VERIFIABLE PAPER-TO-VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern paper-to-video systems can now produce polished and compelling presentations from scientific papers. Yet visual quality does not guarantee faithfulness to the source. Existing evaluations focus largely on presentation quality and informativeness, with much less attention to whether the claims in a generated video are actually supported by the paper. We introduce PRESENTFAITH, a benchmark that compares claims recovered from the generated video with a fine-grained Evidence Index of the source paper. Our evaluation shows that current systems can perform well on standard presentation metrics while still producing substantial unsupported and contradictory content. These findings motivate a change in how the presentation itself is generated. We propose EVI-IR, an evidence-linked intermediate representation that stores each generated claim in a structured form together with its claim type, evidence pointer, and relevant quantities. Keeping this information explicit makes each claim traceable to the paper and its supporting evidence directly verifiable. EVI-IR then checks these links whenever possible and repairs unsupported or inconsistent claims before rendering. Experiments on PRESENTFAITH demonstrate that EVI-IR reduces the unfaithful claim rate from 17.6% to 12.5% relative to the strongest baseline, while increasing supported claims per minute from 11.7 to 13.6 and preserving existing presentation-quality metrics.

1 INTRODUCTION

Researchers often struggle to turn a technical paper into a convincing conference presentation. Increasingly, machines can help. Recent agent-based systems can generate complete presentation videos directly from scientific papers. To automate the process, a paper-to-video system has to decide both how to present the work and which scientific claims to present. Recent systems have made remarkable progress on the first, generating complete talks with slides, narration, speech, and talking heads (Zhu et al., 2025; Shi et al., 2025), building on earlier work on automatic slide and poster generation (Fu et al., 2022; Zheng et al., 2025; Ge et al., 2025; Pang et al., 2026). Existing metrics follow the same emphasis, measuring similarity to author-made presentations, how much the audience retains, and which presentation the audience prefers (Zhu et al., 2025; Shi et al., 2025). All of these describe how well a presentation communicates. Whether the claims it communicates are supported by the paper is left unmeasured.

This missing dimension matters in practice. Figure 1 shows three examples from the official implementations of Zhu et al. (2025) and Shi et al. (2025): a numeric value that is not stated in the paper, a comparison whose direction is reversed, and a result attributed to the wrong method. The videos remain fluent and well presented, and the errors are not reflected by the standard presentation metrics. A viewer can therefore come away with a fluent, persuasive account of the paper built on claims the paper never made. Measuring how often that happens is where we start. We introduce PRESENTFAITH, a benchmark that compares claims recovered from the delivered video with a fine-grained Evidence Index built from the source paper. The index represents evidence at the level of paragraphs, equations, figures, and individual table cells, allowing numeric and comparative claims to be checked against the paper in a structured way. Because PRESENTFAITH evaluates the delivered video, it can be applied consistently to existing systems without requiring access to their internal generation process. Our evaluation shows that current systems still produce substantial unsupported and contradictory content even when they perform well on standard presentation metrics. Faithfulness and presentation quality thus behave as separate axes.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

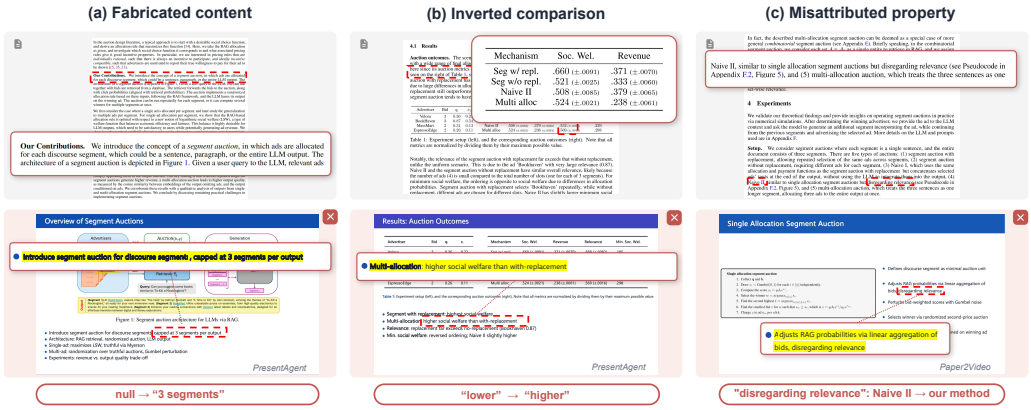


Figure 1: Three faithfulness failures of existing paper-to-video systems. Each column pairs the source paper's own evidence (top) with the claim a delivered video makes about it (bottom, marked ×), with the system named underneath. The four presentation metrics used by prior work stay at their pipeline's average on all three videos; the unfaithful-claim rate (UCR) separates them.

These results motivate a different way to build the pipeline: a generated claim should stay connected to the evidence that supports it, and that connection should be checkable by machine. The closest prior system, OmniPresent(Ma et al., 2026), plans from an explicit evidence stream and verifies plans before rendering, yet its pointers stop at sections and figures and its checks are VLM-judged.

We therefore propose EVI-IR, an evidence-linked intermediate representation for paper-to-video generation. Each assertive unit is emitted in structured form together with its claim type, one or more evidence pointers, and the quantities it mentions, and every claim carries its own record of support. EVI-IR first applies deterministic checks to resolve evidence pointers, validate evidence types, and verify numeric consistency whenever possible. Claims that remain unsupported or inconsistent are revised through a bounded repair process, and only then does the representation pass to rendering. the evidence link lives in the representation from the moment it is written, and the presentation is produced from a representation that already knows its sources.

We evaluate EVI-IR on 40 papers. Under the same renderer used for the baselines, EVI-IR reduces the unfaithful-claim rate from 17.6% to 12.5% and numeric contradictions from 9.7% to 2.6% relative to PAPERTALKER, while increasing supported claims per minute from 11.7 to 13.6 and preserving existing presentation-quality metrics. Together, these results show that paper-to-video generation can remain effective as a presentation task while making its scientific claims substantially more traceable and verifiable.

Contributions

1. To our knowledge, we are the first to formulate paper-to-video generation as a problem of *verifiable faithfulness*: generated claims should not only be faithful to the source paper, but also remain traceable to evidence that can be checked after the video is produced.
2. We introduce PRESENTFAITH, a benchmark that scores claims recovered from delivered videos against a fine-grained Evidence Index of the source paper, with structured numeric checks and paired anti-gaming metrics.
3. We introduce EVI-IR, which emits claims with explicit evidence links, verifies them symbolically before rendering, and repairs what fails; on PRESENTFAITH it improves faithfulness while preserving presentation quality.

