From Fluent to Useful: Generative AI That Models Purpose, Audience, and Presenter for Scientific Communication

Ishani Mondal
University of Maryland, College Park
imondal@umd.edu

Abstract

Modern generative AI produces fluent text, polished slides, and clean diagrams — yet still fails when an artifact must serve a specific purpose for a specific reader, used by a specific presenter. The missing piece is not fluency but a model of why content is being produced, for whom (presenter and audience alike), and how it should adapt as goals shift. My completed and published work develops five systems across the scientific communication pipeline: ADAPTIVE IE for intent-driven extraction; Persona-Aware Slide Generation for audience reframing rather than blanket simplification; GPA for reconciling divergent group preferences; SciDoc2Diagrammer-MAF, whose multi-aspect critics distinguish purposeful abstraction from genuine omission or hallucination; and SMART-Editor, which models cascading edits across multimodal layouts. Together they show that aligning with intent, audience, and structure is necessary — but cannot answer whether the resulting artifacts actually communicate. I therefore propose three directions in priority order: (RQ1) a goal-driven framework that measures the educational utility of document-to-video generation through IRT-calibrated diagnostic questions, validated against measured learning outcomes and accompanied by inter-annotator agreement studies on human effectiveness judgments; (RQ2) presenter-side personalization that treats the presenter — not just the audience — as a first-class user; and (RQ3) a unified SuperPersonalization benchmark for transferable user preferences. RQ3 is scoped to be deferrable to post-dissertation work if RQ1 expands. The thesis shifts the target from generative AI that produces content that looks correct to systems whose outputs demonstrably communicate.

1 Introduction

Generative AI systems have become ubiquitous tools for content creation, powering applications from automated email responses (Navarro et al., 2025), education (Chukwuere, 2024), journalism (Pavlik, 2023), social media posts (Bail, 2024; Ziems et al., 2024) to slide generation (Bandyopadhyay et al., 2024; Zheng et al., 2025b) and scientific summarization (Azher et al., 2024; Marturi and Elwazzan, 2025). These systems promise to democratize expertise—offering polished language, crisp visuals, and time-saving templates for users across disciplines. Fueled by large language models (LLMs) and diffusion-based generators (Weligalle, 2025; Yin et al., 2025), such systems now assist in creating scientific presentations, research summaries, and interactive educational content.

But a generated artifact is not just a visual object. It is an act of communication, and an act of communication has three structuring questions behind it: why is this being produced, for whom, and how should it change as the goal, the source, or the feedback shifts? Today's systems mostly ignore these questions, treating generation as a one-size-fits-all surface task (Lucy et al., 2024). That gap is the widest in scientific communication, where the same underlying study may need to surface methodological caveats for a peer reviewer, mechanism intuitions for a graduate student still building vocabulary, deployment risk framing for a clinical collaborator, regulatory implications for an agency program officer, and a 30-second hook for a science journalist — and where the presenter using the artifact is also a user, with their own habits and goals, not merely a conduit. We would like to highlight three short examples to make the gap concrete:

1. **Audience flattening.** A researcher generates a slide deck for a museum public lecture. The deck is fluent but pitched at a uniform middle: too much jargon for the museum audience, too little technical scaffolding for the graduate seminar she also wants to give from the same deck. Either she edits twice from scratch, or one audience is shortchanged.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 995–1006

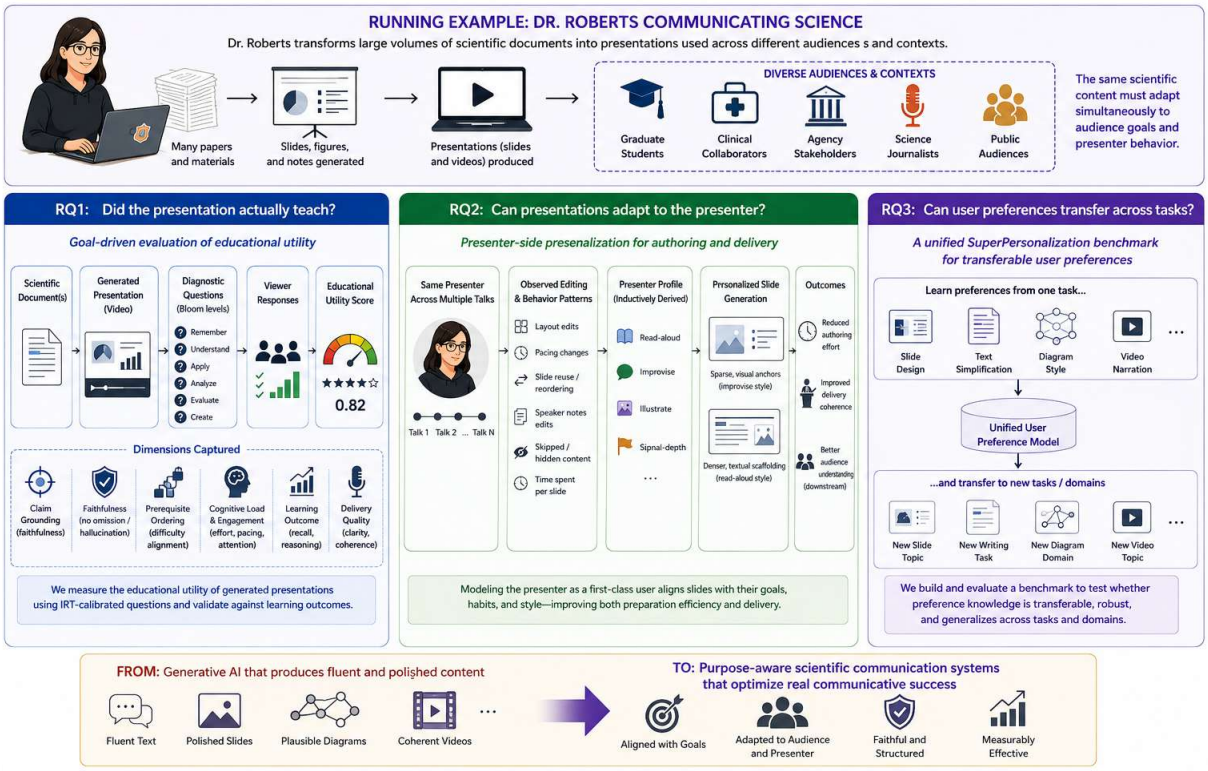

Figure 1: A scientist aims to transform large volumes of documents into a presentation that effectively educates an audience (implicit goal). While modern generative models (e.g., Gemini) can produce visually coherent outputs such as conference talks, there is currently no reliable automatic method to determine whether the generated video actually fulfills its purpose—i.e., whether it meaningfully supports audience understanding.

