TRANSLATING RESEARCH TO SLIDES: AUDIENCE-GROUNDED PRESENTATION GENERATION VIA VERIFIABLE TRANSDUCTION GRAPHS

Anonymous authors
Paper under double-blind review

ABSTRACT

Generating a scientific talk from a paper is a constrained translation problem: one shared deck must serve listeners with different knowledge and goals while preserving evidence, required conditions, and claim scope. Prior systems address important components of presentation generation—including source-content organization, audience-conditioned planning, and render-grounded refinement—but generally leave implicit how a shared narrative links source meaning to audience purpose, presentation order, and preservation requirements. We formulate this setting as *joint-audience semantic transduction*: audience needs may change expression and order, but cannot authorize unsupported claims. AudienceDeck aligns source-anchored semantic units and audience specifications into a frozen contract defining evidence bindings, audience goals, required conditions, and presentation budgets. Each adaptation becomes an executable transduction graph that records source meaning, transformation operation, story role, and conditions that must remain visible. Complete candidates are filtered by fidelity, core coverage, structure, slide, and time budgets; claims exposed before their required conditions are rejected before selection. Among feasible graphs, a prefix-sensitive selector estimates inference risk over visible narrative prefixes and terminal goal attainment after the complete narrative. The selected graph is compiled into traceable page contracts and slide objects, then checked through rendered-output verification and bounded repair. On Academic50, AudienceDeck ranks first among seven evaluated systems in macro coverage (0.7811), correctness (0.9744), and correctness-discounted time efficiency (1.9604), while ranking second in uncorrected time efficiency (2.015). On ArcBench, Judge and Quiz means reach 86.6 and 83.8. The resulting selection record connects source bindings, gate outcomes, selector decisions, rendered audits, and benchmark results in one auditable presentation process.

1 INTRODUCTION

Turning a research paper into a live presentation changes how scientific evidence reaches its audience. Readers can revisit definitions, inspect qualifications, and choose their own path through a paper; listeners encounter a speaker-guided sequence under limited time and slide budgets (Sun et al., 2021; Fu et al., 2022; Ozden et al., 2026). A shared deck must accommodate listeners who need conceptual explanation, technical evidence, or conditions of use (Mondal et al., 2024; Chen et al., 2026). These needs compete for the same narrative space, requiring coordinated choices about what to explain, how to express it, and when to reveal it.

Recent systems address important parts of this problem. Audience-conditioned planning organizes a talk around its audience and duration (Ozden et al., 2026), while render-grounded refinement uses feedback from realized slides (Tian et al., 2026). A shared narrative also requires each adaptation to connect a source meaning with its audience purpose, presentation location, and preservation requirements. An audience brief and a final deck score alone leave these connections implicit. Representing them during planning makes both the intended transformation and its constraints available for verification after rendering.

Unlike ordinary summarization, this setting requires one deck to adapt to different knowledge states without contradiction, preserve fidelity at each visible prefix, and keep every transformation exe-

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.

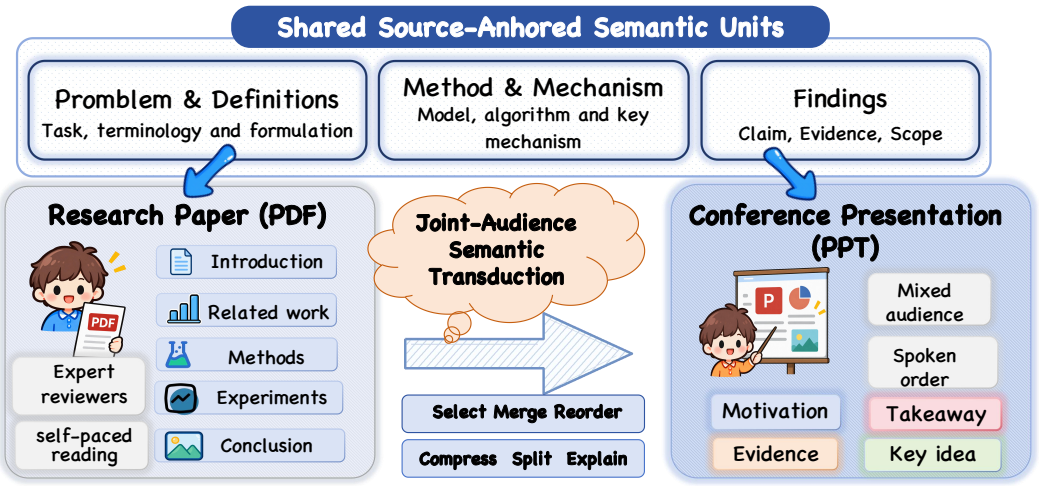


Figure 1: **Paper-to-slides as semantic transduction.** One scientific source supports a paper and a shared talk under different communicative constraints. Expression and order may change; retained claims remain bound to their evidence and scope. Translation concerns communicative form, not language.

cutable from source meaning to audience purpose, narrative location, and visible preservation requirements. These constraints motivate a plan checked before generation and after rendering.

We formulate this requirement as *joint-audience semantic transduction*: adapting source-supported content into one shared narrative while retaining its evidence and scope (Figure 1). Fidelity must hold when a claim becomes public. Presenting a finding before its necessary condition can imply a broader claim than the source supports; a later caveat cannot license that earlier statement. Even a properly qualified claim may invite different inferences across audience views. Accordingly, we first exclude prefix-level fidelity violations, then compare audience inference risk over the visible prefix—the slides exposed so far—and goal attainment after the complete narrative.