2 RELATED WORK

Paper-to-video and presentation generation Automatic presentation generation has moved from slide decks and posters (Fu et al., 2022; Zheng et al., 2025; Ge et al., 2025; Pang et al., 2026; Tan et al., 2027; Xu et al., 2025) to full videos. Direct text-to-video models (Yang et al., 2025) render

a prompt into footage but produce almost no scientific narration. Multi-agent pipelines instead assemble slide synthesis, narration, TTS, and talking-head rendering from off-the-shelf modules (Radford et al., 2023; Chen et al., 2025; Cui et al., 2025), as in PAPERTALKER (Zhu et al., 2025), PRESENTAGENT (Shi et al., 2025), PREACHER (Liu et al., 2025b), EVOPRESENT (Liu et al., 2025a), and OMNIPRESENT (Ma et al., 2026). Two gaps recur across these systems. Verification rests on model judgement, through aesthetic scorers, VLM judges, or reflection agents, which cannot catch a number bound to the wrong table row or a reversed comparison. Provenance stops at sections and figures and never reaches the table cell a claim depends on, leaving such errors with nothing to check against. EVI-IR closes both: every claim is pointed at the cell that supports it, and the pointer is checked with symbolic rules before rendering.

LLM faithfulness evaluation and benchmarks Long-form factuality is commonly evaluated by decomposing a generation into atomic claims and scoring each against a source (Min et al., 2023; Wei et al., 2024; Song et al., 2024; Tang et al., 2024), with roots in summarisation faithfulness (Maynez et al., 2020; Kryściński et al., 2020; Wang et al., 2020) and table-grounded verification (Chen et al., 2019). We adopt the decomposition and change three of its assumptions. The object under test is a video: claims have to be recovered through ASR and OCR and perception errors separated from genuine hallucination. The reference is a single structured document, which lets numeric and comparative claims be decided by unit-aware arithmetic. And the dominant error is relational, a real number attached to the wrong row or a comparison reversed, which existence-style support checks grade as correct. Perception benchmarks (Wang et al., 2024; Guan et al., 2024; Zhang et al., 2024) ask what a model sees in a given video; PRESENTFAITH asks what a generated video asserts about a document. The closest prior work, OmniPreBench (Ma et al., 2026), scores generated suites with VLM-judged FactScores, whereas PRESENTFAITH decides every claim against a structured Evidence Index arithmetically and pairs each metric with an anti-gaming counter-metric.

Citation generation and attribution A parallel line asks models to cite their answers (Menick et al., 2022; Nakano et al., 2021), to be evaluated for citation quality (Gao et al., 2023b; Yue et al., 2023), and to attribute claims to evidence, either post hoc (Gao et al., 2023a; Dhuliawala et al., 2024) or by construction before generating prose (Slobodkin et al., 2024). We take the by-construction stance and push it along three axes. Pointers resolve to individual table cells with row and column paths, and the dominant error class becomes visible. A table-granularity index hides it. Attribution is placed before an irreversible projection, the only point at which it can be decided at all. Given a symbolic index, most numeric and comparative bindings are decided arithmetically, with schema-constrained decoding (Willard & Louf, 2023; Geng et al., 2023) keeping the representation parseable. Bounded monotone repair on verification failure keeps revision evidence-anchored.

3 PRESENTFAITH: A FAITHFULNESS BENCHMARK

3.1 TASK DEFINITION AND OVERVIEW OF FAITHFULNESS

Section 2 left two gaps open: verification rests on model judgement, and provenance stops above the table cell a claim depends on. Neither says how much unfaithfulness reaches the viewer, which is also what the presentation metrics leave unmeasured (Zhu et al., 2025). PRESENTFAITH measures it. Given a paper P (LaTeX or PDF) and the video v of slides and narration that a paper-to-video system produces, the benchmark records, for every claim v makes, whether P supports it.

To score a video, we convert it into *informative atomic claims* $A(v)$, the smallest independently checkable units of assertion, read from the narration and slides. A benchmark-side verdict function

$$y^{\exists} : A(v) \to \{\text{SUPPORTED}, \text{ CONTRADICTED}, \text{ UNSUPPORTED}\} \tag{1}$$

labels each claim SUPPORTED if some evidence unit in $\mathcal{U}(P)$ supports it, CONTRADICTED if some unit contradicts it, and UNSUPPORTED otherwise. Here $\mathcal{U}(P)$ is a typed *Evidence Index* that organizes P into *evidence units* (paragraphs, equations, figure captions, table cells, etc.; Section 3.2). Figure 2 gives an overview of the resulting benchmark.

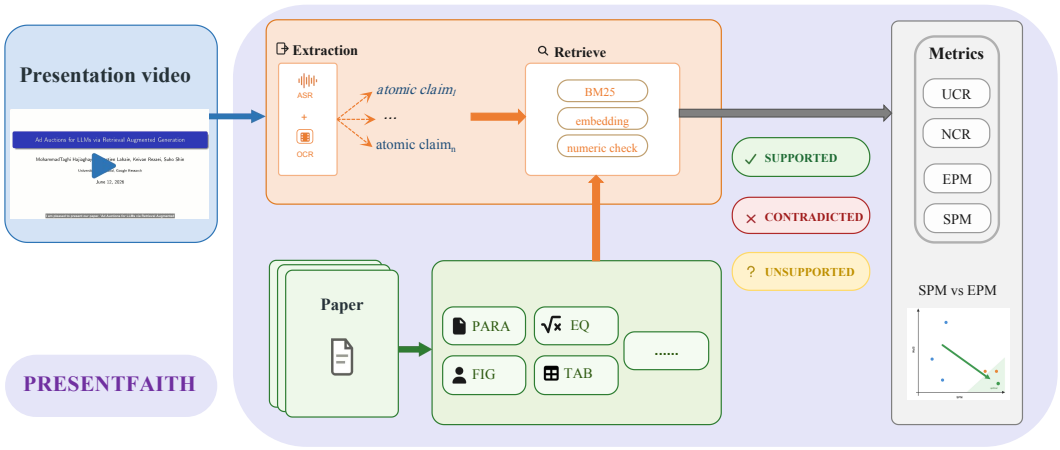


Figure 2: The **PRESENTFAITH** benchmark. Atomic claims are recovered from the delivered video by word-level ASR of the narration and OCR of slide keyframes, then checked against the typed Evidence Index $U(P)$ built from the paper's paragraphs, equations, figures, table cells, and numeric mentions. A retrieval layer proposes candidate units, and a layered verifier labels each claim SUPPORTED, CONTRADICTED, or UNSUPPORTED; the verdicts aggregate into UCR, NCR, EPM, SPM.

3.2 DATASET

Every verdict is read against the index, so we describe it before the pipeline that consults it. PRESENTFAITH draws its papers from Paper2Video (Zhu et al., 2025), the largest public collection of paper–presentation pairs: 101 papers, each accompanied by an author-recorded talk. Author narration provides a human reference for what a faithful video should convey, and the shared source keeps our results comparable to published systems. From these we keep 40 papers, chosen for quantitative density. Every retained paper has at least one table with ≥ 6 numeric cells and one captioned figure, and every body runs 6–12 pages, short enough that truncation cannot confound the evaluation.