2. **Diagram divergence — but which kind?** A pipeline diagram drops two preprocessing steps (a deliberate simplification — they are present in the paper but irrelevant to the talk) and adds a feedback edge the paper never claimed (an unintentional hallucination). A reader staring only at the rendered diagram cannot tell which is which. Existing diagram evaluators reward "looks plausible" and miss both phenomena.

3. **Edit cascade.** A new result arrives a week before submission. Inserting it into the draft poster shifts a downstream figure into a column it overlaps with; a callback reference two slides later now points to content that was reordered out of place. Each edit was locally correct; globally the artifact is broken.

None of these failures are about fluency. The text reads well. The diagram looks fine. The slide is rendered. The failures are about purpose, audience, and structural coherence — and they are the failures I have spent five published systems trying to characterize and fix. I use a single composite scenario throughout the proposal — a researcher we will call Dr. Roberts — not as a substitute for evaluation, but as a running thread that ties otherwise distinct systems to a unified communicative pipeline. Where her scenario adds clarity, I use it; where concrete failure examples make the point better, I use those instead.

Central claim. Fluent, well-formatted outputs do not guarantee communicative success (Clark et al., 2021). Modern generative systems excel at coherence and presentation (Lauriola et al., 2025) but remain inadequate collaborators in contexts that demand adaptability, nuance, and sensitivity to both audience and presenter (Reuel and Undheim, 2024; Gill et al., 2025; Liu et al., 2025b). What is needed is a class of systems that reason about purpose — that foreground rigor where rigor is the point, that compress mechanism where intuition is the point, that know when a diagram should abstract because the speaker will fill it in verbally, and that know when an omission is a mistake rather than a deliberate simplification.

Roadmap. Section 2 reviews related work on personalization, audience-aware generation, and evalu-

ation, situating the dissertation against the broader literature. Section 3 summarizes the five completed and published systems, naming what each fixes and what it leaves unsolved. Section 4 develops the three proposed research questions with explicit timelines, hypotheses, and evaluation plans. Section 5 concludes.

2 Related Work and Position

This dissertation sits at the intersection of three literatures, none of which has been brought together explicitly for scientific communication.

Personalization in language generation. Work on persona-grounded dialogue (Liu et al., 2025b; Magister et al., 2025), preference learning from logs (Liu et al., 2025c), and audience-conditioned simplification (Kim et al., 2025) has established that user adaptation matters and is technically feasible. What remains underdeveloped is *multimodal* personalization that ties together text, layout, and figure-level decisions, and *transferable* personalization that lets preferences learned on one task carry over to another.

Audience-aware and educational content generation. Adaptive learning systems (Gill et al., 2025) and educational explainer generation (Andrist et al., 2014a) have long argued that tailoring to learner background improves outcomes. But these works typically optimize for a single learner profile in isolation; they do not address the case where one artifact must serve overlapping, divergent groups, nor the case where the *presenter*—not just the audience—has preferences the system should respect. This dissertation extends the literature by treating overlapping group preferences (GPA) and presenter-side iterative refinement (SMART-Editor and proposed RQ2) as first-class problems.

Evaluation of generated content. Existing benchmarks for text generation (Clark et al., 2021), video generation (Liu et al., 2024; He et al., 2024; Kim et al., 2025), and slide generation (Zheng et al., 2025b) lean heavily on surface fluency, audiovisual coherence, or schema match. The science-communication literature (Tayeebwa et al., 2022) provides theoretical grounding for what *useful* communication is—concept transfer, learning gain, attitudinal change—but operationalizing it for generated artifacts is open. RQ1 directly addresses this gap.

Position. Where prior work treats personalization, audience adaptation, and evaluation as separate problems, this dissertation argues they are facets of a single underlying competence: *modeling purpose*.

3 Completed Work

I describe the five completed and published systems in pipeline order: from the earliest stage of evidence gathering to the final stage of multimodal layout refinement. For each system I name (a) the specific failure it targets, (b) the technical contribution, and (c) what it leaves unsolved—the unsolved residue is what motivates the proposed work.

3.1 ADAPTIVE IE: Intent-driven extraction under shifting goals

Failure targeted. Real information needs are exploratory and shift over time. During the 2014 Chile earthquake, a Disaster Emergency Response team needs "safe routes" and "transportation options," while a USGS analyst on the same corpus needs "buildings damaged," "people trapped," and "casualties"—and either user's needs may pivot mid-session as the situation develops. Schema-based extractors (Li et al., 2022; Chambers and Jurafsky, 2011; Chambers, 2013; Cheung et al., 2013) assume fixed slot inventories defined in advance; LLM zero/few-shot extraction over a full corpus is too costly and slow for time-sensitive deployment. Neither follows the question as it evolves.

