DOCUMENT-GROUNDED GENERATION OF VERIFIED, INTERACTIVE AND PHOTOREAL PHYSICS EXPERIMENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Most scientific documents convey physical processes through static figures, which leave the dynamics for readers to figure out. Generative models can turn such documents into interactive simulations and videos. These outputs often look plausible, yet it is hard to ensure they obey the physical laws the document states. Our goal is to turn a document's figures into interactive simulations and videos that help readers understand the physics. These simulations must follow the physical laws stated in the document. We represent the document as a graph over its equations, quantities, and validity regimes. Retrieval on this graph continues until the collected equations form a closed, solvable system. This system is recorded as a physics scenario card. The card states what to solve, the governing dynamics, the quantities a reader controls, and their valid ranges. The same card drives generation and supplies tests derived from the document. Generated behavior is therefore checked against the document's own relations. Closure-terminated retrieval recovers substantially more of the required context than similarity search. An evaluator answers figure-specific questions from our interactive artifacts at 0.923, against 0.389 for single-call generation. In photorealistic videos, motion measured from the frames matches the closed form derived from the document. Without physical conditioning, the same video model produces no measurable period. This work enables a scientific document to demonstrate the physical behavior it describes. It opens a path from the static figures of education, science communication, and engineering to interactive, verifiable representations.

1 INTRODUCTION

Technical documents explain physical processes with static figures, leaving readers to imagine how these processes unfold. Dynamic representations make this temporal behavior explicit. A meta-analysis reported a mean learning advantage of instructional animation over static pictures (Höffler & Leutner, 2007). Long continuous animations can also increase working-memory demands (Wong et al., 2012). Interactive simulations let learners vary parameters, control the pace, and observe the consequences. Such simulations are used widely for inquiry and conceptual learning in science and physics (Wieman et al., 2008; De Jong, 2006; De Jong et al., 2013). They are built for a curated set of topics, not for each figure in a document. Most technical documents therefore still present physical processes as static diagrams and photographs. Building a faithful interactive version of each figure remains manual work.

Large language models now support rapid authoring of interactive and animated educational content. Recent work includes AI-generated physics simulations (Ben-Zion et al., 2025; He et al., 2025), agentic systems that produce explanatory animations and videos (Ku et al., 2025; Chen et al., 2025; P et al., 2025), and automated evaluation of AI-generated explorable explanations (Wang et al., 2026a). A central remaining issue is reliability: generated code can execute and appear plausible while implementing the wrong physical relationship. On a benchmark of executable animations for 52 physics concepts, the strongest evaluated model produces code that both runs and matches the target physics in only 21.5% of attempts (Wang et al., 2026b). A document-grounded system must therefore do more than generate an artifact; it must determine which relations are required, expose

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.

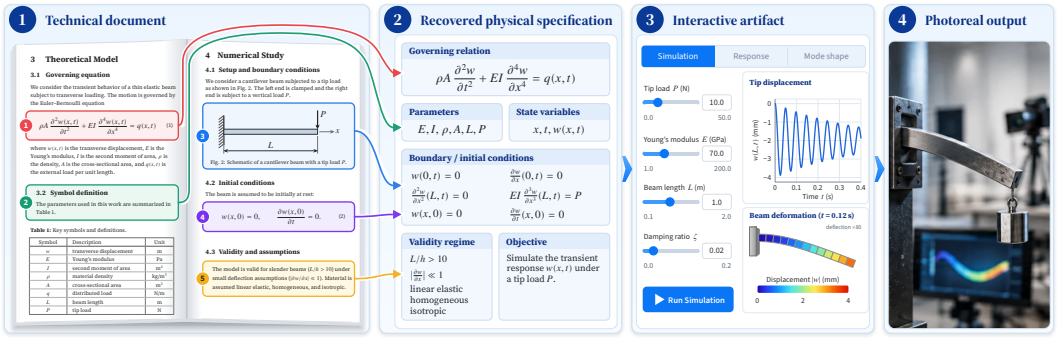

Figure 1: The closure walk turns a document into a physics scenario card (PSC). Colored links trace each card entry to its source passage. The PSC drives the interactive artifact and conditions photoreal generation.

controls consistent with the document's objective, and verify that the generated behavior satisfies those relations.

We address this problem by treating the document itself as an executable specification. For a target figure, we construct a graph over equations, quantities, definitions, and validity regimes, then retrieve from that graph until the accumulated relations form a physically closed system. The result is represented as a *physics scenario card* (PSC), which records the state variables, reader-controlled parameters, governing relations, validity regimes, verification quantities, objective, and provenance. The PSC is used both to generate the artifact and to derive tests that are evaluated independently of the generative model.

The same representation supports two output modalities. For an interactive figure, the PSC conditions code generation and provides the tests used to accept or reject the resulting executable artifact. For a photographic figure, a verified physical trajectory is converted to visual conditioning for a video diffusion model and the relevant physical quantity is measured again from the delivered frames.

Our contributions are threefold. **First**, we introduce objective-anchored retrieval that terminates on physical closure rather than a fixed retrieval budget; on 418 documents it reaches 31.4% sufficiency@10 and 0.449 non-local recall, compared with 21.0% and 0.105 for dense retrieval. **Second**, we introduce a PSC-based synthesis and verification pipeline for interactive scientific figures; across 26 matched figures, the generated artifacts achieve an automated evaluator score of 0.923 versus 0.389 for single-call generation. **Third**, we extend the same specification to physics-conditioned photorealistic generation and verify the delivered output directly; across three control values, measured periods differ from the document-derived closed form by 4.5%, 1.1%, and 1.1%.

2 Related Work

Retrieval over documents. Retrieval-augmented generation combines dense retrieval with generation over a selected set of passages (Lewis et al., 2020). GraphRAG instead constructs an entity–relation graph and community summaries to support local and global question answering (Edge et al., 2025). These methods define relevance operationally through retrieval or summarization, rather than through a domain-specific test of whether the retrieved content is sufficient to instantiate a physical system. Math-aware retrieval has developed expression-search methods and benchmarks (Zanibbi & Blostein, 2012; Mansouri et al., 2021; Zanibbi et al., 2025), while reader interfaces such as ScholarPhi link terms and symbols to definitions elsewhere in a paper (Head et al., 2021). Our setting requires a different output, a closed, executable system whose relations can be checked before generation.

Program synthesis under verification. Self-debugging and self-repair methods use execution feedback or unit tests to improve generated programs (Chen et al., 2024), with gains that depend strongly on the quality of the available feedback (Olausson et al., 2024). In formal reasoning, systems can additionally rely on proof assistants or symbolic deduction engines to check generated proofs (First et al., 2023; Trinh et al., 2024). In those settings, the checker or specification already exists.

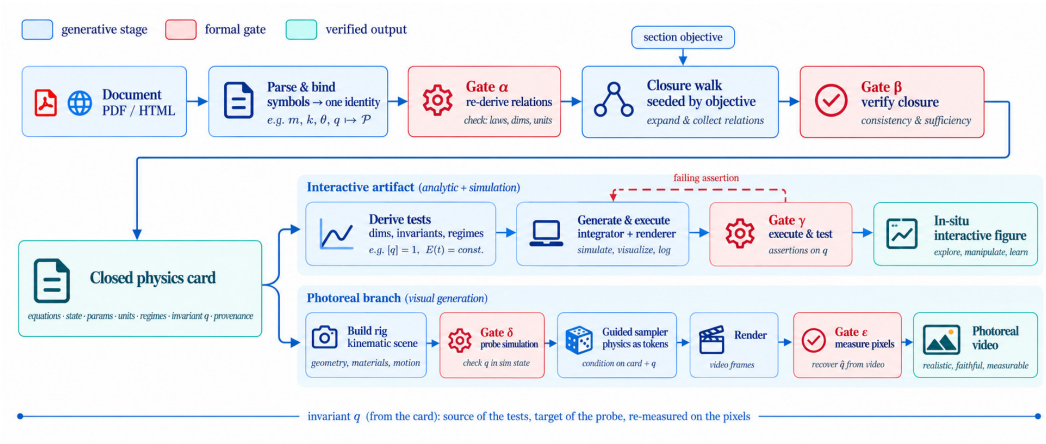

Figure 2: Method overview. Document parsing and symbolic re-derivation feed objective-anchored retrieval until physical closure. The resulting PSC drives both interactive synthesis and physics-conditioned photoreal generation, with Gates α–ε evaluated outside the corresponding generative stages.

Here the source is an ordinary technical document, so the tests must first be derived from the recovered physical relations.

Controllable and physics-aware video generation. Classifier-free guidance (Ho & Salimans, 2022) is widely used to steer diffusion video models (Ho et al., 2022; Blattmann et al., 2023): conditioning architectures such as ControlNet incorporate structural signals (Zhang et al., 2023), and reference-driven animation preserves appearance from a supplied image while following a motion signal (Hu, 2024). These methods target visual or structural control rather than recovery of a governing law from a source document and verification of the delivered video against that law. VideoPhy and PhyGenBench directly evaluate physical commonsense in generated video and report substantial failures among current text-to-video systems (Bansal et al., 2025; Meng et al., 2025). Physics-informed neural networks, by contrast, incorporate governing equations that are supplied to the model (Raissi et al., 2019); our pipeline first has to recover the relevant equations from the document.

AI for education. Large language models are now being explored for educational content creation, tutoring, and personalized support, alongside well-documented concerns about reliability and oversight (Kasneci et al., 2023). Interactive simulations such as PhET provide carefully designed, manipulable representations (Wieman et al., 2008), but they are curated resources rather than artifacts automatically grounded to each figure in a source document. We target this document-specific generation problem.

3 Method

Figure 2 summarizes the pipeline. Starting from a target figure and its surrounding document context, we recover a closed physical specification, represent it as a physics scenario card, and use the card to generate either an interactive artifact or a physics-conditioned photorealistic rendering.

3.1 From a document to a closed physical specification

Let D denote a document region associated with a target figure. We represent its physical content as a system Σ over state variables X, reader-controlled parameters P, and known constants K. The document is parsed into equations, surrounding prose, objectives, and validity regimes. Each symbol occurrence is represented by its host equation, surface form, physical dimension, and prose gloss. Occurrences are first partitioned by exact agreement of surface form and dimension, and semantic judgments are then used only within a compatible partition to resolve whether the glosses denote the same physical quantity. The accepted pairwise bindings are closed by union–find to produce one graph node per quantity; Appendix B gives the complete binding equations and propagation rules.