AudienceDeck implements this formulation through executable transduction graphs. It freezes a reviewed source inventory, three task-conditioned audience views, and a common brief and budget into one contract. Four complete narrative candidates bind adaptation decisions to source evidence, audience goals, and story beats. Hard gates exclude violations of fidelity, core coverage, structure, or budget. Among passing graphs, selection first controls the worst prefix risk across audience views, then favors terminal goal attainment within a specified risk tolerance. The selected graph constrains slide generation and links decisions to visible objects for rendered-output verification, carrying the same scientific obligations from planning to delivery.

We evaluate AudienceDeck on Academic50, the 50-paper academic subset of X+Slides, and the 100-paper ArcBench benchmark (Chen et al., 2026; Ozden et al., 2026). On Academic50, it ranks first among seven systems in three-profile macro coverage (0.7811), correctness (0.9744), and correctness-discounted time efficiency (1.9604), while ranking second in uncorrected time efficiency (2.015). On ArcBench, its Judge and Quiz means are 86.6 and 83.8, compared with 74.9 and 77.6 for the strongest evaluated baseline. A reduced 2×2 ablation further tests candidate selection and coverage-guided revision: enabling both yields the highest Judge-Mean, Quiz-Mean, and Hard scores across the four arms, providing protocol-bound, descriptive support for their joint contribution on the five-paper-per-arm scoring set.

Our contributions are:

- We formulate **joint-audience semantic transduction** for one shared, budgeted scientific talk, coupling audience adaptation with source fidelity and core-coverage requirements.
- We design **executable transduction graphs** that bind source meanings, adaptation decisions, and story beats, carrying preservation requirements from planning to traceable realization and visible verification.
- We introduce **prefix-sensitive candidate selection** under a common reviewed contract, separating hard feasibility from audience inference risk over the narrative and goal attainment at its completion.

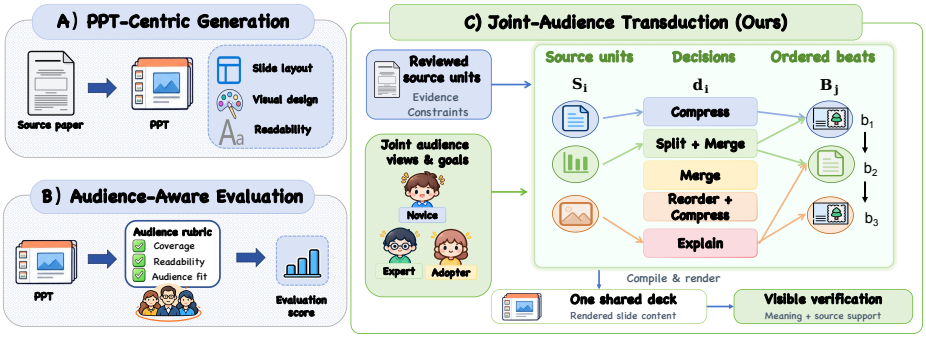


Figure 2: **Conceptual comparison of design emphases.** (A) PPT-centric generation, (B) audience-aware evaluation, and (C) our source–decision–beat planning with visible verification. These emphases are not mutually exclusive; connected boxes instantiate decisions.

2 Related Work

Automated Presentation Generation. Early paper-to-slide systems combine content selection with structured organization (Hu & Wan, 2013; Sun et al., 2021), while DOC2PPT integrates multimodal selection, paraphrasing, and layout prediction (Fu et al., 2022). Recent approaches extend toward deck-level planning and iterative generation: EvoPresent combines storyline construction with aesthetic feedback, ArcDeck conditions outline revision on a global audience- and duration-level commitment, and DeepPresenter iterates from inspected manuscript states and rendered slides (Liu et al., 2025; Ozden et al., 2026; Zheng et al., 2026). Building on these advances, AudienceDeck explicitly represents source-to-slide transformations, connecting each adaptation to its audience purpose and the scientific conditions it must preserve.

Audience-Aware Planning and Explanation. Audience modeling has long driven content and discourse decisions in language generation (Paris, 1988; Moore & Paris, 1993). In scientific presentations, Persona-Aware-D2S conditions topic planning and content selection on expertise and duration, while SlideTailor derives content preferences from example presentations (Mondal et al., 2024; Zeng et al., 2026). As an evaluation benchmark, X+Slides measures audience-conditioned generation using source-grounded probes and audience-specific utility weights (Chen et al., 2026). AudienceDeck instead coordinates heterogeneous listeners within one shared deck: their knowledge, comprehension goals, and potential misconceptions jointly inform content selection, explanation, and ordering under a common presentation budget. The internal planning views are distinct from the benchmark profiles and are not assumed to correspond one-to-one.

Evidence-Grounded Transformation and Verification. Attribute First, then Generate makes evidence selection part of content planning and realization (Slobodkin et al., 2024). In presentation generation, Auto-Slides verifies an existing plan against source content (Yang et al., 2026), while ReDeck combines atomic editing, render-grounded feedback, and deck-level fidelity checks (Tian et al., 2026). Its initial-draft pipeline does not condition planning on an explicit model of audience knowledge or goals (Tian et al., 2026, Appendix A). AudienceDeck checks the shared source interpretation and audience specification before narrative planning. Its executable transduction graph then binds transformations to evidence, preservation requirements, and presentation locations, supporting rendered-output checks of whether the intended meaning and scope are retained in visible slide content.