For each paper, we build a typed evidence index $\mathcal{U}(P)$ from its LaTeX source, writing $\mathcal{U}$ when the paper is clear from context. Deterministic parsing splits the paper into five kinds of *evidence unit*: paragraphs, equations, figure captions, table cells, and numeric mentions. Annotators then attach to every unit the fields the verdict function Equation (1) reads: an immutable hash identifier, the unit type, its path through the paper's hierarchy, and its verbatim content. Numeric units carry two fields. The first is a normalized (value, unit, precision) triple; the second, the quantity the number denotes, read off from its row and column context, and it is the one that makes comparisons checkable. A numeric claim then be tested against the quantity it names, and a comparative claim such as "a 2.5-point gain" against the two cells it subtracts; a near miss that lands in the right table but the wrong column still registers as an error. The schema, protocol, and statistics are provided in Section A. Because the parsing is deterministic, its failures can be audited directly Section 5.2.

3.3 EVALUATION AND METRICS

Evaluation The index makes faithfulness checkable in principle, but it does not say what a video asserts, so a delivered video still has to be turned into explicit assertions first. A fixed pipeline does the conversion: word-level ASR transcribes the narration, OCR reads the slide text, and both streams are split into atomic claims, with non-assertive discourse (agendas, transitions, hedged speculation) exempted from scoring. Each remaining claim is matched to $\mathcal{U}(P)$ by claim type. Numeric and comparative claims take the number path: the pipeline looks up the cells in the index and recomputes the arithmetic, so a claim such as "a 2.5-point gain over the baseline" is decided by locating the two cells it subtracts. All other claims take the retrieval path: candidates are retrieved by BM25 and embedding similarity, and the claim is decided by entailment against them. Every claim thus receives one of three verdicts: SUPPORTED, CONTRADICTED, or UNSUPPORTED. Extraction, retrieval, and verdict protocols are in Section A.2. Scoring the delivered video and nothing else is what lets PRESENTFAITH be applied to closed systems; it also fixes what the benchmark can ask, since once a video exists only the existence-level question survives. Section 4 takes up the cost of that restriction.

Metrics We evaluate faithfulness at two levels. At the **Existence-Level**, a claim is faithful if some unit in $\mathcal{U}(P)$ supports it. Measurable from the video, it forms the main benchmark evaluation.

Let $A^\star(v)$ denote the scored claims in video v, $A^\star\mathrm{num}(v)$ its numeric and comparative claims, and $E(v) = a \in A^\star(v) : y^\exists(a) \in \{\textsc{contradicted}, \textsc{unsupported}\}$ its unfaithful claims. For a video of duration $d(v)$,

$$\begin{aligned} \mathrm{UCR} &= \frac{|E(v)|}{|A^\star(v)|}, \qquad \mathrm{NCR} = \frac{\left|\{\, a \in A^\star_{\mathrm{num}}(v) : y^\exists(a) = \textsc{contradicted} \,\}\right|}{|A^\star_{\mathrm{num}}(v)|}, \\ \mathrm{EPM} &= \frac{|E(v)|}{d(v)}, \qquad \mathrm{SPM} = \frac{\left|\{\, a \in A^\star(v) : y^\exists(a) = \textsc{supported} \,\}\right|}{d(v)}. \end{aligned} \tag{2}$$

UCR (unfaithful claim rate) is the fraction of scored claims that are contradicted or that no unit supports. NCR (numeric contradiction rate) restricts the same count to numeric and comparative claims, where values and comparison directions can be checked with the least ambiguity, and counts outright contradictions only. EPM and SPM move from rates to densities, counting unfaithful and supported claims per rendered minute, so that error density is never read without informational yield. Scores are macro-averaged over papers, with confidence intervals bootstrapped and clustered by paper. None of these ratios means much alone: EPM falls when a system says less, UCR falls when it says only what is safe. We therefore read results on the $(\mathrm{SPM}, \mathrm{EPM})$ plane. The desired region has lower EPM and higher SPM, more of what the paper supports at lower error density, and no single-coordinate strategy reaches it. Section 5 develops this reading on real systems.

At the **Attribution-Level**, for systems that emit provenance, we report *pointer accuracy* (the unit pointed at lies in the set annotators accept), *cell-exact accuracy* (right table, row, and column), and the *abstention rate*. These quantities describe the intermediate representation.

Result Applied to existing P2V systems, PRESENTFAITH shows that fluency and faithfulness come apart: even the strongest baseline leaves 17.6% of scored claims unsupported or contradicted and misstates 9.7% of its numeric ones, while the naive one-shot pipeline reaches 30.5% on UCR and 21.8% on NCR Section 5. All of these are measured after the fact, and that is what limits them: by the time an error is measurable it is already in the artifact, and correcting it means regenerating the slide. Section 4 moves the same decision upstream, to a generator that emits each claim together with the unit meant to support it.

4 **EVI-IR**: EVIDENCE-LINKED INTERMEDIATE REPRESENTATION

4.1 OVERVIEW

Existing paper-to-video pipelines generate content first and check its support afterwards. Section 3 shows what the habit costs: the strongest baseline misstates 9.7% of its numeric claims, and by the time a check could catch one of them the video is already on screen, while the provenance the check would need was never recorded. EVI-IR inverts the order. A claim is emitted together with a pointer to the unit of evidence that supports it, and the pointer is stored with the claim from the moment it is written Figure 3. The representation stays deliberately small. An atom records its surface text, a claim type, a pointer π into the Evidence Index $U(P)$ of Section 3.2, and whatever quantities it mentions. Reusing $U(P)$ as the target space keeps verification simple: a pointer resolves to a typed unit, and a typed unit admits a fixed set of checks. Most bindings are decided symbolically, and the rest by a single entailment decision that reads exactly one unit. Failures go back to the generator, which rewrites the atoms that failed and leaves the rest alone.

Because each atom also carries the render directives its slide needs, nothing downstream has to infer them. EVI-IR is written once and read twice: the verifier consumes it while provenance is still intact Sections 4.2 and 4.3, and the renderer then lays out every slide and produces the video Section C.1.

4.2 GENERATION: EVI-IR

EVI-IR is the single artifact the pipeline writes and never re-derives: a sequence of slide specifications, each holding the atoms that the slide asserts and the directives that place them on screen.

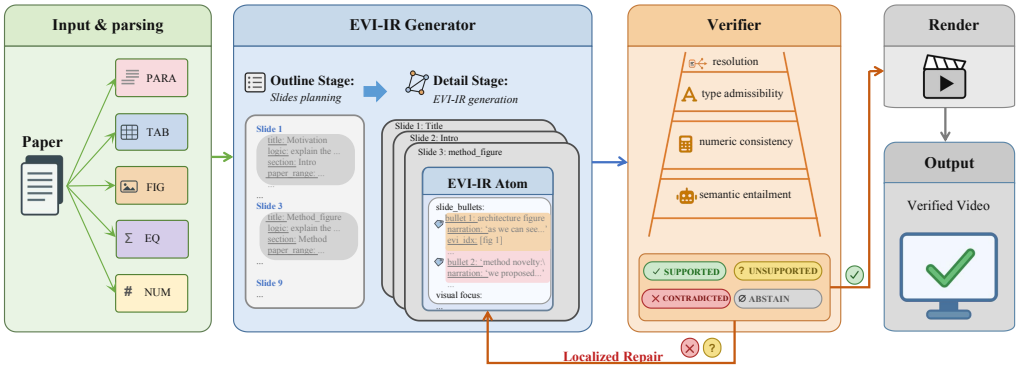


Figure 3: The **EVI-IR** framework. The paper is parsed into a typed Evidence Index $U(P)$ of paragraphs, tables, figures, equations, and numeric mentions. A two-stage generator plans an outline and then emits each slide's atoms in structured form, together with evidence pointer, parsed quantities, and render directives. The verifier applies the layered checks V_1–V_4 in order and labels every atom SUPPORTED, CONTRADICTED, UNSUPPORTED, or ABSTAIN; failing atoms return to the generator through localized repair. Passed atoms reach the deterministic renderer to produce the video.

