alternatively, a high correlation on some pairs would indicate dependency. For example, aggressive compression to address redundancy may degrade faithfulness and coherence together. The claim CoSeCh makes is not that all four axes are mutually orthogonal but that they do not collapse to one dimension. We believe that the correlation matrix is the evidence for that claim. Additionally, the correlation matrix argues that existing coverage-oriented metrics, which are aligned most closely with recoverability/coverage, cannot measure the faithfulness and coherence axes as defined here.

4 Experimental Setup

4.1 Dataset

We evaluate CoSeCh on the Paper2Poster(Pang et al., 2026) dataset which contains 100 source papers paired with author-designed posters from NeurIPS, ICML and ICLR (2022-2024). These author designed posters let us report $\Phi(S_h, U)$ for every axis rather than only inter-generator comparison. A set of selected units is extracted from each generator in isolation from layout. All diagnostic axes are manually validated in a subset of papers.

4.2 Axes Implementation

Mapping generator’s selection to superset

Each generator selects and renders content in its own format which can reword, truncate, summarize the paper’s original sentence. To recover the paper’s actual, unsummarized text behind each selected unit, we map every selected unit back onto its source paragraph in the paper’s canonical content tree using a majority voting between TF-IDF, n-gram keyword matching. An LLM is called as a fallback, i.e. only when these two mappers fail to map the selected unit to any of the superset units

Faithfulness Claim–qualifier pairs are extracted from the source paper via a single LLM pass, restricted to five types: scope, conditional, statistical, magnitude, and assumption. Since a generator’s selected-content text is often a paraphrase of the source, each selected unit is first mapped back onto the paragraph it was drawn from and this original text substitutes for the generator’s, falling back to the unit’s own (summarized) form only when no confident match exists. Survival is then checked independently for each claim and qualifier in this mapped selection: a claim is present if its tokens exceed a containment threshold against the best-matching sentence, and its qualifier survives if a matching span — verbatim or via a fallback list of marker phrases for the five types — appears within a three-sentence window of that claim. A pair is scored as dropped only when the claim is selected but no such qualifier span is found in its window.

Coherence For each paper, a shared directed dependency graph is built once over the paper’s own hierarchical structure, with reference edges detected by regex over figure/table/equation/section mentions; with figure/table nodes attached from one generator’s own asset list as a shared caption source so every schema’s captions resolve against the same canonical targets. A section/subsection reference is treated as satisfied if either the heading itself or any of its descendant paragraphs was selected, since selection happens at paragraph granularity and headings are otherwise almost never selected on their own; units that fail to map onto the canonical content tree U are still scanned for their own outgoing dependencies rather than dropped, so a generator’s selection cannot appear more coherent simply because its unmapped units are hidden from the calculation.

Recoverability Each generator’s selected content units are made to answer PaperQuiz questions provided in the Paper2Poster dataset using gpt-4o. We report direct matching accuracy for each generators selected units for each paper.

Efficiency Text units are embedded with SBERT (all-MiniLM-L6-v2); all units, including visual ones, are embedded with a CLIP-style model for cross-modal comparison. A pair is flagged as redundant if cosine similarity exceeds a calibrated threshold ($\tau = 0.8$) or if the two units share an extracted numeric token, and clusters are formed by union-find over flagged pairs. The threshold was chosen by sweeping and plotting efficiency rate against threshold per generator, fixing the value at the point where rankings stabilized.

Validation For a random sample of 25 papers, we manually validate the mapping of generator’s selected content units to the canonical superset-tree, claim-qualifier pairs extracted by the LLM from the paper, referencing of the edges done to evaluate coherence, recoverability answers given by the model, and the clusters formed by similarity matching to evaluate efficiency. Table 6 reports accuracy against this single expert annotator for every component.

Please refer to appendix A for more details.