The resulting graph $G = (V, E)$ contains document sections, equations, physical quantities, and validity regimes. Structural edges record which quantities an equation references and which quantity its left-hand side defines. Relations *between* equations require a separate verification step. A language model proposes a candidate relation $h = (e_i, e_j, \rho)$, where ρ denotes the proposed relation type (e.g., definition, rearrangement, or combination). Gate α accepts the edge only when an independent symbolic derivation can reproduce it:

$$\mathrm{Gate}_\alpha(h) = \mathbf{1}[\mathrm{ReDerive}(e_i, e_j; \rho) = \mathrm{true} \ \wedge \ \mathrm{DimOK}(e_i, e_j)]\,. \tag{1}$$

ReDerive attempts the corresponding symbolic solve, substitution, or elimination and requires the proposed relation to follow exactly after simplification. Contradicted or unverifiable proposals are discarded.

Retrieval is anchored to the objective o stated by the figure or its surrounding text. We resolve o to a seed equation e_0 from the document and expand from $S_0 = \{e_0\}$ by gathering equations that define quantities still unresolved by the current set. Let $U(S)$ denote quantities that are neither defined by S nor included in K, $n(S)$ the number of non-parameter unknowns, and $R(S)$ the number of independent relations. Retrieval terminates at the first set satisfying

$$\mathrm{Closed}(S) \equiv [\,U(S) \subseteq P\,] \wedge [\,R(S) \geq n(S)\,] \wedge \mathrm{Consistent}(S) \wedge \mathrm{DimOK}(S) \wedge \mathrm{WellPosed}(S). \tag{2}$$

The additional predicates require a common solution, dimensional consistency, and a well-posed dynamical system. Appendices C and D give their implementation and the complete walk state. Validity regimes recovered with the equations are retained as executable constraints on the admissible parameter range.

3.2 Physics scenario card and verified artifact generation

The closed system is converted into a *physics scenario card*

$$\mathcal{C} = (X, P, K; \Sigma; \mathcal{R}, \mathcal{I}, \mathcal{L}, \mathcal{D}; o, \pi),$$

where $\mathcal{R}$ contains validity regimes, $\mathcal{I}$ physical invariants and conserved quantities, $\mathcal{L}$ limiting cases, $\mathcal{D}$ dimensional constraints, and π provenance. The PSC is the interface between retrieval and all downstream generation stages; its full schema and per-stage read sets are given in Appendix E.

The PSC deterministically induces executable tests rather than relying on a second model judgment. Dimensional tests require additive terms to reduce to a common SI-base-dimension vector. For every conserved quantity $E \in \mathcal{I}$ and generated trajectory $Y = \{y_k\}_{k=0}^{N}$, the conservation assertion is

$$\max_{0 \leq k \leq N} \left| \frac{E(y_k) - E(y_0)}{E(y_0)} \right| \leq \epsilon_E, \tag{3}$$

where ϵ_E is stored in the PSC. Stated limiting relations become limiting-case tests, and validity regimes become executable predicates over the exposed controls. Appendix C specifies the corresponding consistency and well-posedness checks.

A code-generation model receives the PSC and the derived test set $\mathcal{T}(\mathcal{C})$, but not the source prose, and emits an integrator and renderer $G_\theta(\mathcal{C})$. Gate γ executes the program to obtain Y and admits it only when every derived assertion passes:

$$\mathrm{Gate}_\gamma(G_\theta) = \mathbf{1}\left[\mathrm{Execute}(G_\theta) \ \wedge \bigwedge_{t \in \mathcal{T}(\mathcal{C})} t(Y)\right]. \tag{4}$$

A failing assertion and its measured value are returned as structured feedback for a bounded number of repair attempts.

An accepted interactive artifact is substituted into the region occupied by the original figure. Placement is determined from the document's figure coordinates, while internal geometry is recovered from the source figure. The resulting controls therefore correspond directly to quantities in P, and the recovered validity regimes remain active during interaction. Implementation details for integration and in-situ substitution are provided in Appendices F and V.

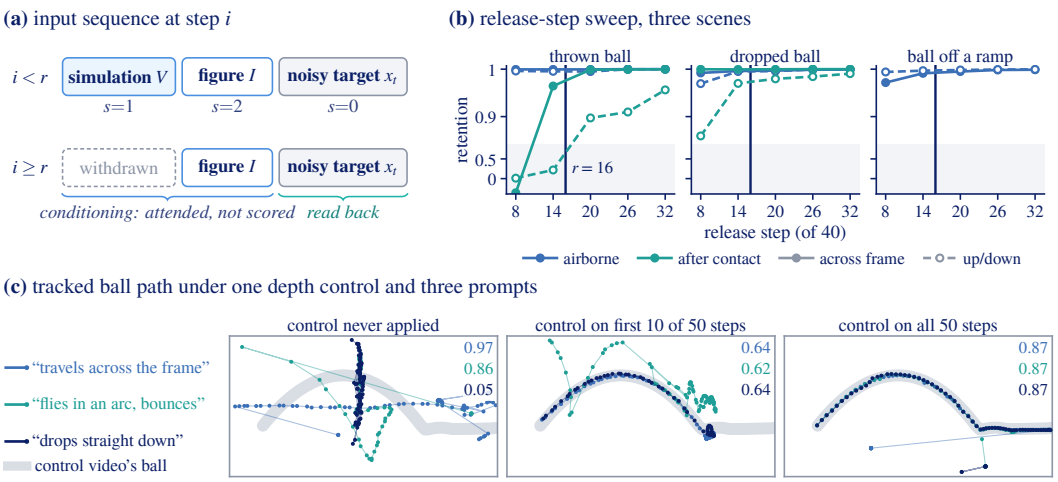

Figure 3: Physical conditioning and its release. (a) Each conditioning stream joins the noisy target with its own source id. The simulation tokens are withdrawn from step r onward. (b) Retention is the correlation between the tracked path and the simulation, split at first ground contact. Releasing at step 8 reverses the thrown ball's motion after contact. The grey band marks retention below the 0.7 threshold. (c) One depth control with three prompts; each number gives a clip's sideways travel. Without control, the path follows the prompt. With control throughout, all three prompts reproduce the control's arc.

3.3 Physics-conditioned photorealistic generation

For photographic figures, the accepted state trajectory $X(t)$ is mapped to a kinematic visual scene $S(t) = \Phi(X(t); P)$, where Φ maps physical state and card parameters to scene pose. Gate δ uses an independently implemented, lightweight probe M_δ to re-measure the PSC's invariant of record from the simulated scene,

$$\hat{q}_\delta = M_\delta(S_{0:T}), \qquad \mathrm{Gate}_\delta(S) = \mathbf{1}\left[\frac{|\hat{q}_\delta - q(\mathcal{C})|}{|q(\mathcal{C})|} \le \tau_\delta\right]. \tag{5}$$

Only a scene passing this pre-render probe is sent to the video model. Each scenario family defines its own invariant and tolerance (Appendix G).

The verified simulation is then encoded as conditioning tokens for the video diffusion model together with the source figure and text prompt. If $c_1, \ldots, c_k$ denote conditioning streams and x_t the noisy target latent, the packed sequence is

$$z^{(C)} = [c_1; \ldots; c_k; x_t], \qquad \hat{\varepsilon}^{(C)} = f_\phi(z^{(C)}, \tau_i, c_{\mathrm{txt}})[\mathbf{m}], \tag{6}$$

where $\mathbf{m}$ selects the target positions. The simulation conditioning is retained during early denoising and released at step r so that it constrains motion formation without fixing the final appearance to the synthetic simulation. The full guidance formulation, sampler configuration, and release sweep are given in Appendices N and U. At step r the simulation tokens leave the sequence, and the multistep solver's history is reset. Figure 3 shows the packed sequence and the effect of the release step.

Finally, Gate ε re-measures the invariant of record directly from the delivered frames using independently implemented instruments. For the oscillatory case evaluated in Section 4.5, a render is accepted when

$$|\hat{T}_1 - \hat{T}_2| \le \delta_{\mathrm{inst}} \quad \wedge \quad \max_{j \in \{1,2\}} \frac{|\hat{T}_j - T|}{T} \le \tau_\varepsilon. \tag{7}$$

Table 1: Full-context retrieval on 418 documents with 28,535 queries per configuration at k=10. Non-local recall is pooled across queries. The final two columns report sufficiency@10 at the cleanest (L_0) and noisiest (L_4) query conditions; the same ordering holds at intermediate noise levels. Additional precision, F_1, and bootstrap results are reported in Appendix P.

		pooled		suff.@10 by level	
	retrieval method	suff.@10	non-loc. rec.	L_0	L_4
baselines	BM25 (lexical)	0.178	0.049	0.202	0.163
	dense (embedding)	0.210	0.105	0.240	0.201
ablations	PTBG only (symbol graph)	0.249	0.284	0.283	0.209
	+ graph rerank	0.272	0.335	0.312	0.229
this work	**closure walk**	**0.314**	**0.449**	**0.361**	**0.263**
	Δ over best baseline	+0.104	+0.344	+0.122	+0.062

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Retrieval. The corpus contains 448 physics documents drawn from an openly licensed university physics text (Moebs et al., 2016) and arXiv preprints. Of these, 418 parse without error and form the analysis set. Retrieval is evaluated on 5,707 equations under five levels of lexical degradation, L_0 (clean) to L_4. This gives 28,535 queries per configuration and 142,675 scored query–configuration pairs. We compare against BM25 and dense embedding retrieval. Two ablations isolate our components: the symbol graph alone (PTBG), and the symbol graph with reranking. The primary measure is sufficiency@10. It is the fraction of queries whose retrieved set contains the target equation's complete annotated dependency closure. Non-local recall is the fraction of required parts recovered that lie away from the equation in reading order. Confidence intervals use 2,000 document-level bootstrap resamples.

Interactive artifacts. For each matched figure, we build the proposed artifact and a single-call baseline from the same page. The baseline receives the whole page and an instruction to build an interactive figure. An evaluator answers eight multiple-choice questions per figure and never sees the source text. It operates each artifact through a fixed interface: list controls, set a control, read an output, and request the displayed relation. An answer of "cannot tell" counts as incorrect. The printed figure is the reference condition, and an artifact built for a different figure is the floor. TQ subtracts the evaluator's no-evidence score. Rules fixed before the run exclude figures whose no-evidence score reaches one half. They also exclude figures whose printed version scores no better than no evidence (Appendices T and I).