Two-stage generation The generator works from the evidence index of Section 3.2 and proceeds in two stages, as in Figure 3. In the outline stage it fixes the number of slides and gives each one a topic and a scope, a scope being a subset of $\mathcal{U}$. In the detail stage it emits the atoms of each slide inside that slide's scope, drawing every pointer from the handful of candidates the scope makes visible. Restricting the choice this way is what makes errors repairable: a wrong pointer is a wrong choice among a few known units. The model emits atoms in structured form directly. Parsing prose after the fact would put back the surface-to-provenance inference we are trying to remove, and a span that will not conform is emitted with an empty pointer, which abstention later accounts for.

Anatomy of an atom An atom is one assertive unit, a tuple $\langle$surface text, type, $\pi \subseteq \mathrm{ids}(\mathcal{U})$ with $|\pi| \leq 2$, $\{(\text{value}, \text{unit}, \text{precision})\}\rangle$; compound sentences are split as they are emitted. The surface text goes to the viewer, the pointer π names the unit of $\mathcal{U}(P)$ meant to support it, and quantities are stored as parsed triples of value, unit, and source precision. Storing quantities this way makes the arithmetic check of Section 4.3 decidable.

The type routes verification, deciding which targets are admissible and which checks apply. Four assertive types suffice. *Quantitative* atoms carry one or two cells, quantities, or equations, and a comparison is simply a quantitative atom with two pointers. *Definitional* atoms state a definition, *attributive* atoms carry a citation, and *descriptive* atoms carry a claim or a figure. The last of these is the residual class, so the four between them cover everything the narration can assert.

Atoms the model cannot attribute ($\pi = \varnothing$), and spans that assert nothing verifiable in the first place, are labeled ABSTAIN. The label is kept distinct from UNSUPPORTED, because without the distinction a confidently wrong pointer and an honest inability to attribute would score identically Sections C.2 and 5.4. Abstentions therefore count in the coverage denominator and never in the accuracy numerator, so that coverage cannot be traded for accuracy.

4.3 Verification and Localized Repair

A pointer is a claim in its own right, and it can be wrong in three ways: it may fail to resolve in $\mathcal{U}$; it may resolve to a unit of the wrong kind; or it may resolve perfectly well to a unit whose content does not support the sentence attached to it. The verifier sits where all three are still diagnosable, between generation and rendering, and returns a binding-level verdict $y^{\rightarrow}$ from {SUPPORTED, CONTRADICTED, UNSUPPORTED, ABSTAIN}. The vocabulary repeats the one introduced in Equation (2), applied to a narrower object: the existence-level score $y^{\exists}$ records whether *some* unit of $\mathcal{U}(P)$ supports the claim, whereas the binding-level verdict records whether the unit *pointed at* does. The three failure modes above all presuppose that a pointer was produced; ABSTAIN, the one label the binding level adds, is reserved for the remaining case, atoms with $\pi = \varnothing$, for which no binding exists to check. Such verdicts count in the coverage denominator and never in the accuracy numer-

ator. Keeping the pointer through this stage is what makes the second question decidable at all. On a rendered video it collapses into the first, which is the limit Section 3.3 records.

Layered verification Every atom passes four sequential checks, V_1–V_4, ordered by increasing cost. An atom advances only while every earlier check passes, and any failure returns it to repair at once. V_1 (resolution) tests that π names a unit that exists in $U(P)$. V_2 (type admissibility) tests the resolved unit is of a kind the atom's type permits; a quantitative atom, for instance, must point at a unit containing numbers. V_3 (numeric consistency) applies to quantitative atoms only and recomputes the relation the sentence asserts, over differences, ratios, relative gains, and orderings, under unit commensurability and source precision. It targets the dominant numeric failure, values present yet jointly inconsistent, exactly the class an existence-style check must mark SUPPORTED. V_4 (semantic entailment) is a prompted LLM decision with the pointed-at unit as premise and the atom as hypothesis, returning SUPPORTED, CONTRADICTED, or UNSUPPORTED. Nothing but the pointed-at unit is consulted, which keeps the verdict binding-level and aligned with Equation (2).

Localized repair Every atom carries its own verdict, so repair stays local. Only failing atoms are regenerated, within their own slide, and atoms already marked SUPPORTED are left untouched. The loop runs at most r rounds and rejects any round that increases the number of non SUPPORTED atoms, which makes it monotone and guarantees termination. Abstention carries a cost of its own. An ABSTAIN never counts as correct, and we report attribution accuracy, coverage, and end-task quality separately, so that one cannot be traded for another. Section 5 measures what each of these choices contributes, under a renderer shared with the baselines, so that nothing is credited twice.

5 EXPERIMENTS

5.1 SETUP

Baselines Our claim is that explicit evidence linkage improves faithfulness without sacrificing presentation quality. To test it, we compare EVI-IR with five reference systems spanning the design space. **(i) NAIVE LLM** emits the narration script and slide text in one pass. **(ii) PRESENTAGENT** segments the paper, renders slide-style frames, and synthesizes narration with LLMs and TTS (Shi et al., 2025). **(iii) PAPERTALKER** adds tree-search layout, cursor grounding, and parallel talking-head generation, and is the strongest baseline (Zhu et al., 2025). **(iv) MPRESENTER** is a multilingual agentic system with audience-oriented critique and iterative refinement (Han et al., 2026). **(v) HUMANMADE** is the author-recorded presentation, which routinely asserts content absent from the paper. For comparability, every system, including ours, is scored only from its delivered video through one shared extraction-and-verdict pipeline.

Data and Metrics All experiments run on PRESENTFAITH: 40 papers with evidence index $\mathcal{U}(P)$. Faithfulness is measured by the Existence-Level quantities of Equation (2): unfaithful-claim rate (UCR), numeric contradiction rate (NCR), errors per minute (EPM), and the supported claims per minute (SPM). No single metric is meaningful alone, so we read UCR, NCR, and SPM jointly on the $(\mathrm{SPM}, \mathrm{EPM})$ plane. For end-task quality we adopt PAPERTALKER metrics (Meta Similarity, PresentArena, PresentQuiz, IP Memory), so our numbers are directly comparable to prior work.

Implementation The multi-agent pipelines run under their official settings; all systems share one backbone (GPT-4.1, with Gemini-2.5-Flash as the VideoLLM judge) and are scored from the delivered video, so emitting an intermediate representation confers no advantage. Stochastic systems use three seeds; scores are macro-averaged over 40 papers with bootstrapped CIs. We organize the experiments into four groups: **(A)** benchmark validation, **(B)** faithfulness against baselines, **(C)** attribution accuracy, and **(D)** cost, transfer, and ablation.