5 Results & Discussion

We present our observations on each axis, computed as a mean with 95% bootstrap confidence intervals over the corpus (Table 1), along with pairwise significance assessed with a paired bootstrap approach (Table 4 & 5). Random selection is assumed as a floor, while the *Author* poster is the human reference. We assess each generator against the human *Author* to check whether a generator matches the human reference and against *Random* to validate whether a generator is better than the floor. All generators clear the random floor on at least one axis, and most clear it on the majority. We discuss the comparison with *Author* reference in detail (Table 2).

5.1 Faithfulness

Paper2Poster emerges as the most faithful method (0.736) and significantly exceeds the human *Author* (+0.076, CI[0.022,0.131]) since it preserves qualifiers better than the *Author*. In contrast, PosterForest is significantly worse than the *Author*. All

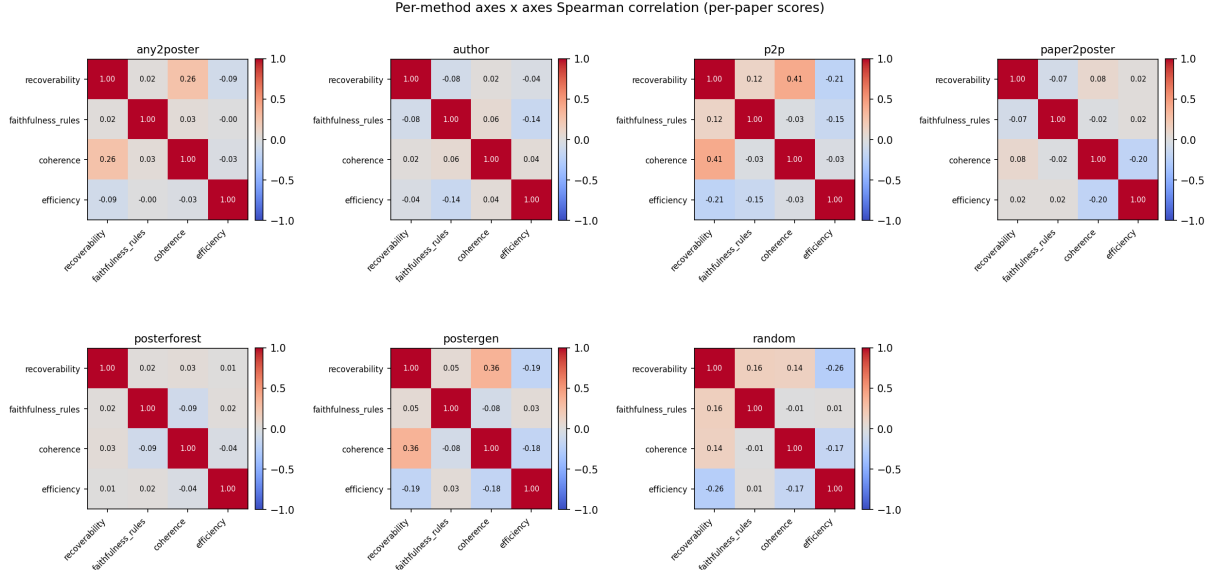


Figure 2: Spearman’s correlation among four axes for every generator. Typically $|\rho| < 0.15$ is considered as negligible.

Method	F	C	R	E
P2P	0.671	0.356	0.719	0.811
Paper2Poster	0.736	0.545	0.758	0.769
PosterForest	0.571	0.751	0.728	0.799
PosterGen	0.644	0.369	0.727	0.873
Any2Poster	0.671	0.181	0.662	0.775
<i>Author</i>	0.668	0.405	0.706	0.842
<i>Random</i>	0.543	0.141	0.646	0.731

Table 1: Per-method mean score with 95% bootstrap CIs (resampled over papers). *Author* and *Random* are references. Per-cell paper counts (n) vary by axis: faithfulness 88–98, coherence 64–86, recoverability 95–100, efficiency 95–100. F=Faithfulness, C=Coherence, R=Recoverability, E=Efficiency (1-Redundancy)