Photorealistic generation. The video backbone is an open reference-video-to-video diffusion model. Each render has 81 frames at 16 fps and uses 40 sampling steps with release at $r = 16$. We render the printed swing photograph at $L = 0.40$, 0.60, and 0.90 m. A segmentation tracker (SAM 2.1; Ravi et al., 2025) and a background-subtraction blob tracker measure each delivered clip. A render passes when it satisfies Equation (7) at $\tau_\varepsilon = 5\%$. Baselines are the same backbone without simulation conditioning, and Veo 3.1 and Runway Gen-4.5 from text prompts. Clips from text prompts carry no metric scale. We therefore compare period ratios $T(L)/T(0.40)$ with $\sqrt{L/0.40}$, which holds at any pixel scale. A clip is readable when it passes a re-timing control and its sine fit reaches $R^2 \geq 0.80$.

4.2 CLOSURE-TERMINATED RETRIEVAL

The closure walk raises non-local recall more than fourfold over dense retrieval, from 0.105 to 0.449 (Table 1). It also leads on sufficiency@10, at 0.314 against 0.210 for dense retrieval and 0.178 for BM25. The lead holds at every noise level: 0.361 against 0.240 at L_0, and 0.263 against 0.201 at L_4. Precision and F_1 follow the same order, with F_1 of 0.227 against 0.164 (Appendix R). The ablations locate the gain. The symbol graph adds 0.179 non-local recall over dense retrieval. Reranking adds 0.051, and the closure walk a further 0.113. Every adjacent difference has a 95% document-level bootstrap interval above zero (Appendix P).

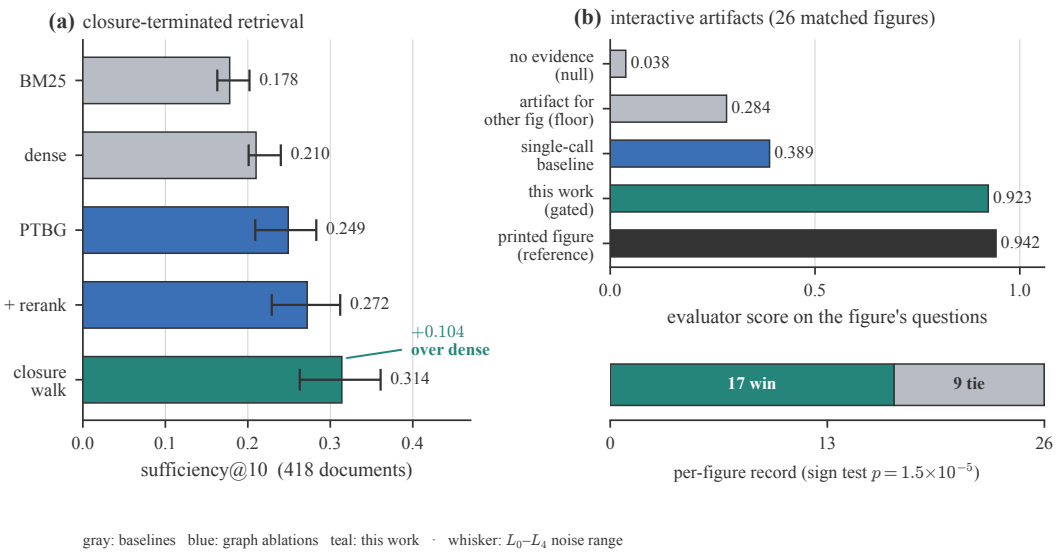

Figure 4: Retrieval and interactive-artifact results. (a) Pooled sufficiency@10 on 418 documents; whiskers span query-noise levels L_0–L_4. (b) Matched evaluation on 26 figures comparing the gated pipeline with single-call generation.

Table 2: Matched comparison against single-call generation on 26 figures with eight questions per figure. The printed figure and different-figure artifact provide reference and floor conditions. TQ subtracts the no-evidence null.

	artifact the evaluator was given	score	TQ	vs. baseline
methods	single-call baseline	0.389	+0.351	—
	this work (gated pipeline)	**0.923**	**+0.885**	**+0.534**
reference	printed figure (upper reference)	0.942	+0.904	—
	artifact for another figure (floor)	0.284	+0.245	—
	no evidence (null)	0.038	—	—

4.3 Interactive artifact evaluation

The gated artifacts score 0.923 on 26 figures, against 0.389 for single-call generation (Table 2, Figure 4b). This is within 0.02 of the printed figure's 0.942. Per figure, the pipeline wins 17, ties 9, and loses none (exact two-sided sign test, $p = 1.5 \times 10^{-5}$). The different-figure floor scores 0.284. The two pre-registered alternative thresholds keep 20 and 19 figures. Their margins are $+0.631$ and $+0.664$, again with no losses. The margin comes from what the artifacts expose. All 46 pipeline and 40 baseline artifacts execute and stay finite over their control sweeps. The pipeline's artifacts display the implemented relation in 96% of cases, against 0% for the baseline. They name the plotted quantity in 85% of cases, against 2%.

4.4 Human reader study

We measured how the artifacts affect human comprehension of textbook figures. Seven figures from the textbook (Moebs et al., 2016) formed the stimulus pool. Each reader completed four blocks, each on a different figure. Every block showed a section excerpt and the printed figure with its caption. Two blocks added a gate-passing artifact. One block added an artifact with one injected physics error. The fourth block showed the printed page alone. After each block, readers answered four multiple-choice questions written from the source section. Accuracy was highest with a gate-passing artifact beside the page: 0.611, against 0.554 for the page alone (Table 3). The table covers 236 answers from 23 readers. Readers flagged 5 of 13 blocks with an injected error, close to their 8 of 29 flags on gate-passing artifacts. Injected errors also did not carry into answers. Only 5 of 54

Table 3: Answer accuracy in the human reader study. Each reader saw every condition, each on a different figure. Intervals are 95% bootstrap intervals over readers. *Flagged as wrong* counts finished blocks in which the reader reported a contradiction with the section.

what the reader saw	accuracy	95% interval	answers	readers	flagged as wrong
printed page only	0.554	[0.386, 0.731]	56	14	3 / 12
page + gate-passing artifact	0.611	[0.485, 0.740]	126	20	8 / 29
page + artifact with injected error	0.537	[0.354, 0.725]	54	14	5 / 13

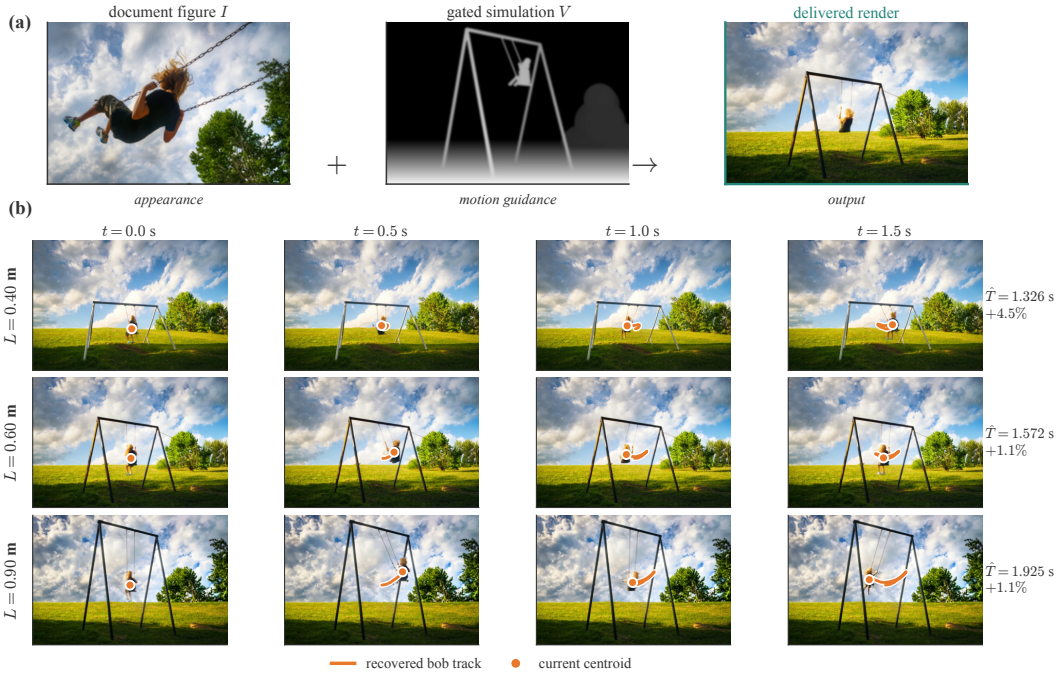

Figure 5: Photorealistic generation from document-derived physical conditioning. (a) The source photograph, a frame from the verified simulation used as motion conditioning, and the corresponding delivered render. (b) Delivered videos for $L = 0.40$, 0.60, and 0.90 m on a common time axis, with the trajectory recovered from the rendered frames.

answers picked the option matching the error, below the 0.130 base rate after gate-passing artifacts. Unaided readers do not reliably detect physics errors in an artifact, so verification has to be automatic. Appendix J gives the full design and further measures.

4.5 Photorealistic physical verification

The delivered videos keep the document's physics (Figure 5). At $L = 0.40/0.60/0.90$ m, the segmentation tracker measures periods of 1.326/1.572/1.925 s. The blob tracker measures 1.33/1.58/1.90 s, against closed-form values of 1.269/1.555/1.904 s from $T = 2\pi\sqrt{L/g}$. The deviations are 4.5%, 1.1%, and 1.1%, and the trackers agree to within 0.025 s. All three renders pass Equation (7).

The physical conditioning is what produces this behavior. Without it, the same backbone yields no recoverable period in any of nine clips. These clips contain motion, but not the simulated trajectory (Appendix L). Against general-purpose video systems, the conditioned branch has a scale error of 4.8% and a correlation of $+0.999$ with $\sqrt{L}$ (Table 4). Veo 3.1 reaches 23.5% and $+0.18$ on its five readable clips of eight. Runway Gen-4.5 yields two readable clips of nine.