5.2 BENCHMARK VALIDATION

The benchmark rests on two artifacts: the evidence index $\mathcal{U}(P)$, which is the ground truth, and the verifier, which is the judge and we audit both of them.

Index validity We first validate the evidence index $\mathcal{U}(P)$ itself. Two annotators check 300 type-stratified units: 95.7% are parsed correctly and 2.3% of spans are mislabeled. The residual errors

System	PRESENTFAITH UCR↓	NCR↓	EPM↓	SPM↑	End-task quality Sim.↑	Arena↑	Quiz↑	IP↑
HUMANMADE *(anchor)*	4.6	0.9	0.7	14.2	1.00	0.50	0.91	—
NAIVE LLM	$30.5_{1.4}$	$21.8_{1.6}$	$4.4_{0.3}$	$8.8_{0.6}$	0.47	0.01	0.59	0.31
PRESENTAGENT	$21.4_{0.9}$	$12.8_{1.1}$	$3.2_{0.2}$	$10.5_{0.4}$	0.51	0.02	0.65	0.13
MPRESENTER	$19.3_{0.9}$	$11.2_{1.0}$	$2.9_{0.2}$	$11.9_{0.4}$	0.56	**0.18**	**0.97**	0.53
PAPERTALKER	$17.6_{0.8}$	$9.7_{0.9}$	$2.7_{0.2}$	$11.7_{0.3}$	**0.61**	0.17	0.95	0.50
EVI-IR (ours)	$\mathbf{12.5}_{0.7}$	$\mathbf{2.6}_{0.8}$	$\mathbf{1.3}_{0.8}$	$\mathbf{13.6}_{0.8}$	0.60	**0.18**	0.96	**0.55**

Table 1: Main results on **PRESENTFAITH**. Left: Existence-Level faithfulness of Equation 2, lower is better for UCR, NCR, EPM and higher for SPM. Right: the four PAPERTALKER end-task metrics under official implementations. All systems are scored from their delivered video in the shared-renderer configuration, so EVI-IR receives no credit for emitting provenance. HUMANMADE is a descriptive anchor and stays ahead of all generated systems on every faithfulness column.

are isolated rather than systematic—a mislabeled span corrupts a single unit, not the index as a whole—so $\mathcal{U}(P)$ is accurate enough to serve as the ground truth against which claims are judged.

Verifier validity We then test whether the verifier judges faithfulness correctly, in two steps. First, agreement: two non-author annotators relabel 1,240 claims without seeing the verifier's output ($\kappa = 0.79$); on a calibration set the verifier agrees with the human majority on 93.4% of claims, and re-ranking the scored systems by human labels instead of the verifier's leaves every rank unchanged ($\tau = 1.00$). Second, sensitivity: agreement could hide shared blind spots, so we corrupt clean claims with injected and fabricated errors and ask both the verifier and two blind annotators to find them. The verifier detects 88.4% of the errors while flagging only 1.6% of clean claims, exceeding its human reference (84.1%, 2.9%). Both audits hold, so we treat PRESENTFAITH's verdicts as the reference throughout. Protocols and secondary audits are in Section B.

5.3 COMPARISON WITH EXISTING SYSTEMS

Main results Having certified that PRESENTFAITH measures faithfulness reliably, we can now measure if EVI-IR is the most faithful system for paper-to-video. We score all the systems on Existence-Level faithfulness metrics and the four end-task metrics Table 1. Against the strongest baseline PAPERTALKER on faithfulness, EVI-IR lowers every error metric (UCR 17.6→12.5, NCR 9.7→2.6) while asserting more supported content per minute (SPM 11.7→13.6). Figure 4 (left) shows that the gap is systematic. On the (SPM, EPM) plane which is defined in Section 3.3, the post-hoc multi-agent systems move right along a band of near-constant error rate: more content brings proportionally more errors. EVI-IR performs better at this point. Meanwhile, faithfulness also does not come at the cost of quality: EVI-IR is within noise of PAPERTALKER on Meta Similarity and IP Memory and marginally ahead on PresentQuiz and PresentArena.

5.4 ABLATION

Section 5.3 rules out the external explanations; what remains is to account for the gain inside the method. EVI-IR has three parts that could produce it: the structured representation, the evidence links tying each claim to a unit, and the verifier that checks those links. They cannot be removed independently, because the verifier is defined over the representation, so a variant without it has no links to verify. We therefore remove them from the outside in, so that adjacent rows of Table 2 differ by exactly one component. All variants share the backbone, the evidence index, and the renderer.

Verification Firstly, we remove the verifier, and found that verification alone accounts for the largest share of the gain. The effect is large and concentrated where expected: UCR rises from 12.5 to 15.0, and NCR more than triples (2.6→8.4). Verification is thus the large single contributor to faithfulness.

Evidence Links We next isolate the evidence links from the unverified model still costs faithfulness: Both UCR and NCR have increased further, while Arena is unchanged. Because the verifier is no

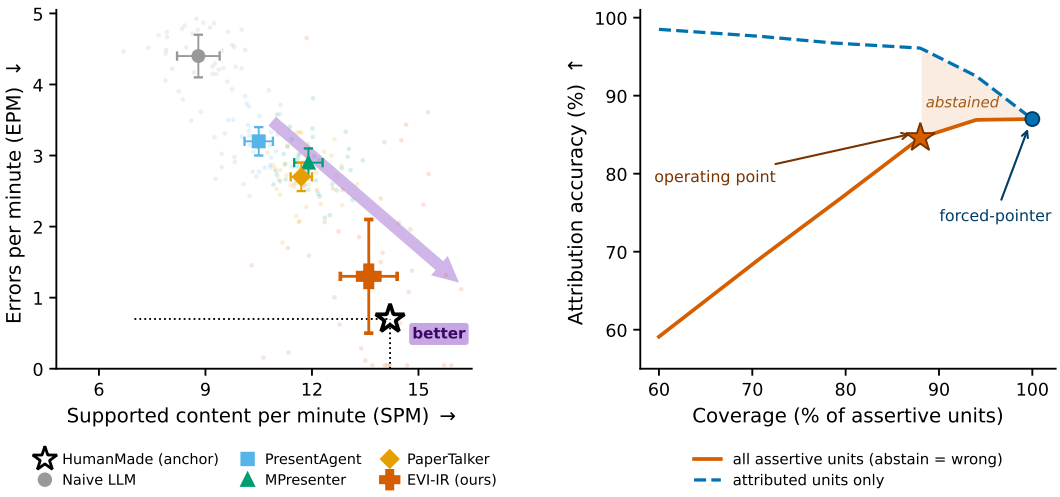


Figure 4: **Left:** One marker per system with per-paper clusters in the (SPM, EPM) plane; higher SPM and lower EPM are better, and the shaded corner marks the best region. Only EVI-IR moves down-right, while the multi-agent baselines move right along a band of near-constant error rate. **Right:** Attribution accuracy against coverage as the abstention threshold is swept; the star marks the operating point used throughout (84.6% accuracy at 88.0% coverage), and the right end is the forced-pointer variant.