Comparison	F	C	R	E
P2P	+0.020 <i>n.s.</i>	-0.091 <i>n.s.</i>	+0.011 <i>n.s.</i>	-0.034 <i>n.s.</i>
Paper2Poster	+0.076	+0.134	+0.052	-0.073
PosterForest	-0.106	+0.316 <i>n.s.</i>	+0.020	-0.043
PosterGen	-0.023 <i>n.s.</i>	-0.019 <i>n.s.</i>	+0.022	+0.035 <i>n.s.</i>
Any2Poster	+0.001 <i>n.s.</i>	-0.275	-0.044	-0.066

Table 2: Paired bootstrap differences vs. the human *Author* (resample papers once, difference within). Positive = generator exceeds the human; *n.s.* = not significant (95% CI includes 0). Paired-paper counts (n) by axis: faithfulness 88–96, coherence 45–72, recoverability 95–100, efficiency 95–100.

generators achieve higher mean faithfulness than the random selection, though for PosterForest and PosterGen the difference is not statistically significant. Faithfulness is observed to be strongly generator-dependent, with some exceeding human and some closer in performance to random selection.

5.2 Coherence

Coherence demonstrates the clearest difference compared with *Author*. PosterForest is the strongest with (0.751). Both PosterForest and Paper2Poster significantly exceed the *Author* (+0.316 & +0.134). At the other extreme, Any2Poster is dramatically worse than *Author* (-0.275) and not significantly better than the random (+0.039, *n.s.*). In other words, Any2Poster breaks the dependen-

cies as often as blind selection. Both P2P and PosterGen match the *Author* and clear the floor comfortably.

$DDR(S)$ is undefined whenever a selection contains no unit with an explicit outgoing dependency edge (§3.3), and such selections are excluded from the coherence means reported in Table 3 rather than assigned a default score. This exclusion is not rare: across the corpus, 18% of P2P, 20% of Paper2Poster, 20% of PosterGen, and 13% of Any2Poster selections have no cross-referencing unit at all, compared to $\approx 34\%$ for PosterForest. This matters for interpretation because the excluded and included selections are not drawn from the same distribution: a selection with zero outgoing edges is, by construction, one that never attempted

a figure/table/section reference in the first place; a sparser or more self-contained subset of that generator’s behavior. PosterForest’s high undefined rate is consistent with the sparse-selection strategy discussed above: by picking fewer, more self-contained units, it both avoids dangling references on the selections it *is* scored on and removes its riskiest, most cross-referencing selections from the coherence denominator altogether. Its reported coherence of 0.751 should therefore be read as coherence *conditional on attempting cross-references*, not as evidence that PosterForest is uniformly coherent across its full output.

5.3 Recoverability

The recoverability axis reports the least distinguishable set of observations with all generators staying close to the human *Author* (all 0.66–0.76, with tight CIs). Even random reaches 0.646, near the human 0.706. We believe that this reflects both the redundancy of the core claims in papers and the generator’s optimization against coverage-oriented metrics. In contrast to coherence, the same generators span just 0.10 here, compared to 0.57 in coherence. This highlights the need for independent evaluation since optimizing for coverage secures recoverability but not other dimensions like coherence and faithfulness.

5.4 Efficiency

Similar to recoverability, efficiency (inverse redundancy) is high and tightly clustered (0.77–0.87). It is observed that humans allocate the space budget more efficiently. Although PosterGen is the most efficient (0.873), most generators are significantly less efficient than the human *Author*, and P2P and PosterGen match the human *author* efficiency. Some redundancy is deliberate, e.g. a significant result is repeated across table and text to aid the reader. Unlike a dropped caveat or a dangling reference, redundant content does not distort the paper, so we report efficiency as a context.