Contribution. ADAPTIVE IE (Mondal et al., 2025b) is a schema-free, human-in-the-loop framework that formalizes *IE on-the-fly*: a function $\text{IE} : \mathcal{N}_t \times C \to \mathcal{T}_t$ mapping a user's current information need to a set of thematic clusters extracted from corpus C at time t. Stage one is QA-guided unsupervised IE: an LLM identifies event triggers per document, generates trigger-conditioned WH-questions whose answers are continuous spans in the source, and clusters the resulting QA pairs into candidate slot types with LLM-generated cluster explanations. Stage two is a four-operation interactive loop—*merge*, *split*, *rearrange*, *delete*—through which users encode must-link and cannot-link constraints, after which clusters are regenerated either by centroid-based reclustering (Rec-Ren) or by name-conditioned reassignment (Ren-Rec). Across three domains (GENEVA, Biomedical Slot Filling, CrisisNLP), ten users improve F1 by up to $+0.17$ over the unsupervised baseline

within a 30-minute session, while running at lower latency and cost than full-corpus GPT-4 prompting.

Residue. ADAPTIVE IE adapts to evolving *information* needs, but says nothing about whether the resulting evidence map will produce useful communication when handed to a downstream generator. Surface-correct extractions can still feed audience-mismatched slides.

3.2 Persona-Aware Slide Generation: Purposeful reframing, not simplification

Failure targeted. A single paper has to be presented multiple ways: a short pitch to a non-technical business client emphasizes use case and impact, while a long technical talk to conference attendees needs methods, equations, and ablation results—and "simplify the jargon" is the wrong rule for the second case, because trainees and conference audiences expect to acquire and use that vocabulary. Prior document-to-slides systems (**?**Sun et al., 2021) optimize a single Rouge-style target against one gold deck and ignore audience and time-budget altogether.

Contribution. Persona-Aware Slide Generation (Mondal et al., 2024b) performs *purposeful reframing* along two axes—audience expertise (expert vs. non-expert) and time budget (short vs. long)—yielding four conditioned configurations per paper. The system is a three-stage pipeline: (1) persona-conditioned slide-outline generation, supervised on a new corpus of 75 *ACL papers each annotated with all four slide-deck configurations and then preference-fine-tuned against expert and non-expert reward models trained on pairwise judgments of *comprehensibility* and *length-fit*; (2) persona-aware content extraction, which first uses a fuzzy section-level filter to bound retrieval candidates and cut GPT calls roughly $8\times$, then performs persona-conditioned span selection with the same SFT-then-preference-optimization recipe; and (3) a summarization-and-alignment pass that converts extractive bullets into a coherent narrative across slides. Human raters distinguish short from long decks 94.4% of the time, confirming that length conditioning carries reliable signal; expert-vs-non-expert decks are harder to tell apart, foreshadowing the heterogeneous-audience problem below.

Residue. The system assumes a single audience per generation; it cannot reconcile a real-world audience that is heterogeneous, nor can it propagate audience-aware decisions consistently across a multimodal artifact (slides + diagrams + layout).

3.3 GPA: Reconciling divergent group preferences

Failure targeted. A typical departmental dry-run includes methods-focused colleagues, an applied collaborator, a senior advisor, and a trainee. Their feedback overlaps on some intents (everyone agrees the opening should motivate the problem) and diverges sharply on others (whether limitations belong in the body or appendix; how much technical depth is right for which segment). Manually balancing this produces inconsistent edits or a bias toward whoever spoke last. Persona-prompting and culture-prompting baselines (Balepur et al., 2025; Li et al., 2024) approximate groups with hand-written profiles rather than learning from how those groups actually react in situ.

Contribution. Group Preference Alignment (Mondal et al., 2025a) learns intent-conditional group preferences directly from human-AI conversation logs by mining satisfaction (SAT) and dissatisfaction (DSAT) judgments at turn level, summarizing the contrast between groups into intent-specific *rubrics*—structured descriptions, scored on a Likert significance scale, of what each group values for a given communicative move (e.g., for novices on debugging: "step-by-step guidance," "verification," "basic explanations"). Rubrics drive two complementary alignment paths: GPA-CT, a training-free path that classifies the user's intent and group at inference time and augments the system prompt with the retrieved rubric; and GPA-FT, which uses rubrics to synthesize contrastive matched-pair training data and aligns one model per group via DPO. Crucially, when groups agree on an intent the rubric is empty and no personalization fires—over-personalizing where preferences actually align is itself a failure mode. On WildChat creative-writing logs and Microsoft Copilot programming logs, evaluated across expert/novice and U.S./India/China splits, both variants beat persona-prompting and static-rubric baselines, and GPA-FT preserves general instruction-following on MT-Bench and Arena-Hard while delivering up to $+10\%$ Arena-Hard win-rate on under-served groups (Novice, China).

Residue. GPA learns preferences but does not measure whether acting on them improves com-

municative outcomes for the audience, nor does it carry preferences across tasks (slides today, video next week).

3.4 SciDoc2Diagrammer-MAF: Telling deliberate abstraction apart from omission

Failure targeted. Diagram generation is where the *intentional-vs-accidental* divergence problem is sharpest. A *good* pipeline diagram for a 60-minute keynote will deliberately leave out preprocessing detail, hyperparameters, and intermediate representations—that omission is a feature, not a bug, because including everything obscures the main flow ("missing the forest for the trees"). What an AI-generated diagram does instead is something different and worse: it drops a *causal* edge that the rest of the talk depends on, or adds a feedback loop that the paper never claimed. Both kinds of divergence look "incomplete" relative to the paper, but only one is wrong. Text-to-image baselines such as DALL-E 3 and Automatikz (Belouadi et al., 2024) either ignore the long-document context or fail to render legible scientific text.