3 Method

3.1 Overview

Given a paper $\mathcal{P}$, its visual assets $\mathcal{V}$, and a brief q specifying purpose, language, audience, and slide/time budgets (L, T), AudienceDeck generates one editable deck for three task-conditioned audience views: *Expert*, *Novice*, and *Adopter*. The goal is not to produce three independent talks, but to execute one shared, source-grounded transduction from paper meaning to an ordered presentation. The central representation is an executable transduction graph: it specifies which source meaning is transformed, by which operation, for which audience goal, in which story beat, and with which conditions preserved. The selected graph is then compiled into page contracts, traced to visible slide objects, and audited after rendering.

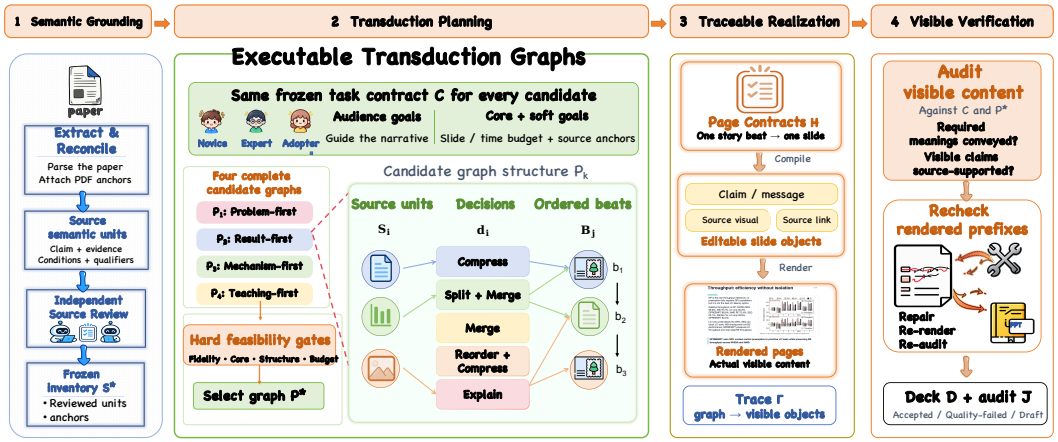


Figure 3: **Overview of AUDIENCEDECK.** Semantic grounding produces a reviewed source inventory and a frozen source–audience contract. Transduction planning constructs four complete executable graphs under the same contract, applies hard feasibility gates, and selects one graph using prefix-sensitive audience criteria. Traceable realization compiles each story beat into an editable slide and records a graph-to-object trace. Visible verification audits the rendered deck against the unchanged contract and selected graph. Bounded repair is restricted to page realization and is followed by re-rendering and re-auditing.

Figure 3 summarizes the four stages from reviewed source grounding and contract freezing to graph selection, traceable realization, and visible verification. In the evaluated implementation, each ordered story beat is compiled into one slide, so planned and rendered prefixes share the same beat-level checkpoints.

3.2 SEMANTIC GROUNDING AND THE FROZEN TASK CONTRACT

We first extract and reconcile source-anchored meanings from the paper and its visual assets, associating each unit with source anchors, protected claim attributes, and prerequisites. Evidence may be distributed across prose, captions, tables, equations, and figures. An independent source review checks these bindings against the original material before freezing the inventory $\mathcal{S}^\star$; inclusion makes a unit available for transduction but does not require its appearance in the final deck. The complete unit schema and review protocol are provided in the separately packaged appendix.

We model the shared audience through three task-conditioned views, $\mathcal{A}^\star = \{\text{Expert}, \text{Novice}, \text{Adopter}\}$. For each view, K_a denotes assumed knowledge, G_a its goals, M_a potential misconceptions, and w_{ag} goal weights; a goal g is specified by an intended understanding action ϕ_g, a visible success signal σ_g, and required source support $S_g \subseteq \mathcal{S}^\star$. Expert emphasizes precision and evidence, Novice prerequisites and explanation, and Adopter applicability and conditions of use. These are task-conditioned assumptions, not measurements of actual populations, and K_a cannot supply evidence absent from $\mathcal{S}^\star$.

The reviewed source inventory and audience views are frozen into the task contract

$$\mathcal{C} = (\mathcal{S}^\star, \mathcal{A}^\star, G_{\text{core}}, G_{\text{soft}}, q, L, T), \qquad G_{\text{core}} \cap G_{\text{soft}} = \varnothing. \tag{1}$$

G_{core} contains source-supported scientific obligations required for a faithful account; G_{soft} contains explanatory preferences that may compete under the common budget. Every candidate receives the same frozen contract: source fidelity and core coverage cannot be traded for utility, one explanation may serve several views without duplication, and the budgets (L, T) remain fixed through realization and repair.

3.3 EXECUTABLE TRANSDUCTION GRAPHS

The contract defines what must be preserved; the executable graph defines how those meanings are transformed into a shared narrative. We construct four complete candidates using problem-first, result-first, mechanism-first, and teaching-first strategies:

$$\mathcal{Q}(\mathcal{C}) = \{P_k\}_{k=1}^{4}, \qquad P_k = (\mathcal{S}^\star, \mathcal{D}_k, \mathcal{B}_k, \mathcal{E}_k), \tag{2}$$

where $\mathcal{D}_k$ is a set of transduction decisions, $\mathcal{B}_k$ is an ordered sequence of story beats, and $\mathcal{E}_k$ contains explicit source–decision–beat alignments. Candidates differ in coverage, expression, and order, but not in the scientific meanings they are allowed to claim.