Table 4: General-purpose video systems and the conditioned branch on the same target behavior. *Readable* clips pass the re-timing control and reach $R^2 \geq 0.80$. *Scale error* is the mean deviation of $T(L)/T(0.40)$ from $\sqrt{L/0.40}$.

	system	readable	scale error	r vs $\sqrt{L}$
prompt only	Veo 3.1	5 / 8	23.5%	$+0.18$
	Runway Gen-4.5	2 / 9	not reported	not reported
this work	**conditioned branch**	**3 / 3**	$\mathbf{4.8\%}$	$\mathbf{+0.999}$

4.6 Release of the physical conditioning

The simulation decides the motion, and the prompt decides only the appearance (Figure 3c). One depth control of a thrown ball is paired with three prompts that request different motions. Without the control, the prompt sets the path, and sideways travel ranges from 0.05 to 0.97. The range attributable to the prompt is 0.875, against 0.252 between seeds. With the control on all 50 steps, every prompt reproduces the control's arc at a travel of 0.87. The range attributable to the prompt falls to 0.001. With the control on only the first 10 steps, the path holds the arc through the flight and departs at landing.

The release sweep locates when the trajectory commits (Figure 3b). Releasing at step 8 leaves the airborne flight intact and changes only the motion after ground contact. Over 16 seeds, airborne horizontal retention drops by 0.003 against never releasing (95% CI 0.001 to 0.004). After contact, horizontal motion runs opposite to the simulation on 9 of 16 seeds, against 0 of 16 without release. From step 14 onward, across-frame retention stays at 0.98 or above in every scene. The vertical component commits later and reaches 0.89 at step 20. The flight is fixed early in sampling, and contact is the last phase to commit. Releasing the simulation lets appearance follow the photograph. With a release, 6 of 6 of the hardest scenes reached photoreal output, against 2 of 8 without (Appendix U).

5 Conclusion

We introduced a framework for transforming static figures in technical documents into executable, physically grounded artifacts. The central idea is to treat the document as the source of the physical specification: retrieval continues until the relevant equations form a closed system, the recovered system is represented as a physics scenario card, and the same card is used to generate and verify both interactive and photorealistic outputs.

The experiments evaluate this chain at three levels. Closure-terminated retrieval improves recovery of the equation context required to instantiate a physical system. The resulting interactive artifacts preserve enough of the source figure's information to answer figure-specific questions substantially more accurately than single-call generation. For photorealistic generation, independently measured motion in the delivered videos follows the document-derived closed form across multiple control values.

These results support three conclusions. Physical closure is an effective stopping rule for retrieval. It recovers the non-local context that similarity search misses. Tests derived from the document make generated artifacts expose the physics that a reader's questions require. Simulation conditioning carries the document's dynamics into a video model, and the delivered frames measure back to the closed form. One physics scenario card links all three stages, so every output stays traceable to the equations on its page.

The framework applies wherever a document states the equations behind its figures. Its pre-render gates already cover seven scenario families, from pendulums to collisions (Appendix G). Future work will complete the human study, extend output measurement beyond the pendulum, and improve retrieval sufficiency.

AI USE STATEMENT

The method uses a language model as a component. It extracts symbols, judges bindings, proposes relations between equations, resolves objectives, and synthesizes code. Symbolic re-derivation checks every proposed relation. Execution and PSC-derived tests check every generated artifact. The model is used for inference only, without task-specific training or fine-tuning.

We used generative AI tools for drafting and editing the manuscript, preparing figures, and routine code assistance. Figure preparation included reading plotted values from vector figures and writing plotting code. We checked the extracted values against the numbers stated in the source text. The authors determined the research questions, experimental designs, measurements, and conclusions. Every reported number comes from the experiment records. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The evaluation corpus is openly licensed text and public preprints. The human reader study (§4.4) recorded no names or contact details. Answers were stored under a random token. Participants gave informed consent, were paid, and could stop at any time. Some participants were colleagues of the authors. The debrief disclosed the injected errors. We intend the system to support instruction, not to replace it. Because we generate the artifact rather than authoring it, we embed provenance back to the originating relations and page in every artifact, so a reader or instructor can audit what a figure came from. Generative video can depict a physical process convincingly while violating it, which is why §3.3 requires post-hoc measurement by two independent instruments.

REPRODUCIBILITY STATEMENT

The corpus is 448 physics documents built from an openly licensed university physics text (Moebs et al., 2016) and arXiv preprints; §4.1 states the 418-document analysis set, the exclusion that produced it, the query construction, the five noise levels, the counts (5,707 equations, 28,535 queries per configuration, 142,675 scored rows) and the document-level bootstrap. The matched artifact comparison states its construction, evaluator interface, pre-registered exclusion rules and both alternative thresholds. Gate tolerances are given per scenario family in Appendix G; sampler settings (N=40, timestep shift 5.0, β=0.875, r=16, guidance weights) are in Appendix N; Appendix U reports the release-index study. Appendix Q describes the measurement implementation and its self-test. Appendix O lists the full technology stack, including the permissively licensed video backbone and the segmentation instruments. We will release source and configuration.

REFERENCES

Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. Videophy: Evaluating physical commonsense for video generation. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 102075–102121, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/fce2d8a485746f76aac7b5650db2679d-Paper-Conference.pdf.

Yossi Ben-Zion, Roi Einhorn Zarzecki, Joshua Glazer, and Noah D Finkelstein. Leveraging ai for rapid generation of physics simulations in education: Building your own virtual lab. *The Physics Teacher*, 63(6):424–427, 2025.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets, 2023. URL https://arxiv.org/abs/2311.15127.

Xinyun Chen, Maxwell Lin, Nathanael Schaerli, and Denny Zhou. Teaching large language models to self-debug. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun

(eds.), *International Conference on Learning Representations*, volume 2024, pp. 8746–8825, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/2460396f2d0d421885997dd1612ac56b-Paper-Conference.pdf.

Yanzhe Chen, Kevin Qinghong Lin, and Mike Zheng Shou. Code2video: A code-centric paradigm for educational video generation, 2025. URL https://arxiv.org/abs/2510.01174.

Ton De Jong. Technological advances in inquiry learning. *Science*, 312(5773):532–533, 2006.

Ton De Jong, Marcia C Linn, and Zacharias C Zacharia. Physical and virtual laboratories in science and engineering education. *Science*, 340(6130):305–308, 2013.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. From local to global: A graph rag approach to query-focused summarization, 2025. URL https://arxiv.org/abs/2404.16130.

Emily First, Markus N Rabe, Talia Ringer, and Yuriy Brun. Baldur: Whole-proof generation and repair with large language models. In *Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pp. 1229–1241, 2023.

Yuze He, Jie Yang, and Yonghe Zheng. Creating your own interactive simulation using an llm: A simple pendulum case study. *The Physics Teacher*, 63(8):702–703, 2025.

Andrew Head, Kyle Lo, Dongyeop Kang, Raymond Fok, Sam Skjonsberg, Daniel S Weld, and Marti A Hearst. Augmenting scientific papers with just-in-time, position-sensitive definitions of terms and symbols. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–18, 2021.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022. URL https://arxiv.org/abs/2207.12598.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David Fleet. Video diffusion models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 8633–8646. Curran Associates, Inc., 2022. doi: 10.52202/068431-0628. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/39235c56aef13fb05a6adc95eb9d8d66-Paper-Conference.pdf.

Tim N Höffler and Detlev Leutner. Instructional animation versus static pictures: A meta-analysis. *Learning and instruction*, 17(6):722–738, 2007.

Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8153–8163, 2024.

Enkelejda Kasneci, Kathrin Sessler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, Stephan Krusche, Gitta Kutyniok, Tilman Michaeli, Claudia Nerdel, Jürgen Pfeffer, Oleksandra Poquet, Michael Sailer, Albrecht Schmidt, Tina Seidel, Matthias Stadler, Jochen Weller, Jochen Kuhn, and Gjergji Kasneci. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274, April 2023. ISSN 1041-6080. doi: 10.1016/j.lindif.2023.102274. URL http://dx.doi.org/10.1016/j.lindif.2023.102274.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4015–4026, October 2023.

Max Ku, Cheuk Hei Chong, Jonathan Leung, Krish Shah, Alvin Yu, and Wenhu Chen. TheoremExplainAgent: Towards video-based multimodal explanations for LLM theorem understanding. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*

(Volume 1: Long Papers), pp. 6663–6684, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.332. URL https://aclanthology.org/2025.acl-long.332/.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf.

Behrooz Mansouri, Anurag Agarwal, Douglas W. Oard, and Richard Zanibbi. Advancing math-aware search: The arqmath-2 lab at clef 2021. In Djoerd Hiemstra, Marie-Francine Moens, Josiane Mothe, Raffaele Perego, Martin Potthast, and Fabrizio Sebastiani (eds.), *Advances in Information Retrieval*, pp. 631–638, Cham, 2021. Springer International Publishing. ISBN 978-3-030-72240-1.

Fanqing Meng, Jiaqi Liao, Xinyu Tan, Quanfeng Lu, Wenqi Shao, Kaipeng Zhang, Yu Cheng, Dianqi Li, and Ping Luo. Towards world simulator: Crafting physical commonsense-based benchmark for video generation. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 43781–43806. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/meng25c.html.

Aaron Meurer, Christopher P. Smith, Mateusz Paprocki, Ondřej Čertík, Sergey B. Kirpichev, Matthew Rocklin, Amit Kumar, Sergiu Ivanov, Jason K. Moore, Sartaj Singh, Thilina Rathnayake, Sean Vig, Brian E. Granger, Richard P. Muller, Francesco Bonazzi, Harsh Gupta, Shivam Vats, Fredrik Johansson, Fabian Pedregosa, Matthew J. Curry, Andy R. Terrel, Štěpán Roučka, Ashutosh Saboo, Isuru Fernando, Sumith Kulal, Robert Cimrman, and Anthony Scopatz. Sympy: symbolic computing in python. *PeerJ Computer Science*, 3:e103, January 2017. ISSN 2376-5992. doi: 10.7717/peerj-cs.103. URL https://doi.org/10.7717/peerj-cs.103.

William Moebs, Samuel J. Ling, and Jeff Sanny. *University Physics Volume 1*. OpenStax, Houston, Texas, 2016. URL https://openstax.org/books/university-physics-volume-1/pages/1-introduction.

Theo X. Olausson, Jeevana Priya Inala, Chenglong Wang, Jianfeng Gao, and Armando Solar-Lezama. Is self-repair a silver bullet for code generation? In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 36545–36593, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/9ddc141bdbf9d1db510cefff56c586ad-Paper-Conference.pdf.

