Consensus-Aware Bridge Maintenance Planning with Auditable Evidence and Multi-Stakeholder AI Evaluation

Takayuki Shinohara [1] **Hidetaka Saomoto** [1] **Jun Katagiri** [1]

Abstract

Bridge-maintenance planning under annual budgets must balance safety, cost, disruption, and equity while remaining explainable in public review. Working with engineers from a *railway operator* and a *major general contractor*, we translate three adoption barriers, namely hidden burdens, justification overhead, and value conflicts, into measurable criteria. We propose a decision-ready framework that combines a lightweight digital twin, constrained multi-objective optimization, synthetic *Virtual Citizens* for distribution-sensitive burden metrics, and an evidence-restricted LLM evaluator for multi-persona acceptability and disagreement. The LLM scores only auditable plan-summary JSON with deterministic decoding and caching. In an offline case study on public data, the framework outputs reviewable plan artifacts that support deliberation, audit, and accountability rather than replacing human decision makers.

1. Introduction

Bridge-network maintenance is inherently a public decision where agencies must schedule interventions under strict budgets while balancing safety, life-cycle costs, mobility disruption, and equity (Jaafaru & Agbelie, 2022; Batty, 2018; Bell & Reed, 2022). However, this is not merely a technical optimization task: plans must withstand committee review and public scrutiny over why some communities face repeated disruptions or why certain risk reductions justify the spending. In practice, Pareto-efficient plans are often rejected because aggregate objectives hide concentrated burdens and value conflicts (Vineis et al., 2025; Rask & Shin, 2024). The central challenge is therefore not only to predict deterioration, but to generate decision-ready policies that are technically defensible and socially legitimate.

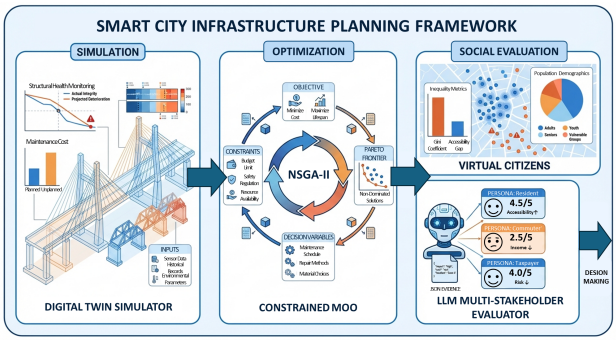

Figure 1. End-to-end infrastructure planning framework. A lightweight digital-twin simulator produces auditable evidence traces; constrained multi-objective optimization (MOO) with NSGA-II searches feasible programs; social evaluation combines Virtual Citizens and an evidence-restricted LLM evaluator.

Multi-objective optimization (MOO) is widely used to navigate these trade-offs, yet standard formulations remain incomplete because aggregated objectives hide distributional inequality and cannot represent contextual vetoes or justification criteria (Wang et al., 2017; Shih & Olson, 2022; Brandt et al., 2016). Distribution-sensitive metrics address part of this gap, but they still miss committee-level rejection patterns such as excessive concurrent closures, insufficient risk reduction for the spending, or benefits perceived as unevenly justified. Grounded LLM workflows suggest a way to operationalize such heterogeneous criteria, but free-form LLM reasoning is risky in safety-critical infrastructure because it can introduce hallucinations and unverifiable rationales (Lewis et al., 2020; Hu et al., 2023; Choi et al., 2024). For accountable governance, social evaluation must therefore be grounded in explicit engineering evidence and expressed as structured signals that can be compared and optimized.

To bridge this gap, we propose a decision-ready planning framework developed in collaboration with engineers from a railway operator and a major general contractor. Our approach couples a lightweight digital-twin simulator and constrained MOO with two explicit social-feasibility signals: a *Virtual Citizens* module that quantifies distributional fairness (e.g., tail burdens and inequality), and an *evidence-*

[1]National Institute of Advanced Industrial Science and Technology, Tsukuba, Japan. Correspondence to: Takayuki Shinohara <shinohara.takayuki@aist.go.jp>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026.

restricted LLM evaluator that scores multi-stakeholder consensus. Crucially, the LLM is not used to generate plans or narratives; instead, it scores auditable plan-summary JSON containing cost, risk, and disruption traces. This turns otherwise hard-to-code rejection categories into explicit objectives under strict evidence constraints while preserving an audit trail. The result is a Pareto search that yields programs that are not only technically efficient but also explicitly optimized for consensus feasibility and ready for committee review.

Problem, participants, method, evaluation, and impact. The real-world problem tackled in this paper is multi-year bridge maintenance planning under annual budgets, where technically efficient plans may still fail because disruption is distributed inequitably or because the rationale is not reviewable by committees. The direct collaborators in this study are engineers from a railway operator and a major general contractor; their participation is in problem formulation, requirement elicitation, and the design of auditable evidence fields and review-facing artifacts. The proposed methodology combines a lightweight digital twin, constrained NSGA-II search, Virtual-Citizens fairness metrics, and an evidence-restricted LLM evaluator. We evaluate the framework as an offline case study on public infrastructure data using engineering, fairness, and consensus metrics, including hypervolume, feasibility rate, CFR, P90 burden, Gini, and acceptance-cost premium. The intended real-world impact is not automated decision replacement, but reviewable candidate plans and plan cards that reduce justification overhead and make trade-offs explicit in deliberation. All claims are therefore about comparative decision support under common assumptions, not about replacing agency-calibrated forecasts or stakeholder deliberation.

Collaboration scope, barrier mapping, and contribution. The cross-disciplinary collaboration in this paper is concentrated in the *problem-framing and requirement-design phase*: partner engineers helped identify three recurring adoption barriers, namely hidden distributional burdens, justification overhead, and value conflicts, and shaped how these barriers are translated into measurable objectives and auditable evidence fields. The present empirical study is therefore an *offline decision-support case study* on public data rather than a live deployment or a full prospective committee trial. We map these barriers to explicit components: Virtual Citizens for tail and inequality, auditable artifacts for review, and an evidence-restricted evaluator for multi-persona consensus and disagreement. Methodologically, the core contribution is to formulate *decision readiness* itself as an optimization target while keeping the simulator-to-evaluator interface auditable and replaceable, so the same optimization layer can wrap lightweight or higher-fidelity back-ends without changing the social-feasibility objectives or review-facing artifacts. Stronger human-in-the-loop validation, richer simulators, and larger portfolios remain important next steps.

2. Related Work

Bridge maintenance optimization under uncertainty and constraints. Bridge portfolio planning combines deterioration forecasting with long-horizon programming under annual budgets and operational constraints. Recent work has integrated learning-based prediction with multi-attribute utility modeling and evolutionary multi-objective optimization to generate feasible network-level programs and explore trade-offs among cost, condition, and service impacts (Jaafaru & Agbelie, 2022). Complementarily, stochastic deterioration modeling with Markov/semi-Markov dynamics supports long-horizon optimization of maintenance strategies under uncertainty (Liu et al., 2025). Nonetheless, deployment repeatedly encounters the bottleneck that Pareto-efficient plans can still be rejected due to limited interpretability, missing distributional accounting, or contested value judgments.

Urban digital twins and data-driven decision support. In the smart-city literature, digital twins and data-driven urban analytics are positioned as a foundation for linking sensing, simulation, and governance, while also raising practical questions about transparency and decision legitimacy (Batty, 2018; Kitchin, 2014; Zhang et al., 2025). The conceptual boundary between models and digital twins further clarifies why "data-richness" does not automatically imply decision readiness (Wright & Davidson, 2020). Transportation-oriented digital twins provide concrete examples of real-time data–simulation coupling for operations, yet still leave legitimacy and policy adoption as core barriers (Xu et al., 2023; Kušić et al., 2023; Argota Sánchez-Vaquerizo, 2022). These insights motivate treating the final "selection" step—how a plan becomes adopted policy—as a first-class computational target rather than an after-the-fact discussion.