The graph's executable control unit is a decision $d = (S_d, \boldsymbol{o}_d, B_d, G_d, M_d, \mathcal{I}_d)$, with $\mathcal{E}_k \subseteq \mathcal{S}^\star \times \mathcal{D}_k \times \mathcal{B}_k$. It binds a source scope, operations, destination beats, audience goals, intended visible message M_d, and protected attributes inherited from $\mathcal{C}$. Operations change communicative form rather than source-supported content: omission requires a reason and cannot remove a core obligation, merge preserves condition–claim–comparison bindings, and explanation may expose prerequisites but may not add unsupported facts. The complete operation restrictions are provided in the separately packaged appendix.

The graph is closed in the sense that every source unit receives exactly one disposition, including a reasoned omission; a retained decision may span multiple beats. This is a completeness invariant, not a proof that the source meaning or an omission is correct. The graph is executable because each decision constrains generation before layout through its operation, purpose, destination beats, intended message, and protected attributes.

For a retained decision and its complete planned or visible expression Z_d, source fidelity requires

$$\mathrm{Faith}_{\mathrm{src}}(d, Z_d) \iff [\![S_d]\!] \models [\![Z_d]\!] \wedge \Pi_{\mathcal{I}_d}([\![S_d]\!]) \equiv \Pi_{\mathcal{I}_d}([\![Z_d]\!]). \tag{3}$$

Here $[\![\cdot]\!]$ denotes scientific meaning, $\models$ denotes source support, and $\Pi_{\mathcal{I}_d}$ selects protected attributes and their claim–condition bindings. Each claim must be supported by the public context available when it is presented; a later qualification cannot license an earlier unsupported claim. A supported expression may nevertheless invite an incorrect audience inference, which is handled as audience risk rather than silently treated as a source-fidelity failure.

3.4 Graph Feasibility and Audience-Conditioned Selection

The selected graph must first satisfy hard requirements. Independent PlanQE reviews anonymized candidates against the original source and the contract, retaining

$$\widehat{\mathcal{F}}(\mathcal{C}) = \left\{ P \in \mathcal{Q}(\mathcal{C}) : \begin{array}{ll} V_{\mathrm{fidelity}}(P;\mathcal{C}) = 1, & V_{\mathrm{core}}(P;\mathcal{C}) = 1, \\ V_{\mathrm{struct}}(P;\mathcal{C}) = 1, & V_{\mathrm{budget}}(P;\mathcal{C}) = 1 \end{array} \right\}, \tag{4}$$

where $V_{\mathrm{budget}} = 1$ requires $n_{\mathrm{slides}}(P) \le L$ and $\sum_j t_j(P) \le T$. These gates check source fidelity, core coverage, ownership and prerequisites, structure, and budget. In particular, V_{fidelity} checks each visible claim at the prefix where it is presented, requiring its necessary scope and claim–condition binding to be available at that point; the complete decision-level check then verifies the protected attributes of Z_d. They are hard exclusions: high audience utility cannot rescue a candidate that fails one of them. If the passing pool is empty, the system returns no feasible candidate from this bounded pool rather than claiming that the task is globally infeasible.

Among passing graphs, selection is part of graph planning and is sensitive to the order in which a graph exposes its content. For $n_P = |\mathcal{B}(P)|$, let $\widehat{x}_j = \mathrm{Expose}(P, j)$ contain only the text, visual evidence, and qualifications committed for beat j, excluding future beats, internal rationales, and evaluation references, and let $\widehat{X}_{\le j}(P) = (\widehat{x}_1, \ldots, \widehat{x}_j)$ be the cumulative public prefix. A configured evaluator estimates, for each audience view and goal,

$$(\widehat{r}_{agj}, \widehat{u}_{agj}) = \mathrm{PrefixEval}_\vartheta\left(K_a, \widehat{X}_{\le j}(P); \phi_g, \sigma_g, M_a, S_g\right), \tag{5}$$

where $\widehat{r}_{agj}$ is the risk that the visible prefix induces a source-inconsistent inference, and $\widehat{u}_{agj}$ is support for the intended goal. The risk score concerns an incorrect inference, not merely incomplete understanding; the latter lowers attainment without automatically counting as risk. The arguments after the semicolon are scoring criteria, not information shown to the audience, and content present only there earns no credit. These are direct surrogate judgments, not measurements of human comprehension or a simulated listener state.

Let $\mathcal{J}_g(P)$ contain every checkpoint from the first public content relevant to g through the complete narrative. Terminal attainment is $\widehat{u}_{ag}(P) = \widehat{u}_{ag,n_P}(P)$ for represented goals and zero otherwise,

with $\widehat{r}_{agj}, \widehat{u}_{agj} \in [0,1]$. We aggregate graph-level risk and attainment as

$$\begin{aligned} \widehat{R}_a(P) &= \sum_{g \in G_a} w_{ag} \max_{j \in \mathcal{J}_g(P)} \widehat{r}_{agj}(P), \\ \widehat{U}_a(P) &= \sum_{g \in G_a} w_{ag} \widehat{u}_{ag}(P), \qquad w_{ag} \geq 0, \quad \sum_{g \in G_a} w_{ag} = 1. \end{aligned} \tag{6}$$

Risk preserves the highest goal-specific risk encountered during the story, whereas attainment is evaluated after the complete prefix is public. We use $\max \varnothing = 0$ for a goal with no relevant public content and set its attainment to zero; missing core content remains a hard failure.