Samarth P, Vyoman Jain, Shiva Golugula, and Motamarri Sai Sathvik. Manimator: Transforming research papers into visual explanations, 2025. URL https://arxiv.org/abs/2507.14306.

M. Raissi, P. Perdikaris, and G.E. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, February 2019. ISSN 0021-9991. doi: 10.1016/j.jcp.2018.10.045. URL http://dx.doi.org/10.1016/j.jcp.2018.10.045.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollar, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 28085–28128, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/45c1f6a8cbf2da59ebf2c802b4f742cd-Paper-Conference.pdf.

Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5026–5033. IEEE, October 2012. doi: 10.1109/iros.2012.6386109. URL http://dx.doi.org/10.1109/IROS.2012.6386109.

Trieu H. Trinh, Yuhuai Wu, Quoc V. Le, He He, and Thang Luong. Solving olympiad geometry without human demonstrations. *Nature*, 625(7995):476–482, January 2024. ISSN 1476-4687. doi: 10.1038/s41586-023-06747-5. URL http://dx.doi.org/10.1038/s41586-023-06747-5.

Xiaozao Wang, Zhewei Wang, and Hongyi Wen. *Evaluating Interactivity: Toward Automated Assessment of AI-Generated Explorable Explanations*, pp. 124–138. Springer Nature Switzerland, 2026a. ISBN 9783032297556. doi: 10.1007/978-3-032-29755-6_9. URL http://dx.doi.org/10.1007/978-3-032-29755-6_9.

Yanan Wang, Renxi Wang, Yongxin Wang, Xuezhi Liang, Fajri Koto, Timothy Baldwin, Xiaodan Liang, and Haonan Li. Training and benchmarking code generation for physics-inspired animations, 2026b. URL https://arxiv.org/abs/2602.10840.

Carl E. Wieman, Wendy K. Adams, and Katherine K. Perkins. Phet: Simulations that enhance learning. *Science*, 322(5902):682–683, October 2008. ISSN 1095-9203. doi: 10.1126/science.1161948. URL http://dx.doi.org/10.1126/science.1161948.

Anna Wong, Wayne Leahy, Nadine Marcus, and John Sweller. Cognitive load theory, the transient information effect and e-learning. *Learning and Instruction*, 22(6):449–457, December 2012. ISSN 0959-4752. doi: 10.1016/j.learninstruc.2012.05.004. URL http://dx.doi.org/10.1016/j.learninstruc.2012.05.004.

Richard Zanibbi and Dorothea Blostein. Recognition and retrieval of mathematical expressions. *International Journal on Document Analysis and Recognition (IJDAR)*, 15(4):331–357, Dec 2012. ISSN 1433-2825. doi: 10.1007/s10032-011-0174-4. URL https://doi.org/10.1007/s10032-011-0174-4.

Richard Zanibbi, Behrooz Mansouri, and Anurag Agarwal. Mathematical information retrieval: Search and question answering. *Foundations and Trends® in Information Retrieval*, 19(1-2):1–190, January 2025. ISSN 1554-0677. doi: 10.1561/1500000095. URL http://dx.doi.org/10.1561/1500000095.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3836–3847, October 2023.

A Worked example: pendulum

We use the pendulum section of an openly licensed university physics text (Moebs et al., 2016) to instantiate the objects defined in Section 3.1. All counts and numerical values below are read from the corresponding pipeline record.

Relation verification (Gate α). The language model proposed 83 inter-equation relations. Independent symbolic re-derivation admitted 6, rejected 32 as contradictory, and classified 45 as unverifiable. Only the 6 verified relations are added to the document graph.

Closure (Gate β). The stated objective resolves to the seed $\omega = \sqrt{g/L}$ (eq0004) among the section's 20 equations. The retrieval walk terminates with $|S| = 6$ after reaching 24 of the graph's 63 vertices (Appendix D). Two symbols remain free, $U(S) = \{m, x\}$; both are admissible parameters. The final system contains $R(S) = 7$ independent relations for $n(S) = 2$ unknowns and is therefore marked `closed_parametric` after the consistency, dimensional, and well-posedness checks succeed.

Dynamics and invariant of record. The recovered law of motion is

$$\ddot{\theta} = -\frac{g}{L}\sin\theta \xrightarrow[\theta_0<15^\circ]{\text{regime}} \ddot{\theta} \simeq -\frac{g}{L}\theta, \tag{8}$$

with the small-angle condition retained as an executable regime guard. The conserved quantity used by the generated-artifact test is $E = \frac{1}{2}mL^2\dot{\theta}^2 + mgL(1-\cos\theta)$, and the invariant of record used by the photoreal branch is $q = T = 2\pi\sqrt{L/g}$.

Visual mapping. For the printed photograph, the solved angle $\theta(t)$ is mapped to the seat position $p_{\text{seat}}(t) = p_{\text{pivot}} + L(\sin\theta(t), -\cos\theta(t))$ while the rider remains upright in world coordinates. The resulting visual rig is kinematic; the pre-render probe therefore measures the motion independently of the rig construction.

B Cross-equation symbol binding

Pass 1 emits one record per (equation, symbol) occurrence, $r = (e_r, s_r, d_r, \gamma_r)$: host equation, surface form, dimension vector $d_r \in \mathbb{Q}^7$ over the SI base dimensions, and the gloss extracted from the surrounding prose. A pair of records binds only if

$$r \bowtie r' \;\equiv\; \underbrace{[\,s_r = s_{r'}\,] \wedge [\,d_r = d_{r'}\,]}_{\text{deterministic bucket}} \;\wedge\; \underbrace{\text{same}(\gamma_r, \gamma_{r'})}_{\text{model judgment, on the glosses}}, \tag{9}$$

and identity is the union–find closure of the accepted pairs,

$$[r] \;=\; \{\, r' \,:\, r \bowtie^{*} r' \,\}, \qquad V_{\text{var}} \;=\; \{\,[r]\,\}. \tag{10}$$

The first conjunct of (9) splits homonyms before any model is consulted; the second is judged on what the prose says each symbol *is*, not on the letter. Definitions existing only in the prose attach to $[r]$, and a derivative or component symbol c of a base quantity p propagates definitions upward,

$$[c] \prec [p] \;\wedge\; e \xrightarrow{\text{defines}} [c] \;\Longrightarrow\; e \xrightarrow{\text{defines}} [p], \tag{11}$$

so closure later treats θ as provided once the equation for $\dot{\theta}$ is gathered.

Structural edges of the document graph follow mechanically from the records and the partition,

$$e \xrightarrow{\text{references}} [r] \;\text{ for every } r \text{ with } e_r = e, \qquad e \xrightarrow{\text{defines}} [r] \;\Longleftrightarrow\; \text{lhs}(e) \in [r], \tag{12}$$

plus *constrained-by* edges from a variable to the validity regime the prose attaches to it. The worked section has 96 structural edges (73 references / 19 defines / 4 constrained-by) and 6 gate-α survivors of the 102 edges in total.

C Closure predicates and derived verification tests

The main paper summarizes closure with Equation (2). This appendix specifies the individual predicates and the tests derived from the resulting PSC.

Consistency. For an equation set S, $\text{Consistent}(S)$ requires the symbolic relations to admit a common solution. This check is necessary when $R(S) > n(S)$, where relation count alone can classify an inconsistent overdetermined system as numerically sufficient.

Dimensional consistency. $\text{DimOK}(S)$ reduces every physical quantity to its SI-base-dimension vector and requires all additive terms in an equation residual to have the same dimension. The check is therefore performed by units algebra rather than by string matching on symbols.

Dynamical well-posedness. $\text{WellPosed}(S)$ rejects systems that are algebraically determinate but cannot define the requested dynamics. In the current implementation this includes missing laws of motion for evolving state variables, differential–algebraic forms that cannot be reduced to the required state update, and operators that appear without a definition.

Conservation tests. Equation (3) is evaluated on the executed artifact rather than on the symbolic equations. The tolerance ϵ_E is stored in the PSC, so this test checks whether the implementation produced by the generator preserves the conserved quantity recovered from the document.

Limiting cases and validity regimes. Symbolic limits stated by the document are converted to executable limiting-case tests. Validity regimes are compiled to predicates that return valid, warning, or invalid. These guards restrict the parameter values exposed by the artifact and prevent the interface from presenting an out-of-regime configuration as physically valid.

D THE DOCUMENT GRAPH AND THE STATE OF THE WALK

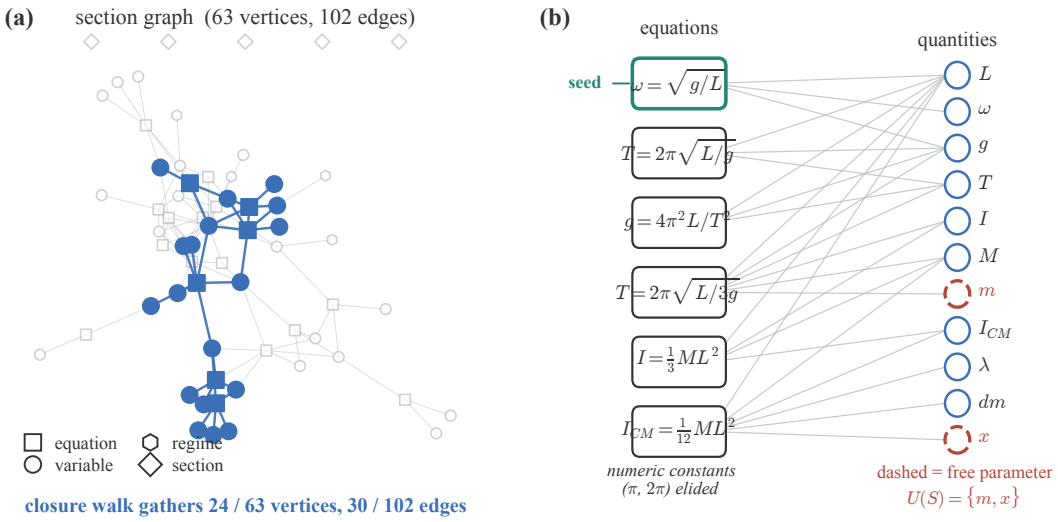

Figure 6: The document graph of the worked section and the closure walk. (a) The graph has 63 vertices; shape encodes vertex type. The walk gathered the highlighted 24 vertices and 30 edges before $\mathrm{Closed}(S)$ held. (b) The gathered subgraph maps six equations to their shared quantities. The seed $\omega = \sqrt{g/L}$ and the free parameters $\{m, x\}$ are marked.