Participatory and deliberative decision making for public legitimacy. A large body of social-science and governance research emphasizes that public legitimacy is not solely an outcome of optimizing aggregates, but of *how* decisions are formed, justified, and perceived through participatory and deliberative processes. Recent reviews and models in participatory governance/collective intelligence highlight the need for inclusiveness, accountability, and decision artifacts that support reason-giving and scrutiny in real public-sector processes (Rask & Shin, 2024; Bell & Reed, 2022). In AI-assisted public decision making, participatory and multi-stakeholder frameworks increasingly formalize decision making as a multi-actor optimization problem, where diverse preferences must be represented and reconciled rather than assumed away (Vineis et al., 2025).

Our work aligns with this perspective by explicitly modeling adoption barriers and producing auditable, decision-facing artifacts (plan cards and evidence traces) that are intended to support committee review, public justification, and audit preparation.

Preference-based and interactive multi-objective optimization vs. post-hoc MCDM. A standard response to "too many Pareto solutions" is to incorporate preferences. Preference modeling and articulation have been surveyed extensively, ranging from a priori weights/reference points to interactive elicitation and preference learning (Wang et al., 2017). Interactive evolutionary MOO methods can learn a decision maker's value function or iteratively steer the search toward a region of interest, reducing the burden of selecting from a large Pareto set (Branke et al., 2015; Lárraga & Miettinen, 2025). However, these approaches typically assume a *single* decision maker (or a unified preference model) and require repeated human interaction, which can impose cognitive/organizational costs in committee-style, multi-stakeholder settings. In contrast, classical multi-criteria decision analysis (MCDA) methods such as TOPSIS rank a posteriori solutions using criterion weights and distances to an ideal point (Shih & Olson, 2022). While widely used, such post-hoc ranking can obscure value pluralism and does not directly quantify *disagreement* or *veto risks* across heterogeneous stakeholders.

Social choice, veto robustness, and consensus metrics. When multiple stakeholders have distinct preferences, the problem resembles social choice: aggregating heterogeneous evaluations into collective decisions. Voting/social-choice research studies aggregation rules and emphasizes that majority efficiency and minority protection can be in tension, motivating notions of veto and minority safeguards (Brandt et al., 2016; Kizilkaya & Kempe, 2025). Inspired by these ideas, we define *consensus feasibility* using worst-persona acceptability and veto-avoidance rates (CFR), which makes coalition-like rejection risks measurable and optimizable. Unlike abstract voting rules that assume direct ballots, our setting derives stakeholder evaluations from auditable engineering evidence and distributional metrics, bridging quantitative optimization with deliberative legitimacy.

LLM/knowledge-grounded assistance for infrastructure O&M and structured evaluation. LLM-based workflows are increasingly explored for infrastructure decision support, especially when grounded by explicit knowledge representations to improve consistency and auditability. For example, knowledge-graph-driven bridge maintenance decision-making has been proposed to structure reasoning and reduce ambiguity in maintenance workflows (Wang et al., 2026b;a). More generally, retrieval-augmented generation and dense retrieval provide a standard mechanism to ground LLM outputs in external evidence and reduce unfaithful generations (Lewis et al., 2020; Karpukhin et al., 2020), and surveys on knowledge-enhanced language models systematize these grounding strategies (Hu et al., 2023). Complementary to evaluation-centered uses, recent work leverages LLMs to extract component-level condition and damage information from narrative bridge inspection reports/PDFs and fit mechanism-aware Markov deterioration models, enabling large-scale predictive maintenance planning in Japan (Shinohara et al., 2025). Our work aligns with the grounding principle, but places the LLM in an *evidence-only evaluation loop* to quantify multi-stakeholder acceptability and disagreement as optimization objectives. Crucially, the LLM is *restricted to an evidence-only plan-summary JSON* and produces structured scores, complementing classical cost–risk–service metrics and distribution-aware disruption accounting while improving auditability in safety-critical settings (Choi et al., 2024).

3. Proposed Method

We propose an end-to-end framework for *decision-ready* bridge maintenance planning that couples constrained multi-objective optimization with two explicit signals of *social feasibility*: (i) distribution-aware disruption metrics computed from *Virtual Citizens*, and (ii) an *LLM-based multi-stakeholder evaluator* that scores plan acceptability and disagreement across personas. A key design choice is to use a *lightweight* bridge digital twin: evolutionary MOO requires evaluating thousands of candidate plans, so the simulator must be fast while still producing *auditable evidence* (cost/risk/disruption traces) for downstream justification. Crucially, the LLM is used only as an *evidence-grounded evaluator* operating on structured plan summaries (*evidence-only plan-summary JSON*), which mitigates hallucination and improves accountability by making every score traceable to explicit inputs. Our notion of *decision-ready* focuses on producing *reviewable artifacts* (evidence traces, plan cards, and consensus signals), rather than claiming high-fidelity absolute forecasts. Figure 2 illustrates the pipeline.

3.1. Problem Setting and Decision Encoding

Let $\mathcal{B} = \{1, \dots, N\}$ be bridges and $\mathcal{T} = \{1, \dots, T\}$ the annual planning horizon. Each bridge i has an initial condition distribution $\mathbf{p}_{i,1} \in \Delta^4$ over $\{\mathrm{CS1}, \mathrm{CS2}, \mathrm{CS3}, \mathrm{CS4}\}$, computed by aggregating element-level inspection quantities into bridge-level ratios. We consider interventions $\mathcal{A} = \{\texttt{none}, \texttt{minor}, \texttt{major}, \texttt{replace}\}$ and restrict to *at most one intervention per bridge* during the horizon for scalability. This simplification is sufficient to test the core question of this paper—how to optimize *consensus feasibility* as an explicit objective given auditable plan evidence—and the

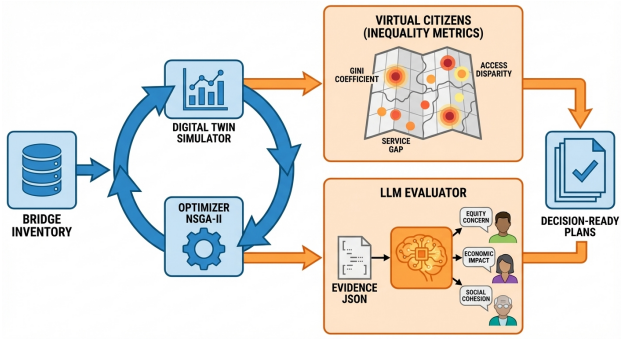

Figure 2. Decision-ready planning framework: a lightweight digital-twin simulator produces evidence (risk/cost/disruption traces) for constrained multi-objective optimization; Virtual Citizens capture distributional burdens (tail and inequality); an LLM evaluator scores multi-persona acceptability using evidence-only plan summaries.

same evidence interface extends to richer encodings (e.g., multiple interventions, bundled projects, or additional resource constraints). A plan is encoded as an integer genome $\mathbf{g} \in \{0, \ldots, 3T\}^N$ and decoded into an action type and execution year for each bridge.

Decision-ready outputs (artifacts and signals). For each candidate plan $\mathbf{x}$, the framework produces (i) auditable annual evidence traces (cost/risk/disruption with budget margins), (ii) distributional burden summaries from *Virtual Citizens* (e.g., tail burden and inequality), and (iii) multi-persona acceptability/disagreement signals from an *evidence-restricted* evaluator. These are compiled into standardized *plan cards* and an *evidence-only plan-summary JSON* suitable for deliberation, review, and audit.

3.2. Lightweight Digital-Twin Simulator

Given a decoded plan, a lightweight simulator produces auditable annual traces of condition/risk, agency cost, and user disruption. We define a scalar risk proxy from a condition distribution:

$$\rho(\mathbf{p}) = \sum_{k=1}^{4} w_k p^{(k)}, \qquad \mathbf{w} = [0,\ 0.1,\ 0.5,\ 1.0], \tag{1}$$

where weights are monotone by condition state and encode the intended severity ordering (CS4 as highest-risk, CS2/CS3 as intermediate). Deterioration is modeled as a monotone Markov process with a risk-coupled rate, and interventions apply a deterministic improvement operator before propagation. This abstraction is not meant to replace high-fidelity deterioration models; rather, it provides a consistent, computationally efficient mapping from interventions to risk trajectories that can be called thousands of times inside NSGA-II. We further verify that key comparative conclusions (e.g., CFR improvements) remain stable under parameter perturbations (e.g., rate scaling and weight variations; see Supplementary Sec. S1.3). Full transition matrices, rate clipping, and action operators are given in Supplementary Sec. S1.3.