We first control the worst audience view and then optimize attainment within a pre-specified risk-tolerance band:

$$\begin{aligned} R(P) &= \max_{a \in \mathcal{A}^\star} \widehat{R}_a(P), \qquad R_{\min} = \min_{P \in \widehat{\mathcal{F}}(\mathcal{C})} R(P), \\ \mathcal{F}_\epsilon(\mathcal{C}) &= \{P \in \widehat{\mathcal{F}}(\mathcal{C}) : R(P) \leq R_{\min} + \epsilon\}, \\ P^\star &\in \operatorname*{arg\,lex\,max}_{P \in \mathcal{F}_\epsilon(\mathcal{C})} \left(\overline{U}(P), \min_a \widehat{U}_a(P), -\operatorname{id}(P) \right), \end{aligned} \tag{7}$$

where $\overline{U}(P) = |\mathcal{A}^\star|^{-1} \sum_a \widehat{U}_a(P)$ and $\epsilon \geq 0$ is a pre-specified tolerance on surrogate risk scores, not a calibrated probability margin. Prefix-sensitive joint-audience selection is a decision rule over executable graphs, not a separate generation stage.

Concrete selection example. Consider a source whose main comparison is valid only under a recorded setting C, and whose metric is unfamiliar to at least one audience view. Two candidates can both preserve C, the same scientific obligations, and the same slide/time budget while exposing that material in different orders. A result-first candidate can show the qualified finding before a short mechanism explanation; a teaching-first candidate can expose the metric and prerequisite before showing the same finding. Both pass source-fidelity and core-coverage gates, so neither is intrinsically more truthful. Their public prefixes still differ: the former may support an immediate comparison while leaving the quantity underdefined, whereas the latter can reduce that inference risk while delaying some terminal goal attainment. Equation (7) therefore compares their worst observed prefix risk first and their completed goal attainment only within the allowed tolerance band. If a candidate exposed the finding before making C available, it would fail the prefix fidelity gate and be removed, rather than receiving only a larger soft risk score. This illustrative case explains why ordering remains meaningful after hard checks; it is not a measured run or a claim that teaching-first always wins.

3.5 Traceable Realization and Visible Verification

The selected graph is compiled into page contracts $\mathcal{H} = \{h_j\}_{j=1}^{n_{P^\star}}$. Each h_j specifies the central message, source alignments, audience goals, required visual evidence, and protected conditions for one story beat. Compilation produces an editable deck and a trace

$$(D, \Gamma) = \mathrm{Compile}(P^\star, \mathcal{H}), \qquad \Gamma \subseteq \mathcal{E}(P^\star) \times \mathcal{O}_{\mathrm{vis}}(D), \tag{8}$$

linking graph commitments to editable and rendered objects and their spatial extents. The trace identifies where a decision is realized; it does not by itself prove that the intended meaning is visible or correctly conveyed.

Independent rendered-output review checks the selected graph in both directions:

$$\begin{aligned} \mathrm{Required}(P^\star) &\subseteq \mathrm{Conveyed}(D), \\ \mathrm{SciClaims}_{\mathrm{vis}}(D) &\subseteq \mathrm{Supported}(\mathcal{C}). \end{aligned} \tag{9}$$

Here $\mathrm{Required}(P^\star)$ comprises core meanings, necessary qualifications, and selected visible commitments, not full attainment of every soft goal. The audit checks all visible scientific claims, including untraced ones, at their rendered prefixes; necessary scope and claim–condition bindings must be visible when each claim appears. Non-visible content cannot satisfy these obligations. Recomputed audience scores guide diagnosis and repair, but do not independently determine acceptance.

Bounded repair changes only page realization, preserving the inventory, contract, selected graph, ownership, and beat order. Each repair is followed by complete re-rendering and independent re-auditing; defects requiring graph changes require replanning or failure. Acceptance requires both

conditions in Eq. (9), compliance with the final slide and time budgets, and completion and passage of all required hard checks on the final artifact. It does not establish human comprehension. Detailed repair permissions and terminal states are provided in the separately packaged appendix.

4 EXPERIMENTS

We evaluate complete systems on ArcBench and Academic50, then examine plan selection and coverage-guided revision in a reduced controlled comparison.

4.1 EXPERIMENTAL SETUP AND METRICS

Data and information access. ArcBench compares seven automated systems on the same 100 papers, with no exclusions from this common set (Ozden et al., 2026). Academic50 is our local academic subset of X+Slides: 50 papers with 5,089 source-grounded probes, evaluated for specialists, learners, and decision makers (Chen et al., 2026). Each paper produces one shared deck evaluated under all three profiles. We aggregate profiles within each paper before computing macro summaries, rather than treating the 150 profile observations as independent papers. Probe questions, answers, probe-level evidence, utility weights, and reference slides are excluded from generation, selection, and repair. Public profile descriptions are permitted; internal goals are derived from the source and brief, not the evaluation bank.

Generation and comparability. AUDIENCEDECK uses a fixed generation configuration with high reasoning effort, a 15-minute talk budget, a preferred 16–20 slides, and a 24-slide ceiling unless the task specifies otherwise. We compare against ArcDeck, DeepPresenter, DeepSlides, EvoPresent, MemSlides, and SlideGen (Ozden et al., 2026; Zheng et al., 2026; Cui et al., 2026; Liu et al., 2025; Jin et al., 2026; Liang et al., 2025). The reproduction preserves native system defaults: audience briefs, slide counts, and talk durations are not matched. We report system-level performance under these configurations. Dataset provenance and execution details are given in the separately packaged appendix.