Contribution. SciDoc2Diagrammer-MAF (Mondal et al., 2024a) introduces SCIDOC2DIAGRAMBENCH, a benchmark of 1,080 diagrams over 89 *ACL papers spanning four extrapolated diagram types (flowcharts, results plots, architectures, summary tables), and operationalizes the abstraction-vs-omission distinction in two stages. First, an intent-conditioned planner classifies the user's diagram type, generates clarification questions, retrieves answers from the paper text and from figures/tables (converting charts to tables when needed), and assembles a diagram plan; a code-generation step then renders the plan via a structured backend (e.g., graphviz, plotnine). Second, a multi-aspect feedback loop refines the rendered diagram through three critic modules: a *completeness* critic that decomposes the user intent into questions and checks whether each component needed to support that intent is present (an intent-relative judgment, not a paper-relative one), a *faithfulness* critic that generates verification questions from the diagram and compares answers against the paper to catch hallucinated edges, and a *layout* critic for readability. Two refinement schedules are studied: SUMMAF, which aggregates feedback across critics simultaneously, and SEQMAF, which applies critics one at a time until each is satisfied; SEQMAF dominates on flowcharts where load-bearing causal structure is at stake, while SUMMAF edges ahead on summary tables that benefit from balanced multi-aspect revision. Across automatic metrics and human evaluation, the refinement loop substantially improves completeness and faithfulness over zero-shot, few-shot, and self-refine baselines.

Residue. Multi-aspect critics improve diagrams locally, but the dissertation does not yet evaluate whether better diagrams produce better *learning* in viewers—the ultimate test of communicative usefulness.

3.5 SMART-Editor: Modeling cascading edits across multimodal layouts

Failure targeted. Slides, posters, and project websites are revised, reordered, repurposed, and ported. A new result inserted into a poster shifts a figure into an overlapping column; a deleted bullet leaves a callback two slides later dangling; reordering a section inverts the narrative arc. Existing tools treat each operation as local (Suri et al., 2024; Mathur et al., 2023).

Contribution. SMART-Editor (Mondal et al., 2026) models cascading effects through three coordinated agents. An *Action Agent* converts user instructions into symbolic layout operations. A *Critique Agent* checks results against constraints spanning overlap, whitespace, alignment, narrative coherence, and cross-section consistency. An *Optimizer Agent* revises using two strategies—Reward-Refine for inference-time loops and Reward-DPO for training-time preference optimization.

Residue. SMART-Editor incorporates revision feedback but does not yet model the *presenter's* cumulative editing patterns across sessions—the same person revising different decks tends to make systematically similar edits, and treating that as user-specific signal is exactly what proposed RQ2 takes up.

3.6 What the completed work establishes—and what it does not

Across these five systems a consistent picture emerges. Generative models can support scientific communication when they reason about user intent (ADAPTIVE IE), audience profile (Persona-Aware Slide Generation), reconciled group preferences (GPA), purpose-conditioned visual abstrac-

Phase	Months
RQ1: Dataset & IRT calibration	Under Submission
RQ2: Presenter longitudinal study	3-5 months
RQ3: SuperPersonalization benchmark	5-8
Buffer / writing	8-9

Table 1: Timeline for the three proposed research questions. RQ3 is scoped to be deferrable to post-dissertation work if RQ1 expands.Timeline for the three proposed research questions. RQ3 is scoped to be deferrable to post-dissertation work if RQ1 expands.

tion (SciDoc2Diagrammer-MAF), and structural cascade (SMART-Editor). Each system fixes a specific failure mode named in Section 1.

But the completed work cannot answer three questions that the dissertation must close out:

1. *Surface-level evaluation is not enough.* Each completed system is evaluated against task-specific surface metrics or expert annotations. None measures whether the resulting artifact, in the hands of a real audience, actually produces understanding. This motivates RQ1.

2. *The presenter is also a user.* The completed work treats audience adaptation as the central personalization axis. Yet a slide deck is primarily a presentation aid for a presenter, who reads from it, glances at it, or improvises around it. Different presenters use the same deck differently—and the same presenter uses the same deck differently across audiences, focusing on different sections and skipping others. This motivates RQ2.

3. *Preferences should transfer.* Each completed system learns preferences in isolation. A serious science of personalization needs to know whether preferences learned for slides transfer to videos, whether group rubrics generalize to new tasks, and whether a unified representation is even possible. This motivates RQ3.

4 Proposed Work

I propose three research questions in priority order, with explicit timelines (Table 1) and a deferrability plan. **RQ1 is the primary commitment of the remaining dissertation**; RQ2 is a planned follow-on; RQ3 is scoped so that it can become post-dissertation work.

4.1 RQ1 (primary): Evaluating whether a generated scientific video communicated the ideas of the underlying paper

A central insight motivating this work was that surface-level similarity does not indicate whether an artifact supports communication. This problem was most acute for video, where visual, textual, auditory, and temporal channels must jointly convey understanding (Andrist et al., 2014b), and where existing benchmarks emphasized visual realism, motion smoothness, or audio–visual coherence (Liu et al., 2024; He et al., 2024; Kim et al., 2025) rather than what the viewer actually learned. Even quiz-based metrics such as PRESENTQUIZ (Zhu et al., 2025) and slide-quality metrics such as PPTEVAL (Zheng et al., 2025a) rewarded content presence over explanatory depth, leaving a gap between benchmark performance and utility.

Why a single goal here? Earlier sections argued that the communicative goal varies by audience. RQ1 did not contradict this—it studied *one slice* of the goal space (instructional utility for scientifically literate viewers, as instantiated by the VISTA corpus), because (a) instructional utility was the most measurable communicative goal and admitted ground-truth signals via downstream learning outcomes (Reiter and Dale, 2000), and (b) the methodology—claim-grounded decomposition, Bloom-stratified questioning, and multi-agent diagnostic scoring—was designed to extend to other goals (persuasion, regulatory framing, accessibility) by swapping in goal-appropriate diagnostic instruments. The single-goal framing was a methodological starting point, not a theoretical commitment.