Table 5 gives the composition of the graph in Figure 6 and the state of the closure walk at each depth on the worked section. Every value is read from the run's own record.

Table 5: The graph of Figure 6 by vertex type and relation, and the state of the walk at each depth.

vertices	63
Equation / Variable / Regime / Section	20 / 34 / 4 / 5
edges	102
structural: references / defines / constrained_by	73 / 19 / 4
gate-α survivors: defines_for / rearrangement / combination	4 / 1 / 1
depth 0 $S = \{$eq0004, eq0005$\}$	
undefined $U(S)$	$\{L, g\}$
$R(S)$ against $n(S)$	2 vs 2
closure status	not closed
depth 1 + eq0007, eq0022, eq0023, eq0033	
undefined $U(S)$, both admissible parameters	$\{m, x\}$
$R(S)$ against $n(S)$	7 vs 2
closure status	**closed_parametric** (physically closed)

The retrieved system, written in the document's own notation with its equation identifiers:

$$\omega = \sqrt{g/L}, \qquad \text{(eq0004, the seed)} \tag{13}$$

$$T = 2\pi\sqrt{L/g}, \qquad \text{(eq0005)} \tag{14}$$

$$g = 4\pi^2 L/T^2, \qquad \text{(eq0007)} \tag{15}$$

$$T = 2\pi\sqrt{I/(mgL)} = 2\pi\sqrt{L/3g}, \qquad \text{(eq0023)} \tag{16}$$

$$I = I_{CM} + \tfrac{1}{4}L^2 M = \tfrac{1}{3}ML^2, \qquad \text{(eq0022)} \tag{17}$$

$$I_{CM} = \int x^2\, dm = \tfrac{1}{12}ML^2. \qquad \text{(eq0033)} \tag{18}$$

E PHYSICS SCENARIO CARD

The PSC stores the recovered physical system in a typed representation:

$$\mathcal{C} = (X, P, K; \Sigma; \mathcal{R}, \mathcal{I}, \mathcal{L}, \mathcal{D}; o, \pi).$$

X, P, and K contain state variables, reader-controlled parameters, and constants together with units, dimension vectors, admissible ranges, glosses, source spans, and code-safe identifiers. Σ contains the retrieved relations verbatim. $\mathcal{R}$ stores validity regimes, $\mathcal{I}$ conserved quantities and measurable invariants, $\mathcal{L}$ limiting cases, and $\mathcal{D}$ dimensional constraints. The objective o records the source figure's stated task, and π stores provenance including document location, equation identifiers, figure coordinates, and the closure record.

Table 6 summarizes which fields are read by each downstream stage. In particular, the code generator receives the PSC and the derived tests rather than the source prose.

Table 6: PSC fields consumed by each downstream stage.

stage	reads
assertion derivation	$\mathcal{D}, \mathcal{I}, \mathcal{L}, \mathcal{R} \rightarrow$ executable tests
artifact generation (gate γ)	X, P, K, Σ, o + the derived tests
regime guards in the artifact	$\mathcal{R}$, compiled to valid / warning / invalid
rig construction + gate δ	Σ for trajectory, P for geometry, $\mathcal{I}$ for probe
render conditioning	source figure from π; render family indexed by P
measurement (gate ε)	invariant from Σ and $\mathcal{I}$

F THE INTEGRATOR IN THE ACCEPTED ARTIFACT

The in-situ artifact of Figure 8 integrates (8) as a first-order system $y = (\theta, \dot{\theta})$ with $\dot{y} = f(y) = (\dot{\theta},\ -(g/L)\sin\theta)$ by fixed-step RK4 at $h = 1/480$ s,

$$y_{k+1} = y_k + \tfrac{h}{6}\left(k_1 + 2k_2 + 2k_3 + k_4\right), \tag{19}$$

with the usual stage evaluations, sub-stepped so the display rate is independent of the integration step. As noted in Section 3.2, the integrator is the generator's choice and not a prescription of the method: acceptance is decided by the derived tests.

G SCENARIO FAMILIES WITH A PRE-RENDER GATE

Table 7 lists the seven scenario families whose pre-render gate (§3.3) is implemented, with the invariant each gate measures and its tolerance. Tolerances run from 1.0% for the pendulum and orbital periods to 3.0% for centripetal acceleration. The collision family also carries the negative control, a near-inelastic contact configured to fail the gate.

Table 7: Scenario families with a gate implemented. The tolerance is set per family from the numerical conditioning of the invariant, not tuned per scene. The collision family carries the negative control: the tested near-inelastic contact is configured to fail it.

family	invariant measured	tolerance τ
pendulum	period $2\pi\sqrt{L/g}$	1.0%
orbit	orbital period	1.0%
spring	period $2\pi\sqrt{m/k}$	1.5%
incline	acceleration $g\sin\varphi$	1.5%
projectile	range and apex	1.5%
collision	linear momentum and restitution	2.0%
centripetal	centripetal acceleration v^2/r	3.0%

H MEASURES OF RECORD

Table 8 defines the five measures reported in the experiments and the comparator each is read against. Sufficiency@k and non-local recall score retrieval (§4.2), TQ scores the interactive artifact (§4.3), and period error and drift score the delivered video (§4.5). No measure is reported without its comparator.

Table 8: Measures of record. Each is reported against a stated comparator rather than in isolation, and no measure below is defined in terms of the component it scores.

measure	what it counts	compared against
sufficiency@k	queries whose k returned parts contain everything the equation requires	lexical (BM25) and dense retrieval baselines
non-local recall	required parts recovered that are remote from the equation in reading order	the same baselines, pooled over queries
TQ	evaluator score on the figure's questions, with the no-evidence null subtracted	a single-call generation baseline on the same figure
period error	fitted period of the delivered motion	the closed form derived from the retrieved system
drift	largest relative excursion of a conserved quantity over the clip	the same quantity under ideal tracking (the floor)

I ADDITIONAL INTERACTIVE-ARTIFACT ANALYSES

The main paper reports the primary matched comparison on 26 figures. Two pre-registered alternative exclusion thresholds give the same qualitative result. Requiring the different-figure floor artifact to score strictly below the printed figure leaves 20 figures and yields a $+0.631$ difference relative to the single-call baseline. Requiring the floor condition to remain at the no-evidence null leaves 19 figures and yields a $+0.664$ difference. Neither subset contains a loss for the proposed method.

We also separate basic functionality from queryability. Across 46 cached pipeline artifacts and 40 baseline artifacts, every artifact executed and remained finite over the control sweep. A reported output changed for 73% of controls in the proposed artifacts and 71% in the baseline artifacts. The proposed artifacts display the implemented physical relation in 96% of cases and identify the plotted quantity in 85%, compared with 0% and 2% for the baseline. These checks support the interpretation that the primary difference is not simply whether the generated artifact runs, but whether it exposes the physical information required by the figure-specific questions.

J HUMAN READER STUDY

This appendix gives the design of the study in Section 4.4 and its further measures. The design is within reader: every reader sees every condition.

Participants. Sixty-eight adults gave consent over four days. Twenty-three readers answered at least one block, and 11 finished all four blocks. Twenty-seven participants stopped at the background

questions, and 30 stopped mid-session. The analysis includes every answered block, whether or not the reader finished. Participants were paid, and some were colleagues of the authors. Forty-one participants answered two background questions on university-level courses. Of these, 28 had taken mechanics and waves, 4 mechanics only, 7 waves only, and 2 neither. Readers who finished took a median of 27 minutes.

Stimuli. The pool contains seven figures from the openly licensed textbook (Moebs et al., 2016) (Table 9). For each figure, the reader saw a frozen excerpt of 179 to 245 words. The excerpt states the governing relation. The reader also saw the printed figure and its caption. We selected the gate-passing artifacts by hand from the 164 artifacts that passed every gate. Thirty of their 36 controls opened at the midpoint of a generic fallback range. We therefore replaced every opening value before the study. A language model proposed each value, and we accepted it only after running the artifact. In two artifacts, some controls have no effect. The energy artifact has 0 of 2 live controls, and the simple pendulum has 2 of 5.

Each injected-error artifact carries exactly one violation. The violation comes from a verified mutant of the same gate-passing artifact. Three error classes were served. Class A flips the sign of the damping term, so the motion grows instead of decaying. Class B multiplies the restoring term by four, so the motion runs at the wrong rate. Class C changes one sign or coefficient in the algebraic relation the artifact computes.

Table 9: Figure pool of the reader study. Columns P, S and M count finished blocks per condition. P is a gate-passing artifact, S the printed page only, and M an injected-error artifact.

textbook section	source relation	error classes	P	S	M
4.6 Carnot cycle	$Q_h = nRT_h \ln(V_N/V_M)$	C	4	2	1
11.4 Charged particle in B	$T = 2\pi m/(qB)$	C	6	2	0
15.3 Energy in SHM	$E = \frac{1}{2}kx^2 + \frac{1}{2}mv^2$	C	4	2	1
15.5 Simple pendulum	$\ddot{\theta} = -(g/L)\,\theta$	B	2	3	1
15.5 Torsional pendulum	$\tau = -\kappa\theta$	C	1	2	5
15.6 Damped oscillations	$ma = -bv - kx$	A, B	7	1	2
15.7 Forced oscillations	$m\ddot{x} = -kx - b\dot{x} + F_0 \sin \omega t$	A, B, C	5	0	3
total			29	12	13

Procedure. The study ran in one sitting with no follow-up.

(i) *Consent.* The text described the task, the recorded data, and anonymity. It stated that some artifacts contain a deliberately introduced physics error.

(ii) *Introduction.* A short text described the four blocks and the controls.

(iii) *Background.* Two questions asked about university courses in mechanics and in waves.

(iv) *Four blocks.* Each block had a study screen, four questions, and a detection probe.

(v) *Awareness check.* Readers guessed which display the researchers built themselves.

(vi) *Debrief.* The text explained the pipeline, the injected errors, and the measures.

The study screen showed the excerpt and the printed figure. Artifact blocks also showed the artifact, with each control labelled by its meaning. The artifact instruction suggested a couple of minutes of use. The printed-page instruction suggested ten minutes. The questions opened after 60 seconds. The page and the artifact stayed on screen while the reader answered. Readers received no feedback. Option order was shuffled per reader and item. The detection probe had three parts. It asked whether the display contradicted the section, on a three-point scale. It then asked which quantity seemed wrong, and the reader's confidence on a five-point scale.