Why lightweight is a design requirement. Our simulator is intentionally lightweight to satisfy three operational requirements:

- **Design requirement (scale & interactivity):** evolutionary MOO evaluates thousands of plans; thus the simulator must support fast what-if exploration and repeated scoring within realistic time budgets (e.g., committee-style review).
- **Relative validity (ranking/trade-off shape):** decision support often relies on *comparative evidence* (plan ranking and trade-off structure) more than on agency-calibrated absolute forecasts; we therefore emphasize consistent, auditable traces for comparing candidate plans under the same assumptions.
- **Modularity (replaceable back-end):** the simulator and disruption proxy define an evidence interface (traces and plan-summary fields) that can be replaced by higher-fidelity twins or traffic-assignment modules when available, while keeping the optimization and social-feasibility evaluation unchanged.

We therefore interpret the current twin as a decision-support back-end for *relative comparison*, not as a claim of historically calibrated predictive fidelity; calibrating or replacing it with a higher-fidelity simulator is a compatible extension of the same framework.

3.3. Objectives and Constraints

We optimize maintenance programs under annual budgets while balancing safety, disruption, distributional burden, and acceptability. The baseline objective vector is

$$\min_{\mathbf{x}} \ \mathbf{F}(\mathbf{x}) = \big[J_{\text{cost}}(\mathbf{x}),\ J_{\text{risk}}(\mathbf{x}),\ J_{\text{peak}}(\mathbf{x}),\ J_{\text{user}}(\mathbf{x})\big], \tag{2}$$

$$\text{s.t. } C_t(\mathbf{x}) \le B_t,\ \forall t \in \mathcal{T}. \tag{3}$$

Here $J_{\text{cost}}(\mathbf{x}) = \sum_t \gamma^t C_t(\mathbf{x})$ is discounted agency spending, J_{risk} is discounted cumulative structural risk, J_{peak} captures worst-case annual risk, and J_{user} measures discounted user disruption. Let $\rho_{i,t}(\mathbf{x})$ be the simulated risk proxy for bridge i in year t and α_i its importance weight. We define $J_{\text{risk}}(\mathbf{x}) = \sum_t \gamma^t \sum_i \alpha_i\, \rho_{i,t}(\mathbf{x})$ and $J_{\text{peak}}(\mathbf{x}) = \max_t \max_i \rho_{i,t}(\mathbf{x})$ to capture high-consequence outliers. $J_{\text{user}}(\mathbf{x}) = \sum_t \gamma^t \sum_i \Delta H_{i,t}(\mathbf{x})$ measures discounted user disruption in interpretable units (vehicle-hours), using detour-time increments estimated from local OSM-based shortest-path recomputation or conservative proxies.

Virtual Citizens: distributional disruption beyond averages. Portfolio-level user delay can obscure *who* bears the burden. Instead of only optimizing an average proxy, we simulate M *Virtual Citizens*[1] with heterogeneous bridge-usage patterns and compute distribution-sensitive metrics. In the default implementation, each citizen is a synthetic archetype spatially anchored near the bridge set and assigned sparse top-k bridge exposures weighted by ADT and distance; Supplementary Sec. S2.2 gives the full construction. Let $\Delta H_m(\mathbf{x})$ be the discounted added hours for citizen m under plan $\mathbf{x}$. We then define, for example,

$$J_{\mathrm{p90}}(\mathbf{x}) = \mathrm{Quantile}_{0.9}\big(\{\Delta H_m(\mathbf{x})\}_{m=1}^{M}\big),$$
$$J_{\mathrm{gini}}(\mathbf{x}) = \mathrm{Gini}\big(\{\Delta H_m(\mathbf{x})\}_{m=1}^{M}\big). \quad (4)$$

These quantities expose tail burdens and inequality of disruption that are invisible to mean-based objectives. When enabled, these Virtual-Citizen objectives are appended to $\mathbf{F}(\mathbf{x})$.

Social meaning and privacy. Virtual Citizens are designed to make *concentrated burdens* visible in planning, surfacing cases where low average disruption still corresponds to large tail impacts or high inequality. They do *not* infer personal attributes or use individual mobility traces; all citizens are simulated, enabling distributional accounting while avoiding privacy risks.

3.4. LLM-based Multi-stakeholder Acceptance Evaluation

Even Pareto-efficient solutions may be rejected due to institutional constraints and contested justifications. We therefore introduce an LLM evaluator that scores plan *acceptability* and *disagreement* across personas. In this framework, the LLM is treated as a *proxy, early-warning evaluator* for potential veto and disagreement risks in committee-style deliberation, not as normative ground truth. In the present paper it is evaluated as an offline proxy layered over auditable plan evidence rather than as a substitute for human committee judgments.

Evidence-only grounding for auditability. For each plan, we generate an evidence-only *plan-summary JSON* containing: intervention list, annual cost and budget margins, risk trajectories, disruption indicators, and (if enabled) Virtual-Citizens distributional metrics. The LLM is instructed to use *only* this JSON as evidence and to output structured persona scores. Because the entire input evidence is explicit and logged, every score can be audited and compared across runs.

[1]We also refer to them as *synthetic citizens* to emphasize that they are fully simulated agents; we use *Virtual Citizens* as the consistent proper name throughout.

Reproducibility and robustness controls. We use **deterministic decoding** (temperature $= 0$) and **disk caching** keyed by a hash of the plan-summary JSON and persona specification, ensuring deterministic reuse across optimization generations. Prompt templates, persona definitions, and the structured output schema are made explicit in the supplementary material so that sensitivity analyses to persona wording or model choice can be reproduced in future work.

Persona specification as a controlled variable. Heterogeneous stakeholder values are represented as a finite set of personas $\mathcal{S}$ with $K = |\mathcal{S}|$ (e.g., *Safety-first engineer*, *Budget watchdog*, *Mobility advocate*). Rather than treating "stakeholder preference" as implicit behavior, we treat persona definitions as explicit variables and define

$$J_{\mathrm{LLM,min}}(\mathbf{x}) = 1 - \min_{k\in\mathcal{S}} \mathtt{accept}_k(\mathbf{x}), \quad (5)$$
$$J_{\mathrm{LLM,mean}}(\mathbf{x}) = 1 - \mathrm{mean}_{k\in\mathcal{S}}\mathtt{accept}_k(\mathbf{x}), \quad (6)$$
$$J_{\mathrm{LLM,dis}}(\mathbf{x}) = \mathrm{Var}_{k\in\mathcal{S}}(\mathtt{accept}_k(\mathbf{x})). \quad (7)$$

Here $J_{\mathrm{LLM,min}}$ is a veto-robustness proxy, $J_{\mathrm{LLM,mean}}$ summarizes average acceptability, and $J_{\mathrm{LLM,dis}}$ captures polarization risk. Prompt template, personas, caching strategy, and structured output format are detailed in Supplementary S2.2.

Complementary roles of Virtual Citizens and LLM evaluation. Virtual Citizens provide *mathematically defined* distributional signals (e.g., tail burden and inequality of disruption). The LLM complements this by scoring *contextual acceptability* (institutional and justification constraints) *only from explicit evidence* in the plan-summary JSON. Using both signals reduces reliance on any single proxy: distributional metrics constrain and contextualize LLM judgments, while LLM scores capture acceptance dynamics not fully specified by numeric objectives alone.

Mitigating LLM-gaming concerns. Optimizing acceptability cannot fabricate evidence: the LLM receives only logged traces and summary fields, so improving its scores requires changing the underlying plan and evidence. Moreover, we jointly optimize disagreement/dispersion signals, discouraging plans that merely cater to a single persona at the expense of broad consensus.