Implementation details. Within each benchmark, the source inventory, audience brief, and slide/time contract are held fixed for the evaluated system. AudienceDeck uses the fixed high-reasoning configuration above, four complete planning strategies, the same hard gates and prefix-sensitive selector, and the same renderer and visible re-audit protocol across its reported runs. The reduced ablation also holds source representation, generation model, renderer, and evaluator fixed across arms. Baselines retain their native extraction and generation defaults; we do not infer a matched evaluator configuration from their method names. Prompt versions, score anchors, evaluator settings, and runtime fields are jointly verified against the complete selection record for each reported run; the corresponding record is the execution authority for the instantiated selection rule, rather than an inference from the final deck.

Metrics. ArcBench Judge assesses technical-detail density (TQ), narrative ordering (NF), visual layout (VL), and whether visuals carry scientific content (VT). Quiz measures source-information recovery through Story, Visuals, Hard, and Depth. Both report 0–100 scores; their means average all four dimensions. Missing slide information counts as an incorrect quiz answer. ROUGE-L measures sequence overlap with the source on a separate $\times 10^3$ scale, not factual correctness.

For Academic50, let w_i^a be probe i's utility for profile a and $r_i(D) \in \{0, 1\}$ its answerability from deck D. For applicable probes $I_a = \{i : w_i^a \geq \theta\}$,

$$U_a(D) = \sum_{i \in I_a} w_i^a r_i(D), \qquad \mathrm{Cov}_a(D) = \frac{U_a(D)}{\sum_{i \in I_a} w_i^a}. \tag{10}$$

The reported scoring protocol uses $\theta = 0.7$. Eff/slide and Eff/time divide recovered utility by slide count and estimated duration, respectively, where $N_{\mathrm{time}}(D) = 0.25 N_{\mathrm{slides}} + N_{\mathrm{words}}/130$ minutes; this is neither allocated nor measured talk time. Correctness $c(D) \in [0, 1]$ independently assesses extracted claims and is reused across profiles. SafeEff/time equals $c(D)U_a(D)/N_{\mathrm{time}}(D)$ per deck and profile before aggregation, not the product of macro-average efficiency and correctness. These metrics assess benchmark recovery and claim support, not observed human comprehension. Detailed scoring and missing-evaluation boundaries are given in the separately packaged appendix.

Table 1: Full ArcBench results on the same 100 papers. Judge reports technical-detail density (TQ), narrative ordering (NF), visual layout (VL), visual content (VT), and their mean; Quiz reports Story, Visuals, Hard, Depth, and their mean. ROUGE-L is scaled by 10^3. The author reference is excluded from ranking. Bold and underline mark the best and second-best automated values per column.

	Judge					Quiz					ROUGE-L
Method	TQ	NF	VL	VT	Mean	Story	Visuals	Hard	Depth	Mean	$\times 10^3$
Author reference	31.7	46.2	80.5	92.0	62.6	–	–	–	–	–	–
DeepPresenter	<u>70.6</u>	<u>63.6</u>	79.2	86.1	<u>74.9</u>	92.7	76.9	<u>69.7</u>	<u>71.2</u>	<u>77.6</u>	<u>67.4</u>
MemSlides	65.7	56.2	71.9	<u>88.8</u>	70.7	79.5	75.6	46.2	50.0	62.8	**69.2**
SlideGen	42.8	46.6	74.8	87.8	63.0	87.1	**85.8**	51.1	54.3	69.6	52.4
DeepSlides	50.7	57.2	65.5	71.8	61.3	91.3	61.8	62.3	61.7	69.3	55.1
ArcDeck	29.1	43.2	<u>85.5</u>	85.9	60.9	<u>92.9</u>	83.2	56.2	55.5	72.0	54.9
EvoPresent	20.6	38.0	69.6	23.2	37.9	81.1	25.4	66.8	55.2	57.1	21.1
AudienceDeck	**83.9**	**79.7**	**93.7**	**89.3**	**86.6**	**94.2**	<u>83.7</u>	**77.9**	**79.3**	**83.8**	60.5

Table 2: Academic50 macro results on 50 papers. Cov. is three-profile macro coverage; E/slide and E/time are recovered utility efficiencies; Corr. is claim correctness; Safe E/time is correctness-discounted time efficiency. Values are rounded to three decimals; bold and underline mark the best and second-best values.

System	Cov.	E/slide	E/time	Corr.	Safe E/time
SlideGen	<u>0.661</u>	0.613	1.114	0.931	1.038
DeepPresenter	0.632	**1.925**	**2.025**	<u>0.939</u>	<u>1.916</u>
EvoPresent	0.618	<u>1.634</u>	1.670	0.872	1.464
ArcDeck	0.602	1.416	1.967	0.898	1.767
DeepSlides	0.518	1.455	1.584	0.831	1.328
MemSlides	0.396	1.203	1.184	0.848	1.034
AudienceDeck	**0.781**	1.587	<u>2.015</u>	**0.974**	**1.960**

4.2 Main results on ArcBench

AudienceDeck achieves the highest Judge-Mean and Quiz-Mean, 86.6 and 83.8, compared with 74.9 and 77.6 for DeepPresenter (Table 1). The dimension-level paired analysis in the separately packaged appendix shows where these mean gains arise; Quiz penalizes both omission and misstatement, so the mean improvement is not a blanket guarantee of factual correctness.