	UCR↓	NCR↓	Ptr. acc.↑	Arena↑
EVI-IR	**12.5**	**2.6**	**84.6**	**51.4**
w/o Verifier	15.0	8.4	77.1	50.8
w/o Verifier and Links	16.2	9.5	—	50.9
w/o Verifier and IR	17.9	9.8	—	50.6

Table 2: **Ablation study** Removals are cumulative from the full method EVI-IR; adjacent rows differ by exactly one component, and each step isolates that component's contribution. UCR/NCR: faithfulness of the rendered video; Arena: end-task quality; Ptr.: pointer accuracy, defined only where pointers exist.

longer in the loop, the difference can only come from generation itself: anchoring each unit to explicit evidence yields more faithful output.

Intermediate Representation Finally, we remove the IR entirely and have the model plan in plain prose is the weakest variant: UCR reaches 17.9 and NCR 9.8. So it is confirmed that the constraint of IR also plays an important role in faithfulness.

6 CONCLUSION

This paper makes faithfulness to the source a first-class property of paper-to-video generation. On the measurement side, PRESENTFAITH scores every system against a fine-grained Evidence Index of the source paper, showing that current systems do well on standard presentation metrics yet still produce substantial unsupported and contradictory content. On the generation side, EVI-IR stores each claim with its type, evidence pointer, and quantities, checks the links with symbolic rules wherever possible, and repairs failures before rendering. Across 40 papers under one renderer, the unfaithful-claim rate falls from 17.6% to 12.5% and numeric contradictions from 9.7% to 2.6% against the strongest baseline, while supported claims per minute rise and presentation quality holds. Because the evidence link lives in the representation the presentation is generated from, each claim is traceable and its support directly checkable.

AI USE STATEMENT

In this work, we used generative AI tools for cleaning and reformatting the datasets used in PRESENTFAITH and for producing the visualizations in this paper. We have not used generative AI tools for generating synthetic data sets, developing the theoretical framework or its mathematical claims, providing critical ingredients for or assisting in proofs, proposing or refining hypotheses, designing the methodology or providing feedback on the experiments, assisting with translation, supporting qualitative or thematic analysis, or interpreting results, and the rest of the required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for creating figures and auxiliary scripts, improving the readability of our own text, and checking LaTeX formatting. We have reviewed all AI-assisted work. Data-processing and plotting code was inspected by the authors and checked against deterministic unit tests before being run on the benchmark, each figure was regenerated from the verified numbers so that no displayed value originates from a model, and suggested edits to the manuscript were checked line by line to ensure that no technical claim was altered. The models evaluated in this paper are components of the systems under study, not tools used to conduct the research. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Our work aims to make the claims of automatically generated presentations verifiable against their sources, and we believe it has a positive impact on the reliability of scientific communication and on the responsible development of paper-to-video systems. However, we also acknowledge potential risks and misuse of our framework. For example, systems of this kind could be used to generate persuasive videos that attribute fabricated claims to the work of real authors. To address such risks, PRESENTFAITH is designed to measure exactly this failure mode rather than enable it, and in real-world deployment it would be necessary to incorporate safeguards (such as machine-generated content labels, displayed provenance, and source-linked captions) to ensure responsible use of our framework.

REPRODUCIBILITY STATEMENT

We will release the public code for EVI-IR and the PRESENTFAITH annotations to reproduce our work; in the meantime, the code and data are available from the authors upon request. All necessary details to reproduce our experiments are described in Sections A, C and 5.1.

REFERENCES

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. Tabfact: A large-scale dataset for table-based fact verification. *arXiv preprint arXiv:1909.02164*, 2019.

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, JianZhao JianZhao, Kai Yu, and Xie Chen. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6255–6271, 2025.

Jiahao Cui, Hui Li, Yao Yao, Hao Zhu, Hanlin Shang, Kaihui Cheng, Hang Zhou, Siyu Zhu, and Jingdong Wang. Hallo2: Long-duration and high-resolution audio-driven portrait image animation. In *International Conference on Learning Representations*, volume 2025, pp. 91659–91671, 2025.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. Chain-of-verification reduces hallucination in large language models. In *Findings of the association for computational linguistics: ACL 2024*, pp. 3563–3578, 2024.

Tsu-Jui Fu, William Yang Wang, Daniel McDuff, and Yale Song. Doc2ppt: Automatic presentation slides generation from scientific documents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 634–642, 2022.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, et al. Rarr: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16477–16508, 2023a.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6465–6488, 2023b.

Jiaxin Ge, Zora Zhiruo Wang, Xuhui Zhou, Yi-Hao Peng, Sanjay Subramanian, Qinyue Tan, Maarten Sap, Alane Suhr, Daniel Fried, Graham Neubig, et al. Autopresent: Designing structured visuals from scratch. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2902–2911. IEEE, 2025.

Saibo Geng, Martin Josifoski, Maxime Peyrard, and Robert West. Grammar-constrained decoding for structured nlp tasks without finetuning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 10932–10952, 2023.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14375–14385. IEEE, 2024.

Wenhan Han, Xiao Xiao, Mykola Pechenizkiy, and Meng Fang. mpresenter: An agentic framework for generating multilingual presentation videos from scientific papers. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 16358–16371, 2026.

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pp. 9332–9346, 2020.

Chengzhi Liu, Yuzhe Yang, Kaiwen Zhou, Zhen Zhang, Yue Fan, Yanan Xie, Peng Qi, and Xin Eric Wang. Presenting a paper is an art: Self-improvement aesthetic agents for academic presentations. *arXiv preprint arXiv:2510.05571*, 2025a.

Jingwei Liu, Ling Yang, Hao Luo, Fan Wang, Hongyan Li, and Mengdi Wang. Preacher: Paper-to-video agentic system. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17129–17139. IEEE, 2025b.

Qianli Ma, Jipeng Xiao, Siyu Wang, Zhiheng Tian, Wangyu Feng, Shibo Wang, Chang Guo, Shuochen Chang, Qingyang Liu, and Zhipeng Zhang. Omnipresent: Generating coherent presentation suites from scientific papers. *arXiv preprint arXiv:2607.02590*, 2026.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 1906–1919, 2020.

Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, et al. Teaching language models to support answers with verified quotes. *arXiv preprint arXiv:2203.11147*, 2022.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 12076–12100, 2023.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.

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

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pp. 28492–28518. PMLR, 2023.

Jingwei Shi, Zeyu Zhang, Biao Wu, Yanjie Liang, Meng Fang, Ling Chen, and Yang Zhao. Presentagent: Multimodal agent for presentation video generation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 760–773, 2025.

Aviv Slobodkin, Eran Hirsch, Arie Cattan, Tal Schuster, and Ido Dagan. Attribute first, then generate: Locally-attributable grounded text generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3309–3344, 2024.

Yixiao Song, Yekyung Kim, and Mohit Iyyer. Veriscore: Evaluating the factuality of verifiable claims in long-form text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 9447–9474, 2024.

Zusheng Tan, Jing-Yu Ji, Wenhui Yu, Ngai Fung Ng, Fan Yang, Jeff Tang, Ken Fong, Jing Li, Sam Kwong, and Billy Chiu. Localize-then-summarize: Enhancing scientific multimodal summarization with facet-aware cross-modal memory. *Information Processing & Management*, 64(1): 104987, 2027.