3.5. Constrained Multi-objective Optimization and Outputs

We approximate the Pareto front with NSGA-II under annual budget constraints. Each individual is decoded, simulated, assigned objective values, and evaluated for constraint violations. We use standard feasibility-first selection; detailed settings (population size, mutation/crossover, integer clipping) are reported in Supplementary S2.2. Algorithm 1

summarizes the evaluation loop used within NSGA-II in the supplementary material, and caching yields high reuse across generations, reducing total LLM calls (see Supplementary S2.2).

Decision-facing outputs. The method outputs nondominated plans together with auditable artifacts: annual traces, distributional disruption statistics, and persona-conditioned acceptability scores, enabling downstream decision workflows that require both quantitative optimality and stakeholder-facing justification.

4. Experiments

4.1. Experimental Setup

4.1.1. Data and portfolio construction

Our primary data source is the U.S. National Bridge Inventory (NBI). From a target region we construct a heterogeneous bridge portfolio with attributes such as location, deck area, ADT, and an initial condition distribution $\mathbf{p}_{i,1} \in \Delta^4$ derived by aggregating element-level condition quantities (CS1–CS4). Unless otherwise stated, we optimize subsets of size $N = 50$ (and report trends for larger N in Supplementary).

4.1.2. Disruption proxy (OSM detours)

When enabled, user disruption is estimated by a local OSM-based detour proxy. For each bridge we tile-cache a road graph within a bounding box (on the order of 10 km), compare shortest-path travel times before/after removing the bridge-linked segment, and conservatively cap detours when no alternate route exists within the local region. This provides a physically interpretable disruption signal in vehicle-hours.

4.1.3. Optimization and reproducibility

We optimize a T-year program under annual budgets using NSGA-II with integer encoding/decoding (Section 3). Unless stated otherwise, we use $T = 10$, population $P = 40$, generations $G = 20$, and discount rate $d = 0.03$. The LLM evaluator uses deterministic decoding (temperature $= 0$) and disk caching keyed by a hash of the plan-summary JSON and persona specification, ensuring deterministic reuse across generations. All quantitative results are reported over R random seeds for initialization (and sampling of bridge subsets/citizens), with identical seeds shared across methods.

Stakeholder involvement and validation scope. The collaborating engineers from a railway operator and a major general contractor contribute to requirement elicitation and artifact design; they are not presented here as a source of large-scale labeled preference data or as a deployed decision committee. Accordingly, the experiments should be read as an offline decision-support case study on public records and synthetic citizens, designed to test whether stakeholder-shaped requirements materially alter the trade-offs produced by the optimizer.

Scaling note. The default $N = 50$ setting keeps ablations, qualitative inspection, and repeated seeded runs tractable. Methodologically, the optimization loop scales with repeated simulator calls and, when enabled, with the number of unique plan-summary hashes scored by the LLM evaluator. Without caching, the number of scored plans grows on the order of $P \times G$ per run; caching and staged evaluation are therefore important for larger portfolios and are discussed further in the supplementary material.

4.2. Baselines, Ablations, and Cross-evaluation

Baselines (decision-making realism). We compare against baselines that mirror common practice and a minimal fairness-aware alternative:

- **(a) Cost-only (scalar)**: minimize discounted life-cycle cost under annual budgets.
- **(b) Classical MOO (engineering trade-off)**: multi-objective optimization over cost and safety risk (cost vs. cumulative/peak risk), representing standard portfolio-level trade-off analysis.
- **(b*) Fairness-aware MOO (constraint/weight)**: incorporate inequality via a simple fairness mechanism, e.g., $J_{\text{gini}}(\mathbf{x}) \leq \tau$ as an additional constraint or a weighted term (engineering objectives $+\lambda\, J_{\text{gini}}$), representing a standard fairness-aware extension without stakeholder modeling.
- **(c) Proposed (Full)**: engineering objectives plus Virtual Citizens (tail/inequality) and LLM-derived acceptability/disagreement objectives.

To isolate each module's contribution, we also include ablations **+VC** and **+LLM**.

Cross-evaluation for fair comparison. Methods differ in which objectives they optimize, but *decision readiness* must be assessed on a common basis. Therefore, after each run we re-score *all* returned solutions using the same evaluation pipeline: the simulator generates objective traces; Virtual Citizens compute distributional metrics; and the LLM evaluator assigns persona scores from the plan-summary JSON. This enables direct comparisons such as: *"How acceptable are plans produced without optimizing acceptability?"* and supports the central claim that small engineering trade-offs can yield large gains in consensus feasibility.

Table 1. Experimental settings (minimum set for reproducibility).

Item	Value
Portfolio size	$N = 50$ bridges (default; larger N in Supplementary)
Planning horizon	$T = 10$ years
Optimizer	NSGA-II (integer encoding/decoding)
Population / generations	$P = 40$, $G = 20$
Discount rate	$d = 0.03$, $\gamma_t = (1 + d)^{-(t-1)}$
Annual budget	$\{B_t\}_{t=1}^{T}$ (uniform unless stated)
Action set	$\{\texttt{none}, \texttt{minor}, \texttt{major}, \texttt{replace}\}$
Virtual Citizens	enabled: $M = 800$, top-$k = 6$ exposures per citizen
LLM evaluator	temperature $= 0$, K personas; caching enabled
OSM detours	tile-cached; capped when no feasible alternate route

Table 2. Compared variants and which objectives are exposed to NSGA-II. All variants are *evaluated* with the full metric suite via cross-evaluation (VC+LLM), even if those objectives were not optimized.

Variant	Virtual Citizens	LLM	Objectives exposed to NSGA-II
(a) Cost-only	–	–	$[J_{\text{cost}}]$
(b) Classical MOO	–	–	$[J_{\text{cost}}, J_{\text{risk}}, J_{\text{peak}}]$
(b*) Fairness-aware MOO	–	–	$[J_{\text{cost}}, J_{\text{risk}}, J_{\text{peak}}, J_{\text{user}}, J_{\text{gini}}]$
+VC	✓	–	$[J_{\text{cost}}, J_{\text{risk}}, J_{\text{peak}}, J_{\text{user}}, J_{\text{p90}}, J_{\text{gini}}]$
+LLM	–	✓	$[J_{\text{cost}}, J_{\text{risk}}, J_{\text{peak}}, J_{\text{user}}, J_{\text{LLM,min}}, J_{\text{LLM,dis}}]$
(c) Full (Proposed)	✓	✓	$[J_{\text{cost}}, J_{\text{risk}}, J_{\text{peak}}, J_{\text{user}}, J_{\text{p90}}, J_{\text{gini}}, J_{\text{LLM,min}}, J_{\text{LLM,dis}}]$

4.3. Results

4.3.1. Quantitative results: trade-offs and consensus feasibility

Figure 3 visualizes representative Pareto projections. Classical baselines (cost-only and engineering MOO) recover efficient cost–risk trade-offs, but cross-evaluation reveals that many solutions concentrate disruption or trigger persona vetoes despite strong engineering scores. By contrast, adding Virtual Citizens shifts the frontier toward lower tail burden and inequality at comparable risk, demonstrating that fairness is not redundant with mean disruption. Adding the LLM objectives shifts the frontier toward higher worst-persona acceptability and lower disagreement, often requiring only a modest cost increase (quantified by $\Delta J_{\text{cost}}(\theta)$). *To highlight impact, we additionally plot a CFR–cost "L-curve" (Supplementary), showing that a small cost increase can sharply reduce rejection ($1-CFR_\theta$) in the practical regime.*

We emphasize two decision-relevant regimes: (i) **low-risk regime** where incremental spending yields diminishing safety returns but large gains in consensus feasibility (high CFR_θ), and (ii) **contentious regime** where average disruption is small yet tail burden (e.g., $P90$) and inequality remain large, causing vetoes from resident-advocate personas. These regimes support the claim that the proposed objectives expose "social feasibility" trade-offs that are invisible to classical engineering MOO.