Phase 1: Benchmark construction with controlled perturbations. Starting from human-authored presentation videos in the VISTA dataset (Liu et al., 2025a), I constructed EFFECTIVEPRESENTATION-EVALBENCH: a controlled benchmark of 20 papers, each paired with seven presentation-style videos (one human-authored reference plus six automatically generated variants), for a total of 140 videos averaging 7.8 minutes and 28.6 slides each. Variants were produced by structured, prompt-driven perturbations applied to a canonical presentation plan, separated into *content-level* interventions (prerequisite omission, concept reordering) and *delivery-level* interventions (temporal misallocation, audio–visual

desynchronization, visual density manipulation), plus joint perturbations that composed both. Because all seven videos for a paper explained the same source document and were evaluated against the same question set, observed differences in viewer understanding could be causally attributed to specific instructional failures rather than topical or stylistic drift.

Phase 2: Citation-grounded, Bloom-stratified diagnostic questions. For each paper I generated two complementary question sets. *Background screening questions* tested prerequisite knowledge and explicitly excluded the paper's novel contributions; annotators had to score $\geq 7/10$ to qualify as evaluators for that paper, ensuring that observed gaps reflected presentation quality rather than prior-knowledge deficits. *Utility-defining evaluation questions* were built from *community-recognized contributions*: I extracted citation contexts from the ten most recent citing papers, normalized each into a structured acknowledgement record (attributed claim, aspect cited, citation stance), and clustered paraphrastic records into consensus contributions. Questions were then expanded across Bloom's taxonomy levels—recall, understand, analyze/apply—so that the battery probed both factual coverage and the multi-step reasoning that distinguishes summarization from teaching. Each question was paired with a paper-grounded gold answer extracted via retrieval over GROBID-parsed paper text and manually verified.

Phase 3: Human utility annotation and inter-annotator agreement. It was not clear *a priori* that human judgments of instructional utility were themselves reliable, so I treated reliability as an empirical question. Each (paper, video) pair was independently annotated by three screened annotators in a within-paper A/B-style protocol; annotators were capped at 20 videos to limit fatigue and learning effects. Of 35 initial participants recruited via Upwork, 24 met the screening criterion and were retained. For each evaluation question, annotators provided a free-text answer (graded against the gold answer on a discrete $\{0, 0.5, 1\}$ correctness scale by an LLM judge), self-reported difficulty, and self-reported viewing effort on 5-point Likert scales. The per-annotator utility combined correctness with multiplicative penalties for difficulty and effort: $u_a(v, q) = s_a(v, q)\,(1 - \tilde{d}_a(v, q))\,(1 - \tilde{e}_a(v, q))$, capturing the intuition that a presentation could technically enable correct answers while still imposing excessive cognitive load. Reliability was reported via Cohen's $\kappa = 0.71$ (pairwise) and Krippendorff's $\alpha = 0.68$ (multi-annotator), with $\kappa = 0.66$ on reasoning-oriented questions, indicating that annotators consistently identified instructional quality differences even for higher-level conceptual explanations. This IAA characterization was itself a methodological contribution: prior video-quality benchmarks rarely reported it.

Phase 4: EFFECTIVEPRESENTATIONSCORER—a multi-agent diagnostic evaluator. I developed EFFECTIVEPRESENTATIONSCORER (EPS), a goal-conditioned, claim-grounded evaluator that decomposed utility into auditable components rather than collapsing it into a single scalar. Each video was first converted into a multimodal representation $\mathcal{V} = \{(s_i, t_i, a_i, d_i)\}_{i=1}^{M}$ aligning slide images, timestamps, narration, and VLM-generated visual descriptions. Given a paper p and question q, EPS proceeded in five stages:

1. *Claim decomposition.* Retrieved the most question-relevant spans from p via Sentence-BERT (Reimers and Gurevych, 2019) similarity, then prompted an LLM to produce a dependency-structured claim set $\mathcal{C}(q)$ derived *only* from retrieved text, preventing model-prior contamination.

2. *Claim importance.* An Importance Agent assigned $I(c) \in [0, 1]$ based on each claim's role in the paper's explanatory structure, so that omissions of core causal claims were penalized more strongly than omissions of supporting context.

3. *Claim-level diagnostics.* A *Presence Agent* returned $\pi(c, v) \in \{0, 1\}$ from the video timeline; a *Faithfulness Agent* returned $F(c, v, p) \in \{0, 1\}$ by comparing video evidence against retrieved paper spans, flagging contradiction, oversimplification, or unsupported generalization.

4. *Question-level diagnostics. Coherence* penalized prerequisite-order violations between the paper-derived dependency graph and the order in which claims first appeared in $\mathcal{V}$. *Delivery* aggregated importance-weighted explanation depth, narration time, and audio–visual alignment as $D_{del}(q, v) = \sum_c I(c)\,\pi(c, v)\,\hat{T}(c, v)\,Q(c, v)\,A(c, v)$. *Engagement* was computed from acoustic sig-

nals (pitch variance, RMS-energy dynamics, tempo proxy) and treated as auxiliary.

5. *Meta-evaluation.* A Meta-Evaluator aggregated components via $U(q, v, p) = \lambda_1\pi + \lambda_2 F + \lambda_3 C + \lambda_4 D_{del} + \lambda_5 E$, with weights calibrated on a held-out 25% split (final values $\lambda = (0.30, 0.25, 0.20, 0.15, 0.10)$, robust to $\pm 20\%$ perturbation) and emitted a structured rationale attributing score differences to missing, distorted, misordered, or weakly explained claims.