Allocation. A searched plan assigns 28 reader groups to the seven figures. Each group sees four distinct figures under conditions P, P, S and M. The order of conditions is specific to each group. Across the plan, each figure appears eight times as P, four times as S, and four times as M. Each block position carries 14 P, 7 S and 7 M blocks. The plan provides 84 slots, three per group, filled in a fixed order. The error class of each M block is fixed at allocation. Within a figure and version, the server serves the least-used instance. The server verifies a hash of the frozen item set at consent.

Items. Each figure has four multiple-choice items, so each reader answers 16 of 28. The items were written from the published sections, not from any generated artifact. Each item targets one of the figure's three learning objectives. Each item has a key and three distractors. Each distractor is tagged with one error class and a named misconception. One item set therefore serves every version of a figure. For example, one damped-oscillation item reads: "Over the next few swings the swing will". The key is "fall by about the same fraction each time". Two distractors are "grow from swing to swing" (A) and "collapse to nothing within a single swing" (B). The third is "stay the same size indefinitely" (C).

Measures and analysis. Accuracy is the fraction of answers that match the key. Error transmission is the fraction of M answers whose chosen distractor matches the served error class. Its reference is the rate of choosing any tagged distractor after a P artifact, divided by three. Detection is any judgement other than "No, it behaves as the section says". Engagement is focused time on the study screen and the number of control moves. Accuracy intervals use 4,000 bootstrap resamples over readers. All answers of a resampled reader move together. Intervals on proportions of blocks or answers are Wilson intervals. The results are descriptive and come from an unplanned interim look at 11 finished readers. The registered analysis is a mixed-effects logistic model. It has random intercepts for reader, item and instance, and has not yet been fitted.

Further measures. Table 10 reports the within-reader difference, error transmission, detection, and engagement. Over 13 readers who saw both conditions, accuracy after P exceeds accuracy after S by $+0.077$. After an M artifact, 5 of 54 answers chose the distractor that matches the injected error. By class, these counts are 4 of 12 (A), 0 of 4 (B), and 1 of 38 (C). Readers moved no control in 16 of 29 P blocks and 7 of 13 M blocks. In the awareness check, 3 of 11 readers chose "I have no idea".

Table 10: Further measures of the reader study. Intervals are 95%: bootstrap over readers for the accuracy difference, Wilson otherwise.

measure	condition	value
accuracy difference, within reader (13 readers)	P − S	$+0.077$ $[-0.078, +0.229]$
error transmission (54 answers)	M	0.093 $[0.040, 0.199]$
	P (reference)	0.130
flagged as contradicting the section	M (5 of 13 blocks)	0.385 $[0.177, 0.645]$
	P (8 of 29 blocks)	0.276 $[0.147, 0.457]$
	S (3 of 12 blocks)	0.250 $[0.089, 0.532]$
mean detection confidence (1–5)	M / P / S	2.9 / 2.7 / 2.7
focused time on study screen, median	P / S / M	142 s / 190 s / 196 s
blocks with no control moved	P / M	16 of 29 / 7 of 13
item latency, median	P / S / M	26 s / 28 s / 41 s

K Delivered periods at three control values

Table 11 lists the measured periods against the closed form, and Figure 7 plots the two trackers' recovered tracks with the fitted and theory-period sinusoids.

Table 11: Measured period of the delivered footage against the closed form obtained from the document, for three separately simulated and separately rendered control values. The two trackers are separately implemented and use distinct tracking procedures; their maximum disagreement is 0.025 s.

	measured		reference	deviation
control value	tracker A (segmentation)	tracker B (blob)	closed form	(tracker A)
0.40 m	1.326 s	1.33 s	1.269 s	$+4.5\%$
0.60 m	1.572 s	1.58 s	1.555 s	$+1.1\%$
0.90 m	1.925 s	1.90 s	1.904 s	$+1.1\%$

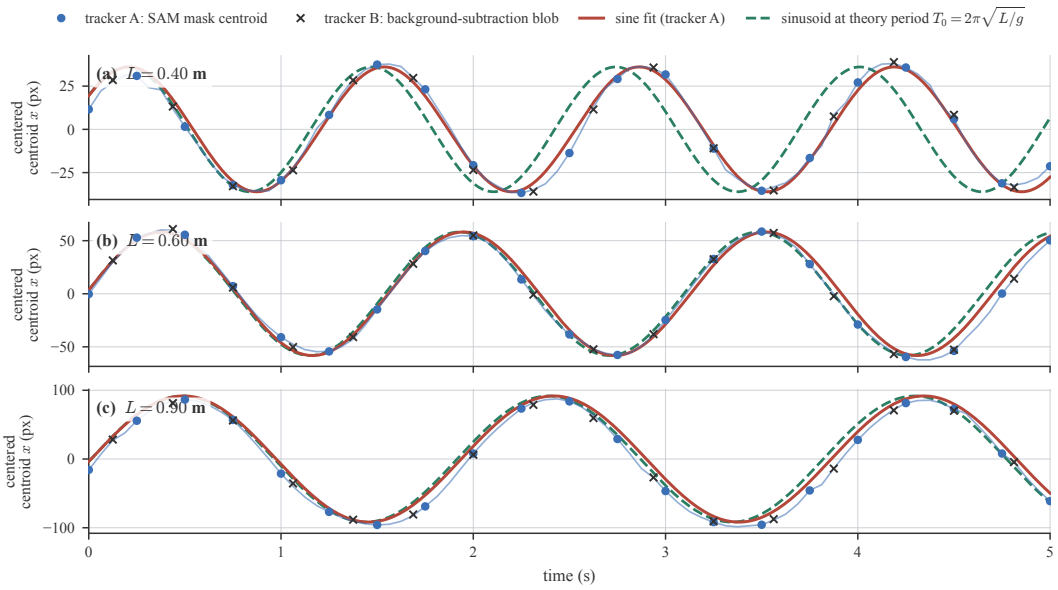

Figure 7: Gate ε on three separately rendered control values, measured by two independent trackers on the delivered frames: SAM mask centroids (blue points) and a background-subtraction blob (dark crosses). Red is the sine fit; green dashed is $T = 2\pi\sqrt{L/g}$ from the retrieved system. The fitted periods lie within $4.5/1.1/1.1\%$ of the closed form, the largest deviation at $L = 0.40$ m.

L UNCONDITIONED VIDEO-GENERATION DIAGNOSTIC

Table 12 gives the unconditioned diagnostic cited in §4.5. The unconditioned clips carry as much motion energy as the known swing (1.96–2.54 against 1.82), but their traveled distance is only 1.2–1.9% against 60.0%, and none of the nine yields a recoverable period. Without physical conditioning the backbone therefore produces motion, but not the displacement of a swinging pendulum.

Table 12: The backbone with no physical conditioning, nine clips (three control values $\times$ three seeds), measured by the same tracker as Table 11. The two reference rows are footage with a known answer, measured by the same code in the same run. Motion energy and moving-region size match the reference; traveled distance does not.

	clips	motion (grey)	moving region	distance traveled
reference	known 1.60 s swing	1.82	0.0104	60.0%
	known static scene	0.00	0.0000	—
unconditioned	prompted $L = 0.40$ m	2.07	0.0054	1.2%
	prompted $L = 0.60$ m	1.96	0.0044	1.9%
	prompted $L = 0.90$ m	2.54	0.0058	1.7%
	period recovered		**none, 0 of 9 clips**	

M MEASUREMENT ERROR AND CONSERVED-QUANTITY DRIFT

On clips carrying no injected error, the drift measured on the generator's own arrays is at most 0.0025, while the same quantity recovered through the tracker reaches 0.159. Under these tracking conditions a conservation figure measured from delivered frames is dominated by the instrument, and applying a gate to it directly could reject simulations that pass the simulation-side conservation check of (3). Applying the gate to the simulation, and using the tracked measurement only as an independent post-hoc check, is what this separation supports.

In the injected-energy condition tested, the tracker-measured quantity nevertheless remained distinguishable: adding 50% additional energy to the driving simulation raised the tracker-measured drift by approximately 46-fold relative to the unperturbed clips, outside the range occupied by tracking error alone on those clips. The intermediate row bounds the operating range — 5% pixel noise alone produces a measured drift of 0.462, which is within the range a genuine violation would occupy, so a measurement taken from noisy frames at that magnitude does not by itself distinguish the two.

Table 13: Drift of the conserved quantity on generated oscillator clips, scored twice: on the generator's own arrays (the value obtained under ideal tracking) and on the delivered frames through the separately implemented tracker. The last row is a negative control in which the driving simulation is deliberately given additional energy.

clips	floor (max)	tracked (median)	tracked (max)
as generated	0.0025	0.028	0.159
with 5% pixel noise	0.0025	0.462	0.754
with 50% energy injected	0.0025	1.296	3.310

N THE CONDITIONING SEQUENCE AND THE RELEASE SCHEDULE

Figure 3 shows the packed input sequence at one step and the release schedule over the run. Source ids beyond the trained range are mapped onto a linear interpolation over $[1, s_{\max}]$, with $s_{\max}$ the largest source id seen in training, so that an arbitrary number of streams remains addressable. The shipped render configuration is $\omega_{\text{vid}} = 1.2$, $\omega_{\text{img}} = 5.5$, $\omega_{\text{txt}} = 2.5$, 81 frames at 16 fps, $N = 40$ steps, timestep shift 5.0, release index $r = 16$ under the *drop* policy of §3.3, with the multistep solver's history reset at $i = r$.