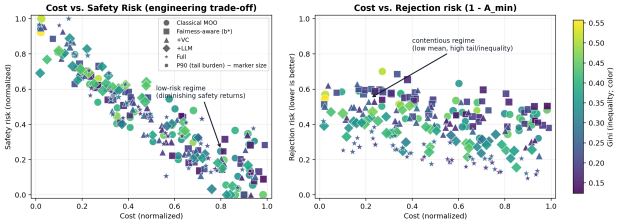

Figure 3. Pareto projections highlighting engineering trade-offs and decision-legitimacy signals. Left: normalized cost vs. normalized safety risk. Right: normalized cost vs. rejection risk $(1 - A_{\min})$ (lower is better), where $A_{\min}$ is worst-persona acceptability. Marker shape denotes the optimization variant (Classical MOO, fairness-aware $b^\star$, +VC, +LLM, Full), marker size encodes tail burden (P90), and marker color encodes inequality (Gini). Annotated regions illustrate a low-risk regime (diminishing safety returns with added cost) and a contentious regime (low average but high tail or inequality associated with elevated rejection risk).

4.3.2. Qualitative results: interpreting LLM rejections and VC failure modes

Quantitative Pareto statistics do not fully capture whether a plan is *interpretable* to a decision committee. We therefore inspect representative plans selected from each Pareto set: minimum-cost, minimum-peak-risk, minimum-inequality, and maximum-worst-acceptability. For each plan we produce a standardized *plan card* (plan-summary JSON rendered into a human-facing table) containing yearly budgets, risk/disruption traces, fairness metrics, and persona scores.

We additionally report *LLM rejection rationales* for a small

Table 3. Selected quantitative comparison across the five primary optimization variants (mean ± std over R=5 seeds). HV is computed after normalization with a common reference point. CFR_θ measures consensus feasibility (no-veto rate). The fairness-aware baseline $b^\star$ is defined in Table 2 and visualized in Figure 3; Table 3 focuses on the main module ablations that support the paper's central claim. Please refer to Supplementary Sec. S2.1 for the complete list of quantitative evaluation metrics and their precise definitions.

Metric	Cost-only	Classical MOO	+VC	+LLM	Full
HV (normalized)	0.41 ± 0.03	0.63 ± 0.02	0.68 ± 0.02	0.66 ± 0.03	$\mathbf{0.74 \pm 0.02}$
Feasible rate (%)	100.0 ± 0.0	98.2 ± 1.3	97.6 ± 1.5	96.8 ± 1.7	95.4 ± 1.2
$\text{CFR}_{0.6}$ (%)	28.4 ± 6.1	41.7 ± 5.4	58.2 ± 4.3	63.9 ± 3.9	$\mathbf{78.6 \pm 4.1}$
Min J_{gini} at matched risk	0.39 ± 0.04	0.33 ± 0.03	$\mathbf{0.22 \pm 0.02}$	0.31 ± 0.03	0.24 ± 0.02
Min $J_{\text{LLM,min}}$ at matched risk	0.54 ± 0.05	0.48 ± 0.04	0.45 ± 0.05	$\mathbf{0.31 \pm 0.03}$	0.33 ± 0.04
Acceptance cost premium $\Delta J_{\text{cost}}(0.6)$	+0.0%	+1.8%	+2.4%	+3.1%	**+3.5%**

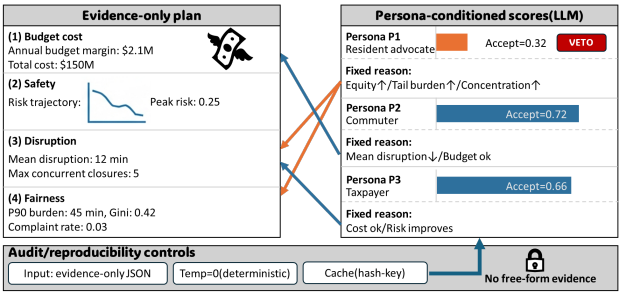

Figure 4. Evidence-grounded interpretability for a vetoed plan. Left: evidence-only plan summary (auditable fields) including budget/cost, safety (risk trajectory/peak risk), disruption (mean disruption/max concurrent closures), and Virtual-Citizens fairness metrics (P90, Gini, complaint rate). Right: persona-conditioned acceptability scores from an evidence-restricted LLM evaluator (proxy), with fixed-label reasons (no free-form text). Arrows link each reason category (e.g., Equity↑, Concentration↑, Budget ok, Risk improves) to the corresponding measurable evidence fields, illustrating the evidence-only interface. Bottom: audit/reproducibility controls (evidence-only JSON input, temperature= 0, cache by hash key) enforcing no free-form evidence.

number of vetoed plans. Because the LLM is evidence-restricted to the plan-summary JSON, the cited reasons map directly to measurable fields (e.g., simultaneous closures, high $P90$ burden, or inequitable allocation across neighborhoods). A particularly informative pattern is a plan with low mean disruption but high tail burden and Gini; such plans are often rejected by resident-advocate personas for perceived unfairness, which aligns with the Virtual-Citizens metrics. We include one representative example in Figure 4.

4.3.3. Ablation: What Each Module Adds

The ablation study isolates when each component is practically meaningful. Virtual Citizens explicitly optimizes tail and inequality, making "low-average but unfair" solutions avoidable. The LLM evaluator introduces an orthogonal axis—*multi-persona consensus feasibility*—which cannot be recovered by engineering proxies alone; cross-evaluation shows that optimizing $J_{\text{LLM,min}}$ and disagreement increases CFR_θ and reduces polarization with modest engineering trade-offs. The fairness-aware baseline ($b^\star$) improves inequality but does not directly quantify *veto risk* or *disagreement*, highlighting why acceptability signals are not reducible to fairness objectives alone.

5. Conclusion

This paper addressed a persistent gap in digital-twin-enabled infrastructure management: although sensing and simulation increase visibility and multi-objective optimization can generate Pareto-efficient maintenance programs, real adoption is often constrained by explainability, distributional burdens, and heterogeneous stakeholder acceptability. We proposed a unified *decision-ready* framework that couples constrained multi-objective optimization with (i) a *Virtual Citizens* module that quantifies tail disruption and inequality and (ii) an *evidence-restricted* LLM-based evaluator that scores multi-persona acceptability and disagreement from *evidence-only plan-summary JSON* artifacts.

Across experiments, introducing fairness- and acceptability-aware objectives reshaped the Pareto frontier beyond classical cost–risk–disruption trade-offs, surfacing plans that are similar in engineering efficiency yet meaningfully improved in distributional burden and veto-robustness. Importantly, our use of "decision-ready" does not claim agency-calibrated predictive accuracy; rather, it denotes the availability of *auditable evidence traces and standardized plan artifacts* that make trade-offs, burdens, and potential veto risks reviewable in committee-style workflows. By keeping constraints explicit and restricting the LLM to structured evidence (with reproducibility controls such as deterministic decoding and caching), the framework supports accountable deliberation in safety-critical, socially sensitive infrastructure planning. *We provide auditable plan artifacts that support deliberation and accountability, not just Pareto sets.*

Acknowledgements

This work was supported by a grant from The Taisei Foundation.

References

Argota Sánchez-Vaquerizo, J. Getting real: The challenge of building and validating a large-scale digital twin of barcelona's traffic with empirical data. *ISPRS International Journal of Geo-Information*, 11(1):24, 2022.

Batty, M. Digital twins. *Environment and Planning B: Urban Analytics and City Science*, 2018.

Bell, K. and Reed, M. The tree of participation: a new model for inclusive decision-making. *Community Development Journal*, 57(4):595–614, 2022. doi: 10.1093/cdj/bsab018. URL https://academic.oup.com/cdj/article/57/4/595/6294808.

Brandt, F., Conitzer, V., Endriss, U., Lang, J., and Procaccia, A. D. Introduction to computational social choice. In *Handbook of Computational Social Choice*, pp. 1–20. Cambridge University Press, 2016. doi: 10.1017/CBO9781107446984.002.