Phase 5: Validation, ablation, and system comparison. I validated EPS along five complementary axes. (1) *Ranking alignment*: paper-level Kendall's τ and pairwise accuracy against human utility, separately for recall and non-recall questions, with comparison against VideoScore, EvalCrafter, PresentQuiz, PPTEval, and LecEval baselines. EPS attained $\tau = 0.58/0.53$ (recall/non-recall) versus $\leq 0.46/ \leq 0.32$ for prior metrics, with the gap widest on reasoning questions where surface metrics failed. (2) *Component ablation*: removing claim decomposition or faithfulness produced the largest degradation ($\Delta\tau \approx -0.07$ to -0.09), confirming that no single signal was sufficient and that reliable utility estimation required combining multiple diagnostic dimensions. (3) *Backbone robustness*: identical pipelines across GPT-4o, Gemini-3, and Qwen-3 confirmed that gains stemmed from the framework rather than from a specific LLM, with EPS retaining the strongest alignment under every backbone. (4) *Diagnostic feedback validity*: EPS rationales and free-form annotator comments were mapped to a shared improvement taxonomy (missing background, poor ordering, dense/rushed, unfaithful, audio–visual misalignment) via an LLM categorizer, yielding micro-averaged $F1 = 0.74$ between EPS-derived and human-derived labels, with strongest agreement on missing background ($F1 = 0.80$) and ordering violations ($F1 = 0.77$). (5) *System comparison*: applied to outputs from PAPER2VIDEO (Zhu et al., 2025), PPTAGENT (Zheng et al., 2025a), and DOC2PPT (Sun et al., 2021) alongside human-authored VISTA references, EPS isolated two recurring failure modes—insufficient prerequisite motivation and passive treatment of figures—that surface metrics consistently overlooked. Human-authored videos retained a persistent utility advantage even when automatic systems matched them on recall and faithfulness, with the gap concentrated in delivery and prerequisite grounding.

Contribution. This work contributed the first goal-conditioned, claim-grounded, interpretable framework for evaluating document-to-video generation through instructional utility rather than surface fluency; a controlled benchmark (EFFECTIVEPRESENTATION-EVALBENCH) that disentangled content failures from delivery failures; and an empirical map of which human video-quality judgments were reliable enough to serve as ground truth. Beyond evaluation, EPS produced actionable diagnostics that could be repurposed as intermediate supervision or reinforcement signals for future paper-to-video generation systems.

4.2 RQ2: Does presenter-side personalization improve authoring and downstream comprehension?

The completed work—and most prior personalization work (Liu et al., 2025b)—concentrates on adapting to the *audience*. Yet a slide deck is fundamentally a presentation aid for a presenter and is generally not meant to stand alone as an educational artifact. Presenters use slides differently: some read directly from them, some use them as memory prompts and improvise, some use them to illustrate specific points, and some use slides partly to signal that the work has more substance than time allows. The same presenter often reuses one deck across audiences, foregrounding different sections and skipping others. None of this is currently modeled. Reframing RQ2 around the presenter side—rather than re-asking whether audience-aware generation helps comprehension, which Persona-Aware Slide Generation already partly addresses—targets a genuinely underexplored axis of personalization.

Presenter typology and creator-side longitudinal study. I will recruit a cohort of researcher-presenters ($n \approx 20$) and observe them preparing and revising decks across multiple sessions. Half use a personalized system that conditions on their prior revisions, layout preferences, pacing patterns, and stated presentation style (read-aloud / improvise / illustrate / signal-depth, derived inductively from the cohort); the other half use a non-personalized baseline. I measure revision counts, edit types, time-to-completion, slide reuse behavior, speaking-outline modifications, and self-reported cognitive load, using mixed-effects regression to

disentangle participant- and topic-level variation. The completed SMART-Editor work showed that creators have stable revision patterns *within* a session; this study tests whether modeling those patterns *across* sessions reduces effort and improves consistency in presentation preparation.

Beyond efficiency, I will study whether presenter-side personalization changes the *relationship* between the presenter and the audience. Prior work in educational psychology suggests that presenters differ not only in delivery style but also in how they externalize knowledge and structure explanations. A presenter who relies heavily on slides may benefit from denser textual scaffolding, whereas an improvisational presenter may prefer sparse visual anchors that preserve conversational flow. I hypothesize that aligning generated decks with these presenter-specific strategies will reduce cognitive overhead during delivery and indirectly improve audience comprehension by producing presentations that are more coherent, confident, and naturally delivered.

To evaluate downstream effects, I will pair presenter-side measurements with audience-side studies. Participants from the audience will watch presentations produced under personalized and non-personalized conditions and complete recall, reasoning, and perceived clarity assessments. This enables testing whether presenter-side adaptation improves not only authoring efficiency but also communicative effectiveness. Importantly, the goal is not merely to optimize slide aesthetics, but to model presentations as collaborative cognitive artifacts jointly shaped by presenter intent, audience needs, and delivery style over time.

4.3 RQ3: A unified SuperPersonalization benchmark for transferable user-preference representations

Across the completed work, users vary along multiple axes: static personas (Persona-Aware D2S), evolving information needs (ADAPTIVE-IE), group-level rubrics (GPA), multi-aspect feedback patterns (SciDoc2Diagrammer-MAF), and stable editing tendencies (SMART-Editor). No existing dataset captures all of these simultaneously, which makes it impossible to test whether preferences *transfer* across tasks.

Plan. I will aggregate diverse personalization datasets into a unified textual schema recording user-profile information, preference-relevant history, task context, and user-specific target outputs. Two evaluation tasks reflect patterns from the completed work: personalized response generation and preference-conditioned rewriting. Three regimes—generic generation, history-based personalization, and rubric-based personalization (using GPA-extracted rubrics)—are compared, with both user-held-out and dataset-held-out splits to test generalization.

Deferrability. A natural alternative would be to run RQ3 first so it could serve as evaluation infrastructure for the rest of the dissertation. I considered this and chose against it for two reasons: (a) the rubric-extraction methodology depends on GPA's design, which is already complete and provides the schema; and (b) RQ1's evaluation framework is the dissertation's primary new technical contribution, and committing the year-1 budget to it preserves the most defensible scope. If RQ1 expands, RQ3 is the cleanest direction to defer to post-dissertation work without breaking the narrative arc: RQ2 would then stand alongside RQ1 as the two completed proposed contributions.