Liyan Tang, Philippe Laban, and Greg Durrett. Minicheck: Efficient fact-checking of llms on grounding documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8818–8847, 2024.

Alex Wang, Kyunghyun Cho, and Mike Lewis. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 5008–5020, 2020.

Yuxuan Wang, Yueqian Wang, Dongyan Zhao, Cihang Xie, and Zilong Zheng. Videohallucer: Evaluating intrinsic and extrinsic hallucinations in large video-language models. *arXiv preprint arXiv:2406.16338*, 2024.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, et al. Long-form factuality in large language models. *Advances in Neural Information Processing Systems*, 37:80756–80827, 2024.

Brandon T Willard and Rémi Louf. Efficient guided generation for large language models. *arXiv preprint arXiv:2307.09702*, 2023.

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

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, volume 2025, pp. 83048–83077, 2025.

Xiang Yue, Boshi Wang, Ziru Chen, Kai Zhang, Yu Su, and Huan Sun. Automatic evaluation of attribution by large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 4615–4635, 2023.

Jiacheng Zhang, Yang Jiao, Shaoxiang Chen, Na Zhao, Zhiyu Tan, Hao Li, Xingjun Ma, and Jingjing Chen. Eventhallusion: Diagnosing event hallucinations in video llms. *arXiv preprint arXiv:2409.16597*, 2024.

Hao Zheng, Xinyan Guan, Hao Kong, Wenkai Zhang, Jia Zheng, Weixiang Zhou, Hongyu Lin, Yaojie Lu, Xianpei Han, and Le Sun. Pptagent: Generating and evaluating presentations beyond text-to-slides. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 14402–14418, 2025.

Zeyu Zhu, Kevin Qinghong Lin, and Mike Zheng Shou. Paper2video: Automatic video generation from scientific papers. *arXiv preprint arXiv:2510.05096*, 2025.

APPENDIX

A PRESENTFAITH CONSTRUCTION

A.1 EVIDENCE INDEX

Parsing We expand macros against a whitelist. We then flatten `tabular`, `tabularx` and `booktabs` environments into (row path, column path) trees using `\multicolumn`/`\multirow` spans, and attach each table to its nearest `\caption` and `\label`. A grammar over {sign, digits, separators, exponent, unit, ±error} parses numeric surfaces, which are stored with their displayed precision.

Schema Each unit stores four fields:

- a content-addressed identifier, `sha1(paper_id, env_index, row_path, col_path)`, which stays the same under re-parsing;
- its type: paragraph, equation, figure caption, table cell, or numeric mention;
- its hierarchical path, stored separately so that near-misses can still be scored;
- its verbatim content.

Numeric units also store a normalised (value, unit, precision) triple and the quantity the number denotes, read from its row and column context. PDF-parsed indices are marked DEGRADED and are never pooled with LaTeX-parsed ones. None of the 40 scored papers uses this mode.

A.2 CLAIM EXTRACTION AND FAITHFULNESS IDENTIFICATION

Extraction We transcribe the narration with word-level ASR and read slide text by OCR on slide keyframes. Both streams are split into atomic claims. Non-assertive spans (agendas, transitions, hedged speculation) are labelled ABSTAIN and not scored.

Retrieval The verdict function decides numeric and comparative claims by index lookup and arithmetic, and all other claims by entailment against retrieved candidates. A delivered video carries no pointers, so the benchmark must *retrieve* candidate units. Retrieval depends on the atom type:

- *Numeric and comparative atoms.* We look up the atom's (entity, metric) pair in the index, which is keyed by row and column path. Entity and metric strings are first normalised against the alias set induced by the paper's own macro expansions.
- *All other atoms.* Candidates are the union of the top-k units by BM25 and the top-k units by embedding similarity over unit text, with $k = 8$.

Verdicts A claim is labelled:

- SUPPORTED if *any* candidate supportes it (a value often appears in a table cell, the abstract and the results prose alike);
- CONTRADICTED if a candidate matches on entity and metric but conflicts in value or direction;
- UNSUPPORTED if no candidate matches;

A.3 ANNOTATION PROTOCOL

Two non-author annotators, each with ≥ 3 years of ML research experience, label every item, and a third resolves disagreements. Annotators receive the paper, the delivered video and the extracted claim. For each claim they mark: (i) whether the span is assertive or exempt; (ii) the *acceptable set* of evidence units, any of which would supporte the claim; and (iii) a verdict. We use a set rather than a single referent because the same number often appears in a cell, the abstract and the results prose.

Table 3: Benchmark validity. We audit the three properties our conclusions depend on: the index $\mathcal{U}(P)$ (ground truth), the verifier (judge), and sensitivity to injected and fabricated errors. τ is Kendall's rank correlation over the scored systems; FPR is the false-positive rate on clean claims. Secondary audits (a second judge model; correlation with end-task metrics) are in Section B.

Property	Measurement	Value
Index integrity ($\mathcal{U}(P)$)	unit accuracy / mislabeled spans (300 units, 2 annotators)	95.7% / 2.3%
Judge fidelity	inter-annotator Cohen's κ (1,240 claims, 2 non-authors)	0.79
	judge vs. human labels (calibration set)	93.4%
	τ (judge ranking vs. human ranking)	1.00
Sensitivity	verifier: detection / FPR (injected & fabricated errors)	88.4% / 1.6%
	blind humans: detection / FPR	84.1% / 2.9%

B Benchmark Validation

Our conclusions rest on two artifacts: the index $\mathcal{U}(P)$, which is the ground truth, and the verifier, which is the judge. This appendix gives the protocols behind Table 3.

B.1 Index integrity

Two annotators check a type-stratified sample of 300 units from the 40 papers, including 200 table cells. Each unit is checked for boundary, type label, and content. 95.7% of units are correct. Most of the remaining 4.3% are boundary errors with the content intact; the few rest have a wrong type label or cell path, caused by nested `multirow` spans and footnote markers.

Span labeling The extraction pipeline also labels every span assertive or ABSTAIN. We re-label 300 spans that it marked ABSTAIN; 2.3% are in fact assertive, mostly hedged comparatives such as "roughly twice as fast". This is the "mislabeled spans" entry of Table 3. A system cannot lower its UCR by speaking vaguely.

B.2 Judge fidelity

Two non-author annotators relabel 1,240 claims without seeing the verifier's output; a third resolves disagreements. The two annotators agree at $\kappa = 0.79$ on verdicts. On a calibration set of 500 claims drawn from the same stratified sample, the verifier agrees with the human majority on 93.4% of claims. Ranking the five scored systems by human labels leaves every rank unchanged ($\tau = 1.00$).

B.3 Sensitivity to injected and fabricated errors

We corrupt clean claims so that every error's location is known. The operators are numeric perturbation, unit swap, entity (baseline) substitution, comparison flip, and qualifier deletion; we also add fabricated claims, which state content absent from the paper. The operators were fixed before the verifier's checks were implemented, and the two annotators who serve as the human reference work blind, with the paper in hand.