The advantage is not uniform: Visuals is below SlideGen (83.7 versus 85.8), and ROUGE-L is below MemSlides (60.5 versus 69.2). Extended paired outcomes and the author-reference comparison are in the separately packaged appendix. The paired analysis gives gains over DeepPresenter of $+11.74$ in Judge-Mean (pointwise 95% CI $[10.03, 13.45]$) and $+6.16$ in Quiz-Mean ($[4.19, 8.15]$), both with Holm-adjusted $p < 0.001$ after correction over 60 underlying pairwise comparisons (six baselines $\times$ ten reported metrics). VT and Visuals are not statistically resolved; this does not establish equivalence. These intervals quantify between-paper variation conditional on frozen scores, not new generation or judge runs.

4.3 Main results on Academic50

AudienceDeck has the highest macro-average coverage (0.7811) and correctness (0.9744), exceeding the strongest corresponding baselines by 0.1205 and 0.0351. Coverage leads in all three profiles, so the macro advantage is not driven by only one audience. Broader recovery is accompanied by higher rubric-scored claim correctness (Table 2).

Profile-level coverage confirms that the macro result is not a pooling artifact: using the rounded profile values in the separately packaged appendix, AudienceDeck improves over the strongest profile baseline by 17.6 percentage points for specialists ($0.715 - 0.539$), 7.5 points for learners ($0.802 - 0.727$), and 11.2 points for decision makers ($0.827 - 0.715$). The largest absolute gain is therefore for specialists, but all three profiles improve on the same shared deck rather than on separately customized talks. The efficiency columns expose the corresponding trade-off: AudienceDeck is strongest for specialist Eff/time and SafeEff/time, while ArcDeck or DeepPresenter remains stronger for selected learner and decision-maker efficiency measures. These benchmark pro-

files are a result decomposition, not a one-to-one validation of the internal Expert–Novice–Adopter views. We therefore use the macro table for the primary claim and treat profile values as diagnostic evidence; the complete table is provided in the separately packaged appendix.

Higher coverage does not imply uniformly higher efficiency. Among the reported systems, AUDIENCEDECK ranks first in macro SafeEff/time, second in Eff/time, and third in Eff/slide. It leads in specialist time efficiency and correctness-discounted time efficiency, but not in either measure for learners or decision makers. These results show coverage–efficiency trade-offs, not a universal profile-level efficiency advantage. Academic50 comparisons remain descriptive; no confidence intervals or significance tests are claimed.

4.4 CONTROLLED ABLATION

We report a reduced 2×2 comparison over plan selection (S) and coverage enforcement (C). All arms share the source set, model, renderer, evaluation protocol, and hard feasibility gates. $S0$ fixes the problem-first plan; $S1$ uses a role-separated selection judge to choose among the available candidates. $C0$ records soft-coverage shortfalls without forcing revision; $C1$ performs coverage-guided plan revision after a soft-coverage failure and before final realization. Fidelity, core coverage, ownership/prerequisite, structure, slide-count, and time-budget gates remain active in every arm. A candidate that fails core coverage is rejected in both $C0$ and $C1$ and does not produce a scored artifact; $C0$ can retain only a hard-gate-passing artifact with a soft-coverage shortfall. Thus $S1$ applies the candidate-selection rule to the available plans, whereas $C1$ adds revision after the unchanged hard gates; neither condition changes the frozen source contract. Each arm has five scored papers. $S1C1$ is the complete-control variant within this reduced protocol, not the original full pipeline.

Table 3: Reduced 2×2 comparison. Each arm has 5/5 scored papers under shared source, model, rendering, and evaluation conditions. Judge-Mean and Quiz-Mean retain their four-dimensional definitions; only Hard and Depth are shown separately. Bold denotes the best value.

Arm	Judge-Mean	Quiz-Mean	Hard	Depth
$S0C0$	88.92	87.58	78.87	81.20
$S0C1$	89.72	86.56	76.44	82.22
$S1C0$	84.38	88.75	81.67	**84.33**
$S1C1$	**92.50**	**92.67**	**93.33**	81.33

The complete-control arm $S1C1$ reaches Judge-Mean 92.50, Quiz-Mean 92.67, and Hard 93.33, gaining 3.58, 5.09, and 14.46 points over $S0C0$; full dimensions are provided in the separately packaged appendix. The one-factor arms expose trade-offs ($S1C0$ has the highest Depth), supporting a joint contribution within this frozen, five-paper-per-arm protocol. The comparison remains descriptive and does not isolate prefix-sensitive from terminal-only scoring.

4.5 SCOPE AND LIMITATIONS

ArcBench and Academic50 measure source-grounded recovery, claim support, and efficiency, not whether a listener forms or remembers the intended mental model. Baselines retain native defaults, and the five-paper-per-arm ablation is descriptive; prefix risk is a surrogate and is not isolated from terminal-only selection. Rendering and re-audit are included in the pipeline, while listener-level and graph-versus-outline studies remain future work. Within the reduced protocol, S1C1 provides the strongest component evidence; the complete benchmarks show coverage across task-conditioned views under one shared AudienceDeck budget.