Branke, J., Greco, S., Słowiński, R., and Zielniewicz, P. Learning value functions in interactive evolutionary multiobjective optimization. *IEEE Transactions on Evolutionary Computation*, 19(1):88–102, 2015. doi: 10.1109/TEVC.2014.2303783. URL https://ieeexplore.ieee.org/document/6729055.

Choi, S. L., Jain, R., Emami, P., Wadsack, K., Ding, F., Sun, H., Gruchalla, K., Hong, J., Zhang, H., Zhu, X., and Kroposki, B. egridgpt: Trustworthy ai in the control room. Technical Report NREL/TP-5D00-87440, National Renewable Energy Laboratory, Golden, CO, May 2024. URL https://www.nrel.gov/docs/fy24osti/87440.pdf.

Federal Highway Administration (FHWA). Focus on bridge preservation (unit-cost ranges for deck repair/rehab/replacement). FHWA Office of Operations, 2011. Summarizes representative unit costs for repair/rehabilitation/replacement used for calibration/ordering.

Federal Highway Administration (FHWA). Bridge replacement unit costs: Rehabilitation cost as a fraction of replacement (technical directive sd-2019). FHWA, 2019. Provides representative rehab-to-replacement relationship for cost anchoring.

Federal Highway Administration (FHWA). Bridge replacement unit costs (technical directive / nbi guidance; unit costs normalized by deck area). FHWA, 2020. Accessed via FHWA public guidance on replacement unit costs and deck-area normalization.

Hu, L., Liu, Z., Zhao, Z., Hou, L., Nie, L., and Li, J. A survey of knowledge enhanced pre-trained language models. *IEEE Transactions on Knowledge and Data Engineering*, 2023.

Jaafaru, M. and Agbelie, B. Bridge maintenance planning framework using machine learning, multi-attribute utility theory and evolutionary optimization models. *Automation in Construction*, 141:104460, 2022. doi: 10.1016/j.autcon.2022.104460.

Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., and Yih, W.-t. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*, 2020.

Kitchin, R. The real-time city? big data and smart urbanism. *Big Data & Society*, 1(1), 2014. doi: 10.1177/2053951714528481.

Kizilkaya, F. E. and Kempe, D. k-approval veto: A spectrum of voting rules balancing metric distortion and minority protection. *CoRR*, abs/2507.17981, 2025. doi: 10.48550/arXiv.2507.17981. URL https://arxiv.org/abs/2507.17981.

Kušić, K., Schumann, R., and Ivanjko, E. A digital twin in transportation: Real-time synergy of traffic data streams and simulation for virtualizing motorway dynamics. *Advanced Engineering Informatics*, 55:101858, 2023.

Lárraga, G. and Miettinen, K. Survey of interactive evolutionary decomposition-based multiobjective optimization methods. *Evolutionary Computation*, pp. 1–39, 2025. doi: 10.1162/evco_a_00366.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

Liu, X., Liu, Y., Sun, Z., Wang, B., and Zhao, Y. Dynamic optimization of maintenance strategy for bridges in regional transportation network through semi-markov processes. *Automation in Construction*, 174:106175, 2025. doi: 10.1016/j.autcon.2025.106175. URL https://doi.org/10.1016/j.autcon.2025.106175.

New Hampshire Department of Transportation. Bridge cost estimate (example of itemized preservation/repair bid items for bridge decks). NHDOT public estimate sheet (PDF), 2024. Used as auditable example evidence for minor repair/preservation pay items.

Rask, M. and Shin, B. Integrating paths: Enhancing deliberative democracy through collective intelligence insights. *Societies*, 14(12):270, 2024. doi: 10.3390/soc14120270. URL https://www.mdpi.com/2075-4698/14/12/270.

Shih, H.-S. and Olson, D. L. *TOPSIS and its Extensions: A Distance-Based MCDM Approach*. Springer, 2022. doi: 10.1007/978-3-031-09577-1. URL https://link.springer.com/book/10.1007/978-3-031-09577-1.

Shinohara, T., J-katagiri, and Saomoto, H. LLM-powered report-driven markov modelling for large-scale predictive bridge maintenance in japan. In *UrbanAI: Harnessing Artificial Intelligence for Smart Cities*, 2025. URL https://openreview.net/forum?id=NE2uiXUAqn.

Vineis, V., Perelli, G., and Tolomei, G. Beyond predictions: A participatory framework for multi-stakeholder decision-making, 2025. URL https://arxiv.org/abs/2502.08542.

Wang, H., Olhofer, M., and Jin, Y. A mini-review on preference modeling and articulation in multi-objective optimization: current status and challenges. *Complex & Intelligent Systems*, 3:233–245, 2017. doi: 10.1007/s40747-017-0053-9. URL https://link.springer.com/article/10.1007/s40747-017-0053-9.

Wang, S., Chen, X., Yang, J., Zhang, L., Hong, W., Diao, W., Wang, W., Chen, W., and Jiao, L. Knowledge graph-driven bridge maintenance decision-making via retrieval-augmented chain-of-thought prompting and large language models. *Automation in Construction*, 177:106632, 2026a. doi: 10.1016/j.autcon.2025.106632.

Wang, Y., Xiong, W., Zhu, Y., and Cai, C. S. Knowledge graph-driven bridge maintenance decision-making via integrating large language models and chain-of-thought reasoning. *Automation in Construction*, 181:106632, 2026b. doi: 10.1016/j.autcon.2025.106632. URL https://doi.org/10.1016/j.autcon.2025.106632.

Wright, L. and Davidson, S. How to tell the difference between a model and a digital twin. *Advanced Modeling and Simulation in Engineering Sciences*, 7:1–13, 2020.

Xu, H., Berres, A., Yoginath, S. B., Sorensen, H., Nugent, P. J., Severino, J., Tennille, S. A., Moore, A., Jones, W., and Sanyal, J. Smart mobility in the cloud: Enabling real-time situational awareness and cyber-physical control through a digital twin for traffic. *IEEE Transactions on Intelligent Transportation Systems*, 24(3):3145–3156, 2023.

Zhang, Y., Lin, Y., Zheng, G., Liu, Y., et al. Metacity: Data-driven sustainable development of complex cities. *The Innovation*, 2025. doi: 10.1016/j.xinn.2024.100775.

Supplementary Material for Decision-Ready Bridge Maintenance Planning with a Lightweight Digital Twin, Synthetic Citizens, and Evidence-Restricted LLM Consensus Proxies

S1. Proposed Method (Extended Details)

S1.1. Positioning against alternative approaches

Table 4 summarizes how our approach differs from interactive preference-based MOO, post-hoc MCDA, and social-choice-style aggregation, highlighting our focus on (i) explicit multi-persona disagreement and veto robustness, and (ii) auditability via evidence-restricted evaluation.

Table 4. Positioning of our framework relative to common alternatives for preference handling and collective decision support.

Approach family	**Strengths**	**Limitations in multi-stakeholder public decisions**	**Typical operational burden**
A posteriori MOO + post-hoc MCDA (e.g., TOPSIS) (Shih & Olson, 2022)	Simple ranking from weights; easy to implement	Weights may hide value conflicts; limited treatment of disagreement/veto; fairness often implicit	Low (one-shot scoring), but requires agreed weights
Preference-based / interactive MOO (Wang et al., 2017; Branke et al., 2015; Lárraga & Miettinen, 2025)	Targets region of interest; can reduce Pareto overload	Often assumes a single DM or unified preference; repeated interactions may be costly in committees	Medium–High (iterative elicitation / learning loop)
Social choice / voting-style aggregation (Brandt et al., 2016; Kizilkaya & Kempe, 2025)	Explicitly studies collective choice; can model veto/minority protection	Typically assumes direct ballots; may lack evidence-grounded audit trail and engineering constraints	Medium (requires preference reporting / ballots)
This work: constrained MOO + fairness + evidence-restricted LLM evaluation	Optimizes engineering + fairness + veto robustness; quantifies disagreement; evidence-auditable artifacts	LLM is a proxy evaluator and requires robustness governance; depends on summary-field design	Medium (LLM scoring mitigated by deterministic decoding/caching)

S1.2. Per-Bridge Intervention Cost Estimation and Calibration

Purpose and scope. This work focuses on *decision-ready* portfolio planning under annual budgets. Accordingly, we use an intervention *agency-cost* model that is (i) lightweight enough to be evaluated thousands of times inside NSGA-II, (ii) auditable (explicit unit-cost assumptions), and (iii) calibratable to agency bid-tab databases. The model targets direct repair/rehabilitation/replacement costs and does not include user costs (which are handled separately through disruption objectives).