All depends on selection? Each generator exhibits a distinct profile towards diagnostic axes. Paper2Poster exceeds the human on faithfulness, coherence, and recoverability and is the strongest overall. PosterForest is the best on coherence, but is significantly below the human on faithfulness. Overall, Any2Poster is the weakest performer, lower than the human on coherence, recoverability, and efficiency, and no better than random on

faithfulness. That generators diverge from human selection on different axes is the central argument for a multi-dimensional suite. We believe that these distinct profiles appear because of the way generators select the content. Any2Poster is as coherent as a random selection points towards a process insensitive to inter-unit dependencies. In contrast, PosterForest has the strongest coherence and the weakest faithfulness. It also has the highest coherence-undefined rate ($\approx 34\%$ of selections contain no cross-referencing unit). By selecting sparsely, it commits to fewer dependencies and avoids dangling references, but also drops more qualifying units in the process. Improving one axis (coherence via sparser selection) can thus degrade another (faithfulness), and no single metric can reveal this tradeoff.

Implications The results highlight the drawback of the existing coverage-focused evaluation methodology. A generator trained only on coverage can bypass skipped qualifying conditions and orphan references, and discrepancies introduced because of this go unpenalized. Faithfulness- and coherence-aware evaluations should accompany coverage, and CoSeCh provides a complementary, auditable scoring mechanism.

5.5 Independence

To validate the independence of diagnostic axes, we compute Spearman’s correlation among the four axes for every generator. We find that the correlations are near zero almost everywhere (see Figure 2). The strongest correlation is ≈ 0.4 between coherence and recoverability for P2P and PosterGen. The recoverability axis, which is the axis most aligned with existing coverage measures, is essentially uncorrelated with faithfulness and coherence. This provides direct evidence that coverage-oriented evaluation cannot substitute for distortions and incoherence.

5.6 Why not use LLMs?

We deliberately quantify each diagnostic axis with rule-based functions and not an LLM judge. The motivation behind CoSeCh is to expose failures that are missed by an LLM-based evaluation setup. We need the diagnostic probes to be transparent to evaluate LLM judge’s unreliability. Rule-based scoring is auditable by definition. Since every rule can be broken down into interpretable decisions, any result can be traced to the corresponding claim

and dependencies. It is also independent of the systems considered for testing, several of which are LLM-based (e.g., multi-agent frameworks in P2P and PosterForest). In this way, we avoid the circularity that creeps in while judging an LLM’s output with an LLM of the same family. Additionally, it is deterministic and easy to reproduce, unlike LLM-based tests, which may become unsuitable or obsolete due to model drift or deprecation. We do use a language model to extract claim-qualifier pairs, to map a selected unit back onto its source paragraph when lexical matching fails, and to answer recoverability questions, followed by human validation. But the scoring that follows is a set of deterministic operations over these extracted structures and selections. The division is clear: learned models are used for perception, and measurement remains transparent.

6 Conclusion

We introduce **CoSeCh**, a suite of four automatic rule-based diagnostics that evaluate the content selected by the generator for poster generation. The four axes (faithfulness, coherence, recoverability, and efficiency) assess the selected content independent of layout and rendering. Applying it to 5 generators over ≈ 100 papers, we find that selection quality is multi-dimensional and generator-specific. Paper2Poster is better at preserving qualifying conditions than human authors (+0.076, significant) and cross-unit dependencies than human authors; PosterForest is the most coherent (0.751), yet drops the qualifiers more than humans (-0.106). In contrast, Any2Poster selections are as coherent as the random choice (+0.039, n.s.). Recoverability, the property that existing coverage measures reward, barely separates the generators. All methods span just 0.10 (0.66–0.76) and even random reaches 0.646 vs. human 0.706. Faithfulness and coherence provide sharp distinguishing factors - on coherence the same methods span 0.57. Additionally, the four axes are near-independent ($|\rho| < 0.15$ almost everywhere, the largest only ≈ 0.4). This, along with the 0.10-vs-0.57 gap, provides the quantified argument that coverage-oriented evaluation misses distortion and incoherence. The formulations for these four axes are rule-based by design and hence traceable to specific claims and dependencies. Consequently, **CoSeCh** offers an auditable complement to LLM- and coverage-based poster evaluation. Its scope is deliberately bounded to the content selection step,

which is implicit and goes unvalidated in most generators. We plan to extend this auditable diagnostic framework to paraphrastic distortions and to layout and rendering as a natural next step.