The verifier detects 88.4% of the planted errors and flags only 1.6% of clean claims; the human reference detects 84.1% with 2.9%. The gap comes from the numeric operators, where the verifier looks up the exact cells and recomputes the arithmetic that humans must do by eye; on qualifier deletion, the only semantic operator, the two are within a point of each other. The verifier is therefore at least as sensitive as its human reference, so it can serve as the judge of record.

The same errors show why the usual alternative cannot judge this task. A retrieval-plus-entailment check detects 79% of numeric perturbations but only 38% of comparison flips and 31% of baseline swaps: in those two cases every value appears in the paper, so checking by existence returns SUPPORTED. Our verifier decides these claims on its number path instead: it looks up the cells the claim names and recomputes the arithmetic.

B.4 SECONDARY AUDITS

Second judge model We swap the entailment judge for Gemini-2.5-Flash and re-run the verdict pipeline on all claims. Verdicts agree on 92.4% of claims; no system's UCR moves by more than 0.7 points, and the system ranking is unchanged.

Correlation with end-task metrics We compute the Spearman correlation between UCR and each of the four end-task presentation metrics across the five systems. No pair reaches significance (largest $|\rho| = 0.26$): presentation quality and faithfulness are separate axes.

C EVI-IR DETAILS AND ADDITIONAL RESULTS

C.1 RENDER-ELEMENT CO-GENERATION

Published pipelines call a model again at render time for layout search, visual choice and cursor grounding. This is their largest cost item. It also adds an unverified step between the artifact that was checked and the artifact the viewer sees.

Each atom already names the unit it asserts, so the layout information is available at generation time. Alongside each atom, EVI-IR emits:

- its slide assignment and layout slot;
- visual directives: which figure or cell to display and highlight;
- TTS markup: numeral reading, units and emphasis.

The renderer then runs. The slide and the narration share one source, so a slide cannot display a different cell from the one its sentence cites.

Render preservation Checking before rendering is meaningful only if rendering preserves what was checked. For every atom with $y^{\rightarrow} = \text{SUPPORTED}$, we test whether its asserted quantities can be recovered from the delivered video by transcript and OCR alignment. We probe every point where assertion-bearing content can change: TTS numeral renormalisation ("84.3"→"eighty four point three"), slide-text overflow, figure cropping. Overall, 99.2% of SUPPORTED atoms are preserved (Section 5.3).

C.2 ATTRIBUTION ACCURACY

Because EVI-IR keeps every claim's evidence pointer explicit in the representation, it admits an evaluation that no baseline supports: checking not just whether a claim is supported, but whether it points at the very evidence that supports it. We evaluate the pointers against human acceptable sets built by two annotators. The criterion is strict—a pointer counts as correct only if it lies in both annotators' acceptable sets, and abstentions are never scored as correct—and the evaluation reads the intermediate representation rather than the delivered video; no baseline emits provenance, so it covers EVI-IR only.

Under this criterion EVI-IR attributes 88.0% of atoms and abstains on 12.0%, with 84.6% pointer accuracy overall—6.6 points below the human ceiling of 91.2% (Table 4). Accuracy tracks matchability: numeric and comparative atoms are the easiest (88.1%) because their values can be matched verbatim, descriptive atoms are the hardest (78.9%) because paraphrase leaves fewer acceptable targets, and definitional and attributive atoms sit in between (82.3%).

The dominant error mode is visible only at cell granularity. On numeric atoms, table-level accuracy reaches 93.0% while cell-exact accuracy is 79.4%. This 13.6-point gap consists mostly of near-miss pointers—right table, wrong cell—which a table-granularity index would hide entirely. The corresponding human ceiling (86.4% cell-exact agreement) suggests that part of the remaining error is ambiguity in the acceptable sets rather than wrong pointers. The pointers are therefore reliable enough for each claim's evidence to be checked directly—the traceability that the evidence-linked representation is designed to provide, and that evaluation at the claim level alone cannot establish.

Table 4: Attribution quality of EVI-IR against human acceptable sets. Of all spans, 88.0% are attributed and 12.0% abstain.

Atom subset	Pointer acc.↑	Cell-exact acc.↑
Numeric & comparative	88.1	79.4
table-level scoring	93.0	—
Definitional & attributive	82.3	—
Descriptive	78.9	—
All assertive	84.6	—
Human ceiling (annotator Jaccard)	91.2	86.4

C.3 COST AND BACKBONE ROBUSTNESS

Attribution aside, a practical system must also be affordable and robust: the gains should not depend on a strong generator or a large budget. We therefore report the token and wall-clock cost of EVI-IR relative to the baselines, and its sensitivity to the generator backbone.

On cost, EVI-IR is the cheapest system to run end to end (Table 5): 421K tokens and 179 seconds per paper, 18% fewer tokens and 20% less wall-clock time than the strongest baseline PAPERTALKER (512K, 224s). PRESENTAGENT spends fewer tokens (286K) but more than twice the wall-clock time (413s), so no baseline is cheaper on both axes. Explicit evidence links thus do not make EVI-IR expensive—the structured representation lets it verify claims without the long iterative pipelines the multi-agent baselines rely on.

On robustness, we swap the generator backbone while keeping everything else fixed. Faithfulness degrades only mildly as the backbone weakens (Table 6): UCR moves from 12.5 (GPT-4.1) to 12.7 (GPT-4o) to 13.1 (Qwen2.5-7B-Instruct), and NCR from 2.6 to 3.9. Even a 7B backbone keeps EVI-IR far ahead of every baseline—the strongest of them stands at 17.6 UCR (Table 1). These results hold for the tested backbones, but the pattern is consistent: the checks in the representation, not the generator's raw capability, carry the faithfulness. The gains therefore survive a much weaker backbone, and come at a cost below the strongest baseline's on both axes.

Table 5: Cost per paper: total tokens and wall-clock time.

System	Tokens (K)	Wall-clock (s)
PRESENTAGENT	286	413
PAPERTALKER	512	224
MPRESENTER	708	201
EVI-IR	**421**	**179**

Table 6: Backbone swap.

Backbone	UCR↓	NCR↓
GPT-4.1	12.5	2.6
GPT-4o	12.7	3.0
Qwen2.5-7B-Instruct	13.1	3.9

D LIMITATIONS

EVI-IR's residual errors concentrate where the symbolic layer does not reach. Relational atoms that span three or more cells comparing gains across several splits, for example fall outside the pairwise arithmetic check and form the largest share of what remains. Descriptive atoms with a genuinely diffuse acceptable set, pointers into figure internals, and extraction artifacts introduced by ASR, OCR, and segmentation account for the rest. A narration is also not a fully symbolisable object: a substantial share of atoms is settled symbolically, a smaller share abstains, and the remainder

depends on an entailment model and inherits its errors, so we report these quantities separately rather than blending them. The referent throughout is the paper, not the world, so a claim that faithfully repeats an erroneous value counts as supported; adjudicating truth is a different problem. The empirical claims are scoped to numerically dense English papers from ML, CV, and NLP and verified under the backbones tested; they transfer only as far as rendering preserves them. The boundary analysis assumes a deterministic renderer, though a renderer that consumed provenance would move the boundary without removing it.