Cost proxy used in optimization. Each bridge i is assigned at most one intervention $a_i \in \{\texttt{none}, \texttt{minor}, \texttt{major}, \texttt{replace}\}$ scheduled at year $\tau_i \in \{1, \dots, T\}$. We compute the annual spending as

$$C_t(\mathbf{x}) = \sum_{i:\ \tau_i = t} c_i(a_i), \qquad t = 1, \dots, T, \tag{8}$$

and enforce feasibility via $C_t(\mathbf{x}) \leq B_t$. Following the main text, we use a linear size-scaled cost proxy

$$c_i(a) = u(a) \cdot \frac{A_i}{10^4}, \tag{9}$$

where A_i is a bridge size proxy (deck area by default), and $u(a)$ are action-type multipliers. The normalization by 10^4 keeps objective magnitudes in a numerically convenient range.

Bridge size proxy (deck-area scaling). We use deck area as the default size proxy,

$$A_i = L_i \cdot W_i, \tag{10}$$

where L_i and W_i are available from standard bridge inventory attributes (e.g., NBI length/width fields). Deck-area scaling is also consistent with common practice in reporting bridge unit costs, where costs are normalized by total deck area to enable cross-project comparability (Federal Highway Administration (FHWA), 2020). If element-level quantities are available, the same linear form can be instantiated using aggregated quantities (e.g., total element quantities) instead of A_i, while preserving the calibration procedure below.

Unit-cost anchors from public guidance (replacement and rehabilitation). To ground the magnitude of $u(\texttt{replace})$ and $u(\texttt{major})$ in widely cited public data, we use FHWA guidance reporting bridge replacement unit costs in \$/ft^2 (deck area) and a rehabilitation-to-replacement ratio. Specifically, FHWA defines replacement unit costs by dividing total project cost by total deck area and reports national/state-level unit-cost values (Federal Highway Administration (FHWA), 2020). The same guidance provides a representative rehabilitation cost level as a fraction of replacement cost (e.g., rehabilitation as a fixed proportion of replacement) (Federal Highway Administration (FHWA), 2019). These references provide a principled anchor for mapping our normalized proxy (Equation 9) to physical currency units.

Minor vs. major actions (preservation/repair/rehabilitation). To support the relative ordering $u(\texttt{minor}) \ll u(\texttt{major}) < u(\texttt{replace})$, we rely on published tables of bridge repair/rehabilitation/replacement unit costs and agency cost-estimate documents. For example, FHWA bridge preservation guidance summarizes representative unit costs for deck repair, deck rehabilitation, and deck replacement (by region) (Federal Highway Administration (FHWA), 2011). Moreover, DOT cost-estimate sheets provide auditable itemized evidence that preservation/repair actions (e.g., deck crack sealing, surface treatments) occur as distinct pay items at substantially smaller magnitudes than full replacement contracts (New Hampshire Department of Transportation, 2024). In practice, we recommend calibrating $u(\texttt{minor})$ from a library of preservation/repair bid items (e.g., sealants, minor patching, localized repairs) and calibrating $u(\texttt{major})$ from rehabilitation-class contracts (e.g., deck rehab/superstructure rehab), both normalized by deck area.

Calibration recipe to agency bid tabs. Let $\widehat{U}(a)$ denote an empirically estimated unit cost for action type a in \$/m^2 (or \$/ft^2), obtained from historical projects by (i) categorizing contracts into action classes (minor/major/replace), (ii) extracting total costs attributable to the class, and (iii) normalizing by total deck area. We then map to our normalized coefficients as

$$u(a) = \frac{\widehat{U}(a)}{U_0}, \tag{11}$$

where U_0 is a chosen scale (e.g., $U_0 = \widehat{U}(\texttt{replace})$ so that $u(\texttt{replace}) \approx 1$), and optionally apply region/structure-type multipliers (steel vs. concrete, urban vs. rural) if such stratification is supported by the data.

Sensitivity and reporting. Because unit costs vary across states, time, and scope, we recommend reporting (i) the chosen $u(a)$ values, (ii) the underlying unit-cost source(s), and (iii) a sensitivity sweep where $u(a)$ is perturbed within plausible ranges inferred from public guidance and/or local bid tabs. This makes the cost model auditable and ensures that qualitative conclusions (e.g., frontier reshaping by fairness/acceptability objectives) are not artifacts of a single cost table.

S1.3. Bridge Digital Twin Simulator

Given a plan $\mathbf{x}$ (decoded from $\mathbf{g}$), we simulate year-by-year evolution of each bridge's condition distribution. The simulator is intentionally lightweight: it is fast enough to be called thousands of times inside NSGA-II, yet expressive enough to generate auditable evidence (trajectories of costs, risks, and disruptions).

S1.3.1. RISK PROXY AND RISK-COUPLED DETERIORATION

We define a scalar risk proxy from a state distribution $\mathbf{p}$:

$$\rho(\mathbf{p}) = \sum_{k=1}^{4} w_k p^{(k)}, \qquad \mathbf{w} = [0,\ 0.1,\ 0.5,\ 1.0], \tag{12}$$

which increases as probability mass shifts to worse condition states. In the prototype implementation, a bridge-specific baseline deterioration rate is computed from interpretable covariates (e.g., age and ADT), and then coupled to current risk:

$$\lambda_{i,t} = \mathrm{clip}\big(\bar{\lambda}_{i,t} + 0.8\,\rho(\mathbf{p}_{i,t}),\ 0,\ 1\big), \tag{13}$$

where $\bar{\lambda}_{i,t}$ is the baseline (e.g., increasing with age and traffic) and the coupling term captures accelerating deterioration in already-degraded states. This coupling mirrors the implementation used to keep the model simple yet behaviorally plausible.

S1.3.2. MONOTONE MARKOV TRANSITION

We model deterioration as a monotone Markov transition (no spontaneous improvement):

$$\mathbf{p}_{i,t+1} = \mathbf{p}_{i,t}\,\mathbf{P}(\lambda_{i,t})\,, \tag{14}$$

where $\mathbf{P}(\lambda)$ is upper-bidiagonal (CS4 is absorbing):

$$\mathbf{P}(\lambda) = \begin{bmatrix} 1 - r_{12}(\lambda) & r_{12}(\lambda) & 0 & 0 \\ 0 & 1 - r_{23}(\lambda) & r_{23}(\lambda) & 0 \\ 0 & 0 & 1 - r_{34}(\lambda) & r_{34}(\lambda) \\ 0 & 0 & 0 & 1 \end{bmatrix}. \tag{15}$$

Transition rates are clipped to remain probabilistically valid and to avoid unrealistically fast transitions. A simple parameterization consistent with the code is:

$$\begin{aligned} r_{12}(\lambda) &= \min(0.25,\ 0.03 + 0.12\lambda), \\ r_{23}(\lambda) &= \min(0.25,\ 0.02 + 0.10\lambda), \\ r_{34}(\lambda) &= \min(0.30,\ 0.02 + 0.14\lambda). \end{aligned} \tag{16}$$

This model is not intended to replace high-fidelity deterioration models; rather, it provides a consistent, computationally efficient mapping from interventions to risk trajectories, enabling large-scale exploratory search.