Limitations

Scope of Rule-based Scoring

CoSeCh’s scoring is intentionally rule-based, and this limits what it can catch. Our faithfulness probe flags a qualifier only when its unit is dropped from the selection. It does not detect a claim that keeps its qualifier unit, which is rephrased, nor does it detect the distortion added during LLM-based claim-qualifier extraction. The coherence axis excludes selections with no explicit outgoing reference rather than scoring them (§5.5), and this exclusion rate varies sharply by generator, so a generator can appear more coherent partly by avoiding cross-references altogether, meaning Table 3’s coherence column should be read alongside each generator’s undefined rate. In addition, all diagnostic probes in this work evaluate content selection, not poster rendering. We trade in recall for subtle distortions to improve precision and interpretability, which is a real ceiling on coverage. Extending the detection of paraphrases, likely with an LLM-based detector validated against the rule-based scores here, is the natural next step.

References

- Jiho Choi, Seojeong Park, Seongjong Song, and Hyun-jung Shim. 2026. Posterforest: Hierarchical multi-agent collaboration for scientific poster generation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 379–401.
- Yifan Gao, Zihang Lin, Chuanbin Liu, Min Zhou, Tiezheng Ge, Bo Zheng, and Hongtao Xie. 2025. Postermaker: Towards high-quality product poster generation with accurate text rendering. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8083–8093.
- Or Honovich, Roei Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansky, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. True: Re-evaluating factual consistency evaluation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3905–3920.
- Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. Layoutlmv3: Pre-training for document ai with unified text and image masking. In

Proceedings of the 30th ACM international conference on multimedia, pages 4083–4091.

Vijay Jaisankar, Sambaran Bandyopadhyay, Kalp Vyas, Varre Chaitanya, and Shwetha Somasundaram. 2024. Postdoc: Generating poster from a long multi-modal document using deep submodular optimization. *arXiv preprint arXiv:2405.20213*.

Yekyung Kim, Yapei Chang, Marzena Karpinska, Aparna Garimella, Varun Manjunatha, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. Fables: Evaluating faithfulness and content selection in book-length summarization. *arXiv preprint arXiv:2404.01261*.

Kalpesh Krishna, Erin Bransom, Bailey Kuehl, Mohit Iyyer, Pradeep Dasigi, Arman Cohan, and Kyle Lo. 2023. Longeval: Guidelines for human evaluation of faithfulness in long-form summarization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1650–1669.

Wei Pang, Kevin Qinghong Lin, Xiangru Jian, Xi He, and Philip Torr. 2026. Paper2poster: Towards multi-modal poster automation from scientific papers. *Advances in Neural Information Processing Systems*, 38.

Kanya Paramita and Masayu Leylia Khodra. 2016. Tailored summary for automatic poster generator. In *2016 International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA)*, pages 1–6. IEEE.

Yu-Ting Qiang, Yan-Wei Fu, Xiao Yu, Yan-Wen Guo, Zhi-Hua Zhou, and Leonid Sigal. 2019. Learning to generate posters of scientific papers by probabilistic graphical models. *Journal of Computer Science and Technology*, 34(1):155–169.

Yuting Qiang, Yanwei Fu, Yanwen Guo, Zhi-Hua Zhou, and Leonid Sigal. 2016. Learning to generate posters of scientific papers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Rohit Saxena, Pasquale Minervini, and Frank Keller. 2025. Postersum: A multimodal benchmark for scientific poster summarization. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 1828–1844.

Alessandro Scirè, Karim Ghonim, and Roberto Navigli. 2024. **FENICE: Factuality evaluation of summarization based on natural language inference and claim extraction**. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14148–14161, Bangkok, Thailand. Association for Computational Linguistics.