5 Conclusion

Fluent and visually coherent outputs are not enough for effective scientific communication. Meaningful communication requires aligning generated content with user intent, audience needs, presenter habits, and multimodal structure. The five completed and published systems of this dissertation demonstrate how each of these alignment problems can be tackled in isolation; the proposed work closes out the dissertation by (RQ1) building the first goal-driven evaluation framework for document-to-video generation grounded in measured learning outcomes, (RQ2) treating the presenter as a first-class user whose authoring experience the system should improve, and (RQ3) consolidating preference signals into a transferable benchmark.

Limitations

While this work advances the understanding of intent-, audience-, and structure-aware generation for scientific communication, several limitations remain. First, many of the proposed evaluation methods—particularly those based on diagnostic questions and learning outcomes—depend on curated datasets and human annotations. This introduces potential biases in question design, annotator judgments, and domain coverage, and may limit scal-

ability to diverse scientific fields or real-world deployment settings. Second, the proposed evaluation of communicative usefulness relies on proxies for understanding (e.g., question answering, information gain), which may not fully capture deeper cognitive processes such as long-term retention, transfer of knowledge, or real-world decision-making. Third, the personalization approaches explored in this work assume access to user preferences, interaction history, or persona descriptions. In practice, such information may be sparse, noisy, or unavailable, raising challenges for cold-start scenarios and for generalizing across unseen users and contexts. Fourth, while the systems span multiple modalities (text, diagrams, layouts), the proposed methods are evaluated primarily in controlled experimental settings. Real-world scientific communication often involves more complex, dynamic, and interactive environments that may introduce additional challenges not captured in this work. Finally, the proposed unified SuperPersonalization benchmark aggregates diverse datasets, but differences in annotation schemes, domains, and task definitions may introduce inconsistencies that affect training and evaluation. Ensuring robustness and fairness across user groups remains an open challenge. Addressing these limitations is an important direction for future work toward building generative AI systems that reliably support real-world scientific communication.

References

Lester Andrist, Valerie Chepp, Paul Dean, and Michael Miller. 2014a. Toward a video pedagogy: A teaching typology with learning goals. *Teaching Sociology*, 42.

Lester Andrist, Valerie Chepp, Paul Dean, and Michael V Miller. 2014b. Toward a video pedagogy: A teaching typology with learning goals. *Teaching Sociology*, 42(3):196–206.

Ibrahim Al Azher, Venkata Devesh Reddy Seethi, Akhil Pandey Akella, and Hamed Alhoori. 2024. Limtopic: Llm-based topic modeling and text summarization for analyzing scientific articles limitations. In *Proceedings of the 24th ACM/IEEE Joint Conference on Digital Libraries*, JCDL '24, page 1–12. ACM.

Christopher A. Bail. 2024. Can generative ai improve social science? *Proceedings of the National Academy of Sciences*, 121(21):e2314021121.

Nishant Balepur, Vishakh Padmakumar, Fumeng Yang, Shi Feng, Rachel Rudinger, and Jordan Lee Boyd-Graber. 2025. Whose boat does it float? improving personalization in preference tuning via inferred user personas. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3371–3393, Vienna, Austria. Association for Computational Linguistics.

Sambaran Bandyopadhyay, Himanshu Maheshwari, Anandhavelu Natarajan, and Apoorv Saxena. 2024. Enhancing presentation slide generation by LLMs with a multi-staged end-to-end approach. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 222–229, Tokyo, Japan. Association for Computational Linguistics.

Jonas Belouadi, Anne Lauscher, and Steffen Eger. 2024. Automatikz: Text-guided synthesis of scientific vector graphics with tikz. *Preprint*, arXiv:2310.00367.

Nathanael Chambers. 2013. Event schema induction with a probabilistic entity-driven model. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Seattle, Washington, USA. Association for Computational Linguistics.

Nathanael Chambers and Dan Jurafsky. 2011. Template-based information extraction without the templates.

Jackie Chi Kit Cheung, Hoifung Poon, and Lucy Vanderwende. 2013. Probabilistic frame induction. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 837–846, Atlanta, Georgia. Association for Computational Linguistics.

Joshua Ebere Chukwuere. 2024. Developing generative ai chatbots conceptual framework for higher education. *Preprint*, arXiv:2403.19303.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All that's 'human' is not gold: Evaluating human evaluation of generated text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7282–7296, Online. Association for Computational Linguistics.

Riya Gill, Ievgeniia Kuzminykh, Maher Salem, and Bogdan Ghita. 2025. Generative ai-enabled adaptive learning platform: How i can help you pass your driving test? *Artificial Intelligence in Education*, 2(1):1–17.

Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, Kai Wang, Quy Duc Do, Yuansheng Ni, Bohan Lyu, Yaswanth Narsupalli, Rongqi Fan, Zhiheng Lyu, Bill Yuchen Lin, and Wenhu Chen. 2024. VideoScore: Building automatic metrics to simulate fine-grained human feedback for video generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language*

Processing, pages 2105–2123, Miami, Florida, USA. Association for Computational Linguistics.

Taewook Kim, Dhruv Agarwal, Jordan Ackerman, and Manaswi Saha. 2025. Steering ai-driven personalization of scientific text for general audiences. *Preprint*, arXiv:2411.09969.

Ivano Lauriola, Stefano Campese, and Alessandro Moschitti. 2025. Analyzing and improving coherence of large language models in question answering. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11740–11755, Albuquerque, New Mexico. Association for Computational Linguistics.

Cheng Li, Mengzhou Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024. Culturellm: Incorporating cultural differences into large language models. *Preprint*, arXiv:2402.10946.

Lishuang Li, Ruiyuan Lian, Hongbin Lu, and Jingyao Tang. 2022. Document-level biomedical relation extraction based on multi-dimensional fusion information and multi-granularity logical reasoning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2098–2107, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Dongqi Liu, Chenxi Whitehouse, Xi Yu, Louis Mahon, Rohit Saxena, Zheng Zhao, Yifu Qiu, Mirella Lapata, and Vera Demberg. 2025a. What is that talk about? a video-to-text summarization dataset for scientific presentations. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6187–6210, Vienna, Austria. Association for Computational Linguistics.