O REQUIRED TECHNOLOGY STACK

- **Ingestion.** HTML/PDF document parsing; LaTeX-to-symbolic conversion through a cascade of parsers with fallbacks.
- **Language models.** Any instruction-following model behind a backend interface, used for symbol extraction, cross-equation binding, relation proposal, objective resolution and code synthesis. Inference only — no training or fine-tuning is required by the method. Our runs used a hosted frontier model through an internal proxy; the pipeline is backend-agnostic and has been exercised against alternative hosted and locally served models.
- **Symbolic and dimensional layer.** A computer algebra system (SymPy) for re-derivation, closure rank, invariants and limiting cases (Meurer et al., 2017); a units library (pint) for dimensional consistency.
- **Graph layer.** An in-memory multigraph (NetworkX) for the document graph and the closure walk.
- **Numerical layer.** A stiff/non-stiff ODE integrator (SciPy) for reference trajectories; fixed-step RK4 in the emitted artifact for interaction-rate integration.
- **Execution sandbox.** A JavaScript runtime (Node) to execute and test the emitted artifact headlessly as part of gate γ.
- **Rigid-body simulation.** MuJoCo (Todorov et al., 2012), for both the driving scene and the independent pre-render probes.
- **Generative video.** A diffusion video model whose conditioning enters as tokens in the input sequence — so a stream can be withdrawn mid-sampling — and which exposes separately weighted control-video, reference-image and text terms. Our implementation uses an open, permissively licensed reference-video-to-video model (Apache-2.0) at a pinned revision, with a runtime patch implementing the release schedule and the solver reset. A flow-matching sampler with a multistep solver is assumed; any renderer meeting these conditions can be substituted.
- **Verification instruments.** A promptable video segmentation model (SAM 2.1) (Ravi et al., 2025; Kirillov et al., 2023) for mask propagation, plus at least one structurally different tracker (background subtraction and blob tracking) so that agreement between instruments is meaningful.

- **Rendering and delivery.** Document-faithful typography and equation setting (KaTeX) for the in-situ artifact; a headless browser for capture and test.
- **Compute.** Multi-GPU (8×A100 class) for the large video model, sharded per scenario; everything upstream of the render runs on CPU.

P BOOTSTRAP INTERVALS ON THE RETRIEVAL COMPARISON

Table 14 gives bootstrap intervals for the adjacent differences in non-local recall in Table 1. All four intervals lie above zero. The largest step is from dense retrieval to the symbol graph ($+0.179$), and the closure walk adds a further $+0.113$ over graph reranking, with an interval of $[0.095,\ 0.132]$. Each component's gain therefore exceeds the variation from resampling documents.

Table 14: Difference in non-local recall between each system and the one below it, with 95% confidence intervals from a bootstrap over 2,000 resamples drawn at document level. Every adjacent comparison excludes zero.

comparison	Δ non-local recall	95% CI
closure walk − graph rerank	$+0.113$	$[0.095,\ 0.132]$
graph rerank − ptbg	$+0.051$	$[0.040,\ 0.063]$
ptbg − dense	$+0.179$	$[0.161,\ 0.200]$
dense − bm25	$+0.056$	$[0.048,\ 0.065]$

Q THE MEASUREMENT IMPLEMENTATION AND ITS SELF-TEST

All arms of §4.5 pass through one measurement implementation, applied to delivered frames only. Two implementation issues were identified and corrected during this work — a frame rate assumed rather than read from the container, and a period estimator carrying no goodness-of-fit test — and every arm reported in the paper was recomputed after the correction.

The corrected implementation passes a self-test on synthetic footage with a known answer: periods of 1.0, 1.6 and 2.4 s are recovered to within 0.4%; a static scene returns no period rather than a number; jitter, jitter with drift, and pure drift reach R^2 of 0.04, 0.56 and 0.71 respectively and so fall below the 0.80 acceptance bar; and on synthetic clips constructed to obey the scale-free target, the implementation recovers it to 0.1%. These self-tests cover the synthetic cases listed; on generated video, the following consistency check is the supporting evidence. Re-measuring the conditioned branch's own footage with the corrected implementation reproduces the periods of Table 11 to within 1% on all three control values, which supports the reading that those periods come from rendered frames rather than from the arrays that drove them. The scale-free error of the conditioned branch is 4.8% under this implementation.

Separately, on eight generated oscillator clips seven yield a period recoverable by both trackers, and on those seven the recovered period lies within 0.20% to 0.71% of the closed form, the two trackers differing from one another by at most 2×10^{-5} s. For the eighth, neither tracker returned a period estimate; it is retained in the characterization as unmeasured.

R RETRIEVAL: PRECISION AND F_1

Table 15 reports precision and F_1 for the retrieval comparison of §4.2. Both follow the same order as sufficiency and non-local recall in Table 1: each added component raises both measures, and the closure walk reaches a precision of 0.159 and an F_1 of 0.227, $+0.046$ and $+0.063$ over the best baseline.

S SCALE-FREE COMPARISON WITH GENERAL-PURPOSE VIDEO MODELS

Each text-to-video system renders whatever apparent string length it chooses, and a generated clip carries no meter scale, so comparing a recovered period against $2\pi\sqrt{L/g}$ would require a pixel-to-

Table 15: Precision and F_1 for the configurations of Table 1, pooled over the five noise levels.

retrieval method	precision	F_1
BM25 (lexical)	0.098	0.143
dense (embedding)	0.113	0.164
PTBG only (symbol graph)	0.120	0.175
+ graph rerank	0.133	0.194
closure walk	**0.159**	**0.227**
Δ over best baseline	$+0.046$	$+0.063$

meter calibration that does not exist for these systems. The quantity that survives without one is the *ratio* of periods across control values: under the small-angle relation, if the rendered length tracks the prompted length then $T(L)/T(L_0) = \sqrt{L/L_0}$ holds at any pixel scale, and the correlation between $\sqrt{L}$ and the recovered period is $+1$. Neither property is assumed; both are what Table 4 measures.

A clip is scored only if it passes the re-timing control (drop every other frame; the recovered period must halve) and the sine fit attains $R^2 \geq 0.80$ on the recovered track. Both conditions are fixed before the clips are scored. The second materially affects how many clips remain: a model can render a plausible-looking pendulum that does not in fact oscillate periodically, and on such a clip a period estimate is an artifact of the estimator rather than a property of the footage. Table 4 therefore reports readable counts alongside the summaries computed from readable clips.

T Evaluation and measurement protocols

Interactive-artifact comparison. The single-call baseline receives the whole page and an instruction to produce an interactive figure for it. The evaluator operates each artifact through a fixed interface — enumerate controls, set a control, read an output, request the displayed relation — and may answer "cannot tell", which is scored as incorrect. Two arms bound the reading: the printed figure's own caption, description and stated relation (the reference arm), and an artifact built for a *different* figure (the floor arm). Figures are excluded, under rules fixed before the run, where the null arm scores at or above one half, or where the printed figure scores no better than the null.

The acceptance rule (7) is only as good as the instrument that applies it, and a tracker that inverts the transform the generator used to draw the frames agrees with that generator arithmetically rather than evidentially. The instrument was therefore characterized separately on a set of generated oscillator clips. A second tracker reads the delivered frames only — thresholding, taking the largest connected component, and computing its centroid in pixels — with no access to the world-to-pixel transform, the scenario parameters or the generator's metadata. Period is the observable both report, because it is recoverable from mean crossings in raw pixel space and the frame rate with no spatial calibration, and it has a closed form to be wrong against. Every quantity is scored twice: once on the generator's own arrays, giving the value obtained under ideal tracking, and once on the tracked frames. The difference between the two is the instrument's contribution, and is what bounds a false failure of the gate.

U Conditioning-release sweep

The release index was swept over $\{8, 14, 20, 26, 32, \text{none}\}$ on three single-object scenes (ball bounce, projectile, ramp), correlating the tracked output trajectory against the driving simulation. At step 8 the projectile is rendered as an approximately straight line, giving correlations $0.988/0.70/0.995$; step 14 is the earliest release tested at which every scene exceeds 0.99, at $1.00/0.997/0.998$; later releases hold fidelity but recover less appearance. Holding the conditioning to the end retains the simulation render's visual characteristics rather than the target photographic appearance: across one 21-scene run only 2 of 8 scenes reached photoreal output without a release, against 6 of 6 of the hardest scenes with one. The setting used, $r = 16$, lies in the 14–17 band that this sweep and the complex-scene run jointly support.

Table 16: Correlation between the rendered trajectory and the driving simulation, as a function of the step at which the simulation tokens are withdrawn. The sweep covered $\{8, 14, 20, 26, 32, \text{none}\}$; the rows below are those reported here. Step 14 is the earliest release at which every scene exceeds a correlation of 0.99. The shipped setting of $r = 16$ lies in the 14–17 band that this sweep and the complex-scene run jointly support.

release step	8	14	20	none
ball bounce	0.988	1.00	1.00	1.00
projectile	0.70	0.997	1.00	1.00
ramp	0.995	0.998	∼1.00	∼1.00

V IN-SITU FIGURE SUBSTITUTION

Figure 8 shows the in-situ substitution of §3.2 on the pendulum figure. The figure's rectangle is located from the PDF's image blocks (a), its drawn geometry is recovered from a 400 dpi render (b), and the artifact is drawn into the same rectangle (c). The artifact therefore occupies the printed figure's position on the page and starts from its drawn geometry.

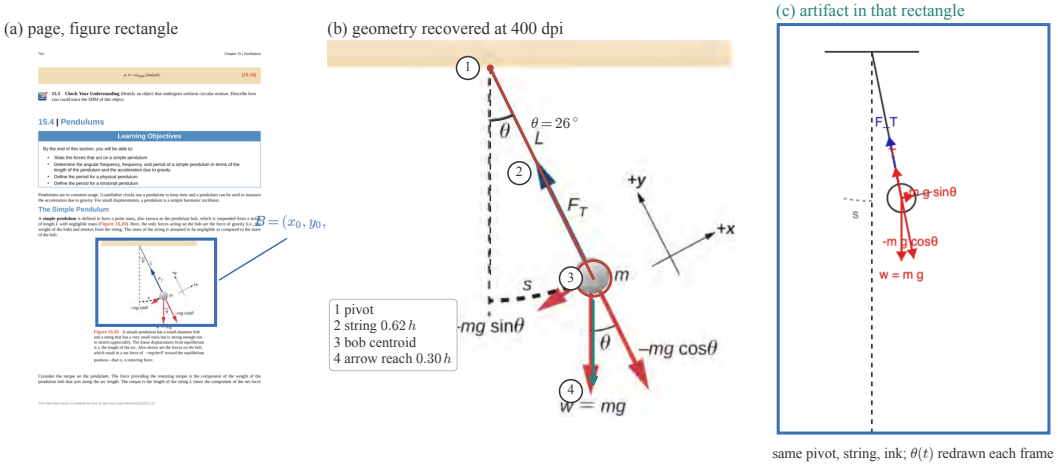

Figure 8: In-situ substitution. (a) the figure's rectangle in normalized page coordinates, taken from the PDF's own image blocks; (b) the printed figure's geometry recovered from a 400 dpi render — ceiling band, pivot, bob centroid, string fraction, printed angle, arrow reach; (c) the artifact drawn into that same rectangle, integrating (8).