Tao Sun, Enhao Pan, Zhengkai Yang, Kaixin Sui, Jiajun Shi, Xianfu Cheng, Tongliang Li, Ge Zhang, Wenhao Huang, Jian Yang, et al. 2026. P2p: Automated paper-to-poster generation and fine-grained

benchmark. In *International Conference on Learning Representations*, volume 2026, pages 18756–18798.

Amogh Vinaykumar, Aiden Li, Suozhi Huang, and Shilong Liu. 2026. Any2poster: Any-source poster generation across modalities and domains. *arXiv preprint arXiv:2606.02915*.

Hao Wang, Shohei Tanaka, and Yoshitaka Ushiku. 2024. Scipostlayout: A dataset for layout analysis and layout generation of scientific posters. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8136–8141.

Sheng Xu and Xiaojun Wan. 2021a. **Neural content extraction for poster generation of scientific papers**. *Preprint*, arXiv:2112.08550.

Sheng Xu and Xiaojun Wan. 2021b. Neural content extraction for poster generation of scientific papers. *arXiv preprint arXiv:2112.08550*.

Zhilin Zhang, Xiang Zhang, Jiaqi Wei, Yiwei Xu, and Chenyu You. 2025. Postergen: Aesthetic-aware multi-modal paper-to-poster generation via multi-agent llms. *arXiv preprint arXiv:2508.17188*.

A Implementation Details

A.1 Content Unit Extraction

We build one canonical content-unit tree per paper — the global superset U , using a docling-based structural decomposer that parses sections/subsections/paragraphs/sentences, extracts figures (paired with captions), tables, and equations as separate units, and tracks heading depth relatively for robustness across papers; this single U is shared across all five generators and the author/random baselines. We break each generator’s pipeline (P2P, Paper2Poster, PosterForest, PosterGen, Any2Poster) at content selection to extract its selected set S , along with the human selection S_h and its size $|S_h|$; the random baseline samples 50 units without replacement over 100 trials, keeping the top- k by inclusion frequency with $k = |S_h|$ for that paper.

A.2 Faithfulness

Claim–qualifier pairs are extracted from the source paper with a single LLM pass, restricted to five qualifier types: scope ("only on English-language papers"), conditional ("under condition Y"), statistical (" $p < 0.05$ "), magnitude ("up to $4\times$ faster"), and assumption ("assuming the set is finite"). Rather than checking survival against the generator’s own selected-content text, which is often a paraphrase or summary of the source, each selected unit is first mapped back onto the paragraph in the source

paper it was drawn from — recovered by matching the unit against the paper’s own paragraph structure — and this original-paper paragraph is substituted in place of the generator’s text; a unit is left in its own (summarized) form only when no confident match to the source is found.

For each pair, we check whether the claim and its qualifier survive in this mapped selection independently. A claim is counted as present if the best-matching sentence in the selection contains its tokens above a fixed containment threshold, and its qualifier is counted as surviving if a matching span appears within a three-sentence window of the claim in the selection. This containment check is supplemented by a fallback list of qualifier-marker phrases matched to these five types, so that a qualifier rephrased using a standard marker is still credited as retained even when it is not a verbatim match. A pair is scored as dropped only when the claim is selected, but no qualifier span, verbatim, or marker-matched, is found in its window.

A.3 Coherence

For each paper, a shared directed dependency graph is built once over the paper’s own hierarchical structure. Reference edges are detected by regex over figure/table/equation/section mentions — including plural and multi-item forms ("Figs. 2, 3"), numeric ranges ("Fig. 2–4"), roman-numeral table numbering, and bare parenthetical equation references ("...follows from (1)") lacking an explicit "Eq." prefix — with figure/table nodes attached from one generator’s own asset list as a shared caption source, so every schema’s captions are resolved against the same canonical targets rather than each schema’s own.