Dongshuo Liu, Zhijing Wu, Dandan Song, and Heyan Huang. 2025b. A persona-aware LLM-enhanced framework for multi-session personalized dialogue generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 103–123, Vienna, Austria. Association for Computational Linguistics.

Junhua Liu, Yong Keat Tan, Bin Fu, and Kwan Hui Lim. 2025c. From intents to conversations: Generating intent-driven dialogues with contrastive learning for multi-turn classification. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, CIKM '25, page 1861–1871, New York, NY, USA. Association for Computing Machinery.

Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. 2024. Evalcrafter: Benchmarking and evaluating large video generation models. *Preprint*, arXiv:2310.11440.

Li Lucy, Su Lin Blodgett, Milad Shokouhi, Hanna Wallach, and Alexandra Olteanu. 2024. "one-size-fits-all"? examining expectations around what constitute "fair" or "good" NLG system behaviors. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1054–1089, Mexico City, Mexico. Association for Computational Linguistics.

Lucie Charlotte Magister, Katherine Metcalf, Yizhe Zhang, and Maartje Ter Hoeve. 2025. On the way to LLM personalization: Learning to remember user conversations. In *Proceedings of the First Workshop on Large Language Model Memorization (L2M2)*, pages 61–77, Vienna, Austria. Association for Computational Linguistics.

Krishna Chaitanya Marturi and Heba H. Elwazzan. 2025. Llm-guided planning and summary-based scientific text simplification: Ds@gt at clef 2025 simpletext. *Preprint*, arXiv:2508.11816.

Puneet Mathur, Rajiv Jain, Jiuxiang Gu, Franck Dernoncourt, Dinesh Manocha, and Vlad I. Morariu. 2023. Docedit: Language-guided document editing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(2):1914–1922.

Ishani Mondal, Meera Bharadwaj, Ayush Roy, Aparna Garimella, and Jordan Lee Boyd-Graber. 2026. SMART-editor: A multi-agent framework for human-like design editing with structural integrity. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 3219–3245, Rabat, Morocco. Association for Computational Linguistics.

Ishani Mondal, Zongxia Li, Yufang Hou, Anandhavelu Natarajan, Aparna Garimella, and Jordan Lee Boyd-Graber. 2024a. SciDoc2Diagrammer-MAF: Towards generation of scientific diagrams from documents guided by multi-aspect feedback refinement. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13342–13375, Miami, Florida, USA. Association for Computational Linguistics.

Ishani Mondal, Shwetha S, Anandhavelu Natarajan, Aparna Garimella, Sambaran Bandyopadhyay, and Jordan Boyd-Graber. 2024b. 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)*, pages 2664–2684, St. Julian's, Malta. Association for Computational Linguistics.

Ishani Mondal, Jack W. Stokes, Sujay Kumar Jauhar, Longqi Yang, Mengting Wan, Xiaofeng Xu, Xia Song, Jordan Lee Boyd-Graber, and Jennifer Neville. 2025a. Group preference alignment: Customizing LLM responses from in-situ conversations only when needed. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 825–849, Suzhou (China). Association for Computational Linguistics.

Ishani Mondal, Michelle Yuan, Anandhavelu N, Aparna Garimella, Francis Ferraro, Andrew Blair-Stanek, Benjamin Van Durme, and Jordan Boyd-Graber. 2025b. ADAPTIVE IE: Investigating the complementarity of human-AI collaboration to adaptively extract information on-the-fly. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5870–5889, Abu Dhabi, UAE. Association for Computational Linguistics.

Andres Navarro, Carlos de Quinto, and José Alberto Hernández. 2025. Email as the interface to generative ai models: Seamless administrative automation. *Preprint*, arXiv:2506.23850.

John V. Pavlik. 2023. Collaborating with chatgpt: Considering the implications of generative artificial intelligence for journalism and media education. *Journalism & Mass Communication Educator*, 78(1):84–93.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of Empirical Methods in Natural Language Processing*.

Ehud Reiter and Robert Dale. 2000. *Building Natural Language Generation Systems*. Cambridge University Press, New York, NY, USA.

Anka Reuel and Trond Arne Undheim. 2024. Generative ai needs adaptive governance. *Preprint*, arXiv:2406.04554.

Edward Sun, Yufang Hou, Dakuo Wang, Yunfeng Zhang, and Nancy X. R. Wang. 2021. D2S: Document-to-slide generation via query-based text summarization. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*.

Manan Suri, Puneet Mathur, Franck Dernoncourt, Rajiv Jain, Vlad I Morariu, Ramit Sawhney, Preslav Nakov, and Dinesh Manocha. 2024. DocEdit-v2: Document structure editing via multimodal LLM grounding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15485–15505, Miami, Florida, USA. Association for Computational Linguistics.

Dr William Tayeebwa, Dr Charles Wendo, and Dr Aisha Sembatya Nakiwala. 2022. Theories and models of science communication. *CABI Books*, page 14–22.

Ashen Weligalle. 2025. Discrete diffusion models for language generation. *Preprint*, arXiv:2507.07050.

Aoxiong Yin, Kai Shen, Yichong Leng, Xu Tan, Xinyu Zhou, Juncheng Li, and Siliang Tang. 2025. The best of both worlds: Integrating language models and diffusion models for video generation. *Preprint*, arXiv:2503.04606.

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

Hao Zheng, Xinyan Guan, Hao Kong, Jia Zheng, Weixiang Zhou, Hongyu Lin, Yaojie Lu, Ben He, Xianpei Han, and Le Sun. 2025b. Pptagent: Generating and evaluating presentations beyond text-to-slides. *Preprint*, arXiv:2501.03936.

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

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.