5 CONCLUSION

We presented AUDIENCEDECK as a joint-audience semantic transduction system for turning a scientific source into one shared, audience-grounded presentation. Its executable graphs keep adaptation decisions traceable to source evidence, audience goals, claim scope, and slide locations. Across ArcBench and Academic50, the system shows strong source recovery, audience-relevant coverage, and claim correctness, with profile-dependent efficiency trade-offs. Within the reduced 2×2 comparison, jointly enabling candidate selection and coverage-guided revision yields the strongest Judge-Mean, Quiz-Mean, and Hard scores, providing protocol-bound component evidence while leaving the prefix-versus-terminal and graph-versus-outline contrasts for future work.

AI USE STATEMENT

Yes, to aid or polish writing. Details are described in the paper.

Yes, for retrieval and discovery (e.g., finding related work). Details are described in the paper.

REFERENCES

Haodong Chen, Xuanhe Zhou, Wei Zhou, Xinyue Shao, Yanbing Zhu, Bo Wang, Jiawei Hong, Anya Jia, and Fan Wu. X+Slides: Benchmarking audience-conditioned slide generation, 2026. URL https://arxiv.org/abs/2606.19256.

Zhiyao Cui, Chenxu Wang, Shuyue Hu, Yiqun Zhang, Wenqi Shao, Qiaosheng Zhang, and Zhen Wang. Design first, code later: Aesthetically pleasing template-free slides generation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 30470–30490, 2026. doi: 10.18653/v1/2026.findings-acl.1524. URL https://aclanthology.org/2026.findings-acl.1524/.

Tsu-Jui Fu, William Yang Wang, Daniel McDuff, and Yale Song. DOC2PPT: Automatic presentation slides generation from scientific documents. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(1):634–642, 2022. doi: 10.1609/aaai.v36i1.19943. URL https://ojs.aaai.org/index.php/AAAI/article/view/19943.

Yue Hu and Xiaojun Wan. PPSGen: Learning to generate presentation slides for academic papers. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, pp. 2099–2105, 2013. URL https://www.ijcai.org/Proceedings/13/Papers/310.pdf.

Ye Jin, Yangyang Xu, Jun Zhu, and Yibo Yang. MemSlides: A hierarchical memory driven agent framework for personalized slide generation with multi-turn local revision, 2026. URL https://arxiv.org/abs/2606.17162.

Xin Liang, Zhilin Zhang, Xiang Zhang, Haoran Su, Yiwei Xu, Siqi Sun, and Chenyu You. SlideGen: Collaborative multimodal agents for scientific slide generation, 2025. URL https://arxiv.org/abs/2512.04529v3. Version 3, revised August 22, 2026.

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, 2025. URL https://arxiv.org/abs/2510.05571.

Ishani Mondal, Shwetha S, Anandhavelu Natarajan, Aparna Garimella, Sambaran Bandyopadhyay, and Jordan Boyd-Graber. Presentations by the humans and for the humans: Harnessing LLMs for generating persona-aware slides from documents. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2664–2684, 2024. doi: 10.18653/v1/2024.eacl-long.163. URL https://aclanthology.org/2024.eacl-long.163/.

Johanna D. Moore and Cecile L. Paris. Planning text for advisory dialogues: Capturing intentional and rhetorical information. *Computational Linguistics*, 19(4):651–694, 1993. URL https://aclanthology.org/J93-4004/.

Tarik Can Ozden, Sachidanand VS, Furkan Horoz, Ozgur Kara, Junho Kim, and James Matthew Rehg. Narrative-driven paper-to-slide generation via ArcDeck, 2026. URL https://arxiv.org/abs/2604.11969.

Cecile L. Paris. Tailoring object descriptions to a user's level of expertise. *Computational Linguistics*, 14(3):64–78, 1988. URL https://aclanthology.org/J88-3006/.

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. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.182. URL https://aclanthology.org/2024.acl-long.182/.

Edward Sun, Yufang Hou, Dakuo Wang, Yunfeng Zhang, and Nancy X. R. Wang. D2S: Document-to-slide generation via query-based text summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1405–1418. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.naacl-main.111. URL https://aclanthology.org/2021.naacl-main.111/.

Muzhao Tian, Zezi Zeng, Yifan Yang, Xin Gao, Yan Li, Zisu Huang, Xiaohua Wang, Changze Lv, Mingxi Cheng, Bei Liu, Kai Qiu, Qi Dai, Dong Chen, Yue Dong, Xiaoqing Zheng, Ji Li, and Chong Luo. ReDeck: Step-level render-grounded refinement for document-to-slide generation, 2026. URL https://arxiv.org/abs/2609.00194v2.

Yuheng Yang, Wenjia Jiang, Yang Wang, Yi Song, Yiwei Wang, and Chi Zhang. Auto-Slides: An interactive multi-agent system for creating and customizing research presentations, 2026. URL https://arxiv.org/abs/2509.11062v3. Version 3, 2026; first released 2025.

Wenzheng Zeng, Mingyu Ouyang, Langyuan Cui, and Hwee Tou Ng. SlideTailor: Personalized presentation slide generation for scientific papers. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(41):34584–34592, 2026. doi: 10.1609/aaai.v40i41.40758. URL https://ojs.aaai.org/index.php/AAAI/article/view/40758.

Hao Zheng, Guozhao Mo, Xinru Yan, Qianhao Yuan, Wenkai Zhang, Xuanang Chen, Yaojie Lu, Hongyu Lin, Xianpei Han, and Le Sun. DeepPresenter: Environment-grounded reflection for agentic presentation generation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 31545–31558, 2026. doi: 10.18653/v1/2026.findings-acl.1578. URL https://aclanthology.org/2026.findings-acl.1578/.