A section/subsection reference is treated as satisfied if either the heading itself or any of its descendant paragraphs was selected, since selection happens at paragraph granularity and the heading node is otherwise almost never selected on its own. Units that fail to map onto the canonical content tree U are still scanned for their own outgoing dependencies rather than being silently dropped, so their references still count toward the denominator (Eq. (4)). This prevents a schema from appearing more coherent than it is simply because its unmapped units vanish from the calculation.

A.4 Recoverability

Recoverability is operationalized directly following PaperQuiz (Pang et al., 2026), the existing

content-recall protocol, rather than a new question set. For each paper, PaperQuiz supplies two batches of 50 four-choice questions generated from the source PDF text: a detail batch (aspects A–M: title/authorship, motivation, objectives, datasets, methodology, hyperparameters, metrics, quantitative results, qualitative findings, comparative/ablation results, conclusions, limitations, definitions) and an understanding batch (aspects A–J: domain, problem, goal/hypothesis, key contributions, methodology overview, headline findings, qualitative insight, implications, limitations, main conclusions). Each question already carries its generating aspect as metadata.

A reader model answers each question in $Q(U)$ using only the generator’s selected content S : the selected text units are concatenated in poster reading order and every selected figure/table image is attached alongside its section label and caption, so the reader sees exactly the same evidence a poster viewer would. The reader returns a single letter (A–D) per question, or "NA" if S offers no basis for an answer; a question is scored correct only on an exact letter match against the gold answer, with "NA" always counted as incorrect. $Rec(S)$ is this accuracy averaged over $Q(U)$.

A.5 Efficiency

Budget efficiency is computed from the same selected-unit set as every other metric: each pair of selected units is compared with SBERT if both are text, or with CLIP if either is a visual unit (so a chart and a bullet point restating its number share CLIP’s joint text/image embedding space rather than being incomparable), and a pair is flagged as a near-duplicate when its cosine similarity is at or above a calibrated threshold τ . Independently of that similarity check, any two units whose text contains the same numeric token — a percentage, a count, a stat — are flagged as a same-claim restatement. These flagged pairs form a graph over the $|S|$ selected units, which is partitioned into k connected components (clusters) via union-find; efficiency is reported as $k/|S|$ — a generator that deliberately restates a result across modalities for emphasis simply spends part of its unit budget on that repetition, and efficiency falls exactly as much for a legitimate cross-modal callback as for an accidental duplicate, without a second pass trying to guess which clusters were "intentional."

Similarity Threshold (τ) itself was calibrated empirically rather than asserted: we swept it from

0.50 to 0.90 in steps of 0.05, recomputing each generators's mean efficiency rate at every step from the same precomputed embeddings (only the flagging/clustering step depends on τ , so embeddings were encoded once, not once per threshold), then averaged the five schemas' rates together at each of the nine threshold values to get one aggregate curve. Quartiles were then taken over those nine averaged rate values themselves not over the threshold axis, which would be tautological and the thresholds whose rate fell in the bottom quartile (the flattest, most conservative end of the curve, where the rate is least sensitive to a further increase in τ) were 0.80, 0.85, and 0.90. We selected 0.8, which was the first τ value falling into bottom quartile.

Method	Faithfulness	Coherence	Recoverability	Efficiency
P2P	0.671 [0.622, 0.715]	0.356 [0.274, 0.440]	0.719 [0.695, 0.740]	0.811 [0.780, 0.840]
Paper2Poster	0.736 [0.694, 0.776]	0.545 [0.455, 0.636]	0.758 [0.744, 0.772]	0.769 [0.734, 0.802]
PosterForest	0.571 [0.513, 0.628]	0.751 [0.683, 0.814]	0.728 [0.711, 0.743]	0.799 [0.767, 0.829]
PosterGen	0.644 [0.589, 0.696]	0.369 [0.286, 0.450]	0.727 [0.710, 0.742]	0.873 [0.850, 0.896]
Any2Poster	0.671 [0.627, 0.713]	0.181 [0.122, 0.247]	0.662 [0.644, 0.681]	0.775 [0.740, 0.810]
<i>Author</i>	0.668 [0.615, 0.721]	0.405 [0.324, 0.487]	0.706 [0.687, 0.724]	0.842 [0.815, 0.869]
<i>Random</i>	0.543 [0.472, 0.611]	0.141 [0.094, 0.192]	0.646 [0.607, 0.681]	0.731 [0.702, 0.759]

Table 3: Per-method mean with 95% bootstrap CIs over papers. Author and Random are references. Per-cell n by axis: faithfulness 88–98, coherence 64–86, recoverability 95–100, efficiency 95–100.

Comparison	Faithfulness	Coherence	Recoverability	Efficiency
P2P	+0.140 [+0.063, +0.220]	+0.197 [+0.100, +0.291]	+0.073 [+0.032, +0.115]	+0.078 [+0.034, +0.121]
Paper2Poster	+0.202 [+0.121, +0.284]	+0.387 [+0.270, +0.500]	+0.113 [+0.080, +0.150]	+0.038 [-0.005, +0.080] ^{ns}
PosterForest	+0.034 [-0.055, +0.121] ^{ns}	+0.590 [+0.495, +0.680]	+0.083 [+0.049, +0.120]	+0.071 [+0.028, +0.112]
PosterGen	+0.090 [-0.001, +0.182] ^{ns}	+0.212 [+0.108, +0.317]	+0.078 [+0.045, +0.112]	+0.141 [+0.108, +0.174]
Any2Poster	+0.123 [+0.038, +0.204]	+0.039 [-0.036, +0.118] ^{ns}	+0.017 [-0.021, +0.056] ^{ns}	+0.045 [-0.003, +0.094] ^{ns}

Table 4: Paired bootstrap differences vs. Random (validity floor). ^{ns} = not significant. Paired n by axis: faithfulness 81–89, coherence 59–78, recoverability 95–100, efficiency 95–100.

Comparison	Faithfulness	Coherence	Recoverability	Efficiency
P2P	+0.020 [-0.033, +0.075] ^{ns}	-0.091 [-0.210, +0.028] ^{ns}	+0.011 [-0.016, +0.034] ^{ns}	-0.034 [-0.066, +0.001] ^{ns}
Paper2Poster	+0.076 [+0.021, +0.131]	+0.134 [+0.025, +0.242]	+0.052 [+0.035, +0.070]	-0.073 [-0.111, -0.035]
PosterForest	-0.106 [-0.187, -0.028]	+0.316 [+0.205, +0.433]	+0.020 [+0.002, +0.040]	-0.043 [-0.079, -0.008]
PosterGen	-0.023 [-0.082, +0.038] ^{ns}	-0.019 [-0.135, +0.099] ^{ns}	+0.022 [+0.004, +0.041]	+0.035 [-0.001, +0.072] ^{ns}
Any2Poster	+0.001 [-0.051, +0.054] ^{ns}	-0.275 [-0.376, -0.175]	-0.044 [-0.063, -0.024]	-0.066 [-0.104, -0.030]

Table 5: Paired bootstrap differences vs. the human *Author*. Positive = generator exceeds the human; ^{ns} = not significant (95% CI includes 0). Paired n by axis: faithfulness 88–96, coherence 57–73, recoverability 95–100, efficiency 95–100.

Axis	Component	Judgment	Accuracy
Mapping	unit → superset paragraph	correct / incorrect	80.9%
Faithfulness	claim–qualifier extraction	valid / invalid	72.3%
Coherence	reference edge genuineness	genuine / spurious	75.6%
Recoverability	reader-model answer correctness	correct / incorrect	83.2%
Efficiency	pairwise redundancy	redundant / distinct	79.8%

Table 6: Manual validation of each CoSeCh component against a single expert annotator (25 papers). **Judgment** is the binary decision the annotator made for each item. **Accuracy** is agreement between the automatic component and the annotator’s label.