S1.3.3. INTERVENTION OPERATOR

If bridge i is maintained at year $t = \tau_i$ with $a_i \neq \texttt{none}$, we apply an instantaneous improvement operator before deterioration:

$$\tilde{\mathbf{p}}_{i,t} = \mathcal{A}_{a_i}(\mathbf{p}_{i,t}), \qquad \mathbf{p}_{i,t+1} = \tilde{\mathbf{p}}_{i,t}\,\mathbf{P}(\lambda_{i,t}). \tag{17}$$

We use linear mass-transfer rules that approximate typical rehabilitation effects:

$$\begin{aligned} \mathcal{A}_{\texttt{minor}} : & \ p^{(3)} \leftarrow p^{(3)} - 0.25p^{(3)},\ \ p^{(2)} \leftarrow p^{(2)} + 0.25p^{(3)}, \\ & \ p^{(2)} \leftarrow p^{(2)} - 0.20p^{(2)},\ \ p^{(1)} \leftarrow p^{(1)} + 0.20p^{(2)}, \end{aligned} \tag{18}$$

$$\begin{aligned} \mathcal{A}_{\texttt{major}} : & \ p^{(4)} \leftarrow p^{(4)} - 0.40p^{(4)},\ \ p^{(3)} \leftarrow p^{(3)} + 0.40p^{(4)}, \\ & \ p^{(3)} \leftarrow p^{(3)} - 0.50p^{(3)},\ \ p^{(2)} \leftarrow p^{(2)} + 0.50p^{(3)}, \\ & \ p^{(2)} \leftarrow p^{(2)} - 0.40p^{(2)},\ \ p^{(1)} \leftarrow p^{(1)} + 0.40p^{(2)}, \end{aligned} \tag{19}$$

$$\mathcal{A}_{\texttt{replace}} : \mathbf{p} \leftarrow [1, 0, 0, 0]. \tag{20}$$

S2. Experimental Results (Additional Details)

S2.1. Metrics

Engineering efficiency. We report (i) hypervolume (HV) of the Pareto set after objective normalization and a common reference point, (ii) feasibility rate under annual budgets, and (iii) risk/disruption statistics at matched cost or matched risk.

Algorithm 1 Pareto Search with Virtual Citizens and LLM Stakeholders

Initialize population $\{\mathbf{g}^{(j)}\}_{j=1}^{P}$ with sparse-biased integer sampling.
for gen $= 1$ **to** G **do**
  for all individuals $\mathbf{g}^{(j)}$ **do**
    Decode $\mathbf{g}^{(j)} \mapsto \mathbf{x}^{(j)}$ (action type and execution year per bridge).
    Simulate dynamics to compute objectives and annual-budget constraint violations.
    if virtual citizens enabled **then**
      Compute distributional metrics (e.g., $P90$, Gini, complaint rate).
    end if
    if LLM evaluation enabled **then**
      Build evidence-only summary $s(\mathbf{x}^{(j)})$ (plan-summary JSON) and query personas.
      Cache LLM outputs keyed by a hash of $s(\mathbf{x}^{(j)})$ to control cost and improve reproducibility.
      Aggregate persona scores (worst/mean/disagreement) into minimization objectives.
    end if
  end for
  Apply NSGA-II selection (nondominated sorting + crowding) with constraint handling.
  Generate offspring via crossover/mutation on integer genomes with clipping.
end for
return nondominated plans with accompanying evidence artifacts.

Consensus-Feasibility and social-legitimacy metrics. We define **Consensus-Feasibility Rate** (CFR) as the fraction of feasible solutions that avoid a veto:

$$\mathrm{CFR}_\theta \;=\; \frac{1}{|\mathcal{P}|}\sum_{\mathbf{x}\in\mathcal{P}} \mathbb{I}[A_{\min}(\mathbf{x}) \geq \theta]\,, \tag{21}$$

where $\mathcal{P}$ is the feasible nondominated set and $A_{\min}(\mathbf{x}) = \min_k \texttt{accept}_k(\mathbf{x})$. We report CFR_θ for a small set of thresholds (e.g., $\theta \in \{0.4, 0.6, 0.8\}$). We also report distributional fairness metrics (e.g., J_{p90}, J_{gini}, complaint rate) to quantify "low-average but unfair" failure modes.

Operational impact proxies (justification and review burden). To reflect justification overhead in decision workflows, we report **Justification completeness**: the fraction of required evidence fields present in the rendered plan card (checklist coverage). We also report a **review burden proxy**: the number of candidate plans that must be inspected to find at least one no-veto solution (e.g., $A_{\min}(\mathbf{x}) \geq \theta$), approximating committee workload.

Acceptance cost (trade-off quantification). To make the main claim decision-relevant, we quantify the minimum engineering sacrifice required to achieve consensus: for a target threshold θ, define the **Acceptance Cost Premium** as

$$\Delta J_{\mathrm{cost}}(\theta) = \min_{\mathbf{x}:\, A_{\min}(\mathbf{x})\geq\theta} J_{\mathrm{cost}}(\mathbf{x}) \;-\; \min_{\mathbf{x}} J_{\mathrm{cost}}(\mathbf{x}), \tag{22}$$

(and analogously at matched risk/disruption). This directly supports statements like: *"a small cost increase yields a large reduction in disagreement and inequity."*

S2.2. Implementation

Planning horizon, actions, and constraints. We plan over a discrete horizon of T years (default $T = 10$ in the prototype). Each bridge i is assigned at most one intervention action and its scheduled year; annual budget feasibility is enforced via inequality constraints applied to each year (e.g., $g_t(\mathbf{x}) \leq 0$ for $t = 1, \ldots, T$), where g_t measures overspending relative to the annual budget.

Multi-objective optimization. We use NSGA-II to approximate the Pareto set under the objectives defined in the main paper. The default optimizer configuration uses population size $P = 40$ and $G = 20$ generations, with integer-aware sampling and variation operators. Each candidate plan is evaluated by simulating yearly deterioration and applying intervention effects, producing discounted objective traces. Discounting uses a constant annual discount rate (default 0.03).

Virtual Citizens (distributional disruption). We instantiate a population of synthetic citizens (default $n_c = 800$), each representing an abstract traveler exposed to bridge closures. Each citizen is assigned a type (e.g., commuter-, logistics-, vulnerability-oriented archetypes) that determines trip intensity and tolerance. Citizens are spatially anchored by sampling "home" locations near the bridge set and adding geographic jitter, producing heterogeneous proximity without requiring mobility traces.

Exposure is operationalized by assigning each citizen j a sparse set of top-k bridges (default $k = 6$). The affinity between citizen j and bridge ℓ is modeled as:

$$w_{j\ell} \;\propto\; \big(\log(1 + \mathrm{ADT}_\ell)\big)^{\alpha} \big(d_{j\ell} + 1\big)^{-\beta}, \tag{23}$$

with default parameters $\alpha = 0.6$ and $\beta = 1.4$, where $d_{j\ell}$ is the great-circle distance between citizen j and bridge ℓ. Normalizing $w_{j\ell}$ yields per-citizen bridge-use probabilities over the selected top-k bridges. Given a plan, closures induce added travel time that is accumulated per citizen across the horizon (with discounting). From the per-citizen burden distribution, we compute tail (e.g., $\mathrm{CVaR}_{0.9}$ or $P90$), inequality (Gini), and complaint signals (hard rate above tolerance and a soft complaint score), which are reported and optionally included as optimization objectives.

LLM-based virtual stakeholder evaluator. We incorporate an LLM as a structured evaluator that scores a plan from multiple stakeholder personas. The key design choice is to restrict the LLM to evidence provided in a machine-readable plan summary JSON, mitigating hallucinations and improving auditability. For each plan, the JSON includes budget compliance signals, cost/risk/disruption aggregates, and (when enabled) Virtual Citizens statistics such as Gini, complaint rate/score, and tail impacts. The LLM is prompted to output *only* JSON persona scores (no prose). We use deterministic decoding (temperature $= 0$) and cache responses keyed by a hash of the plan summary and persona specification, enabling efficient reuse and reproducibility across runs. Persona-level outputs are aggregated into robustness and disagreement indicators (e.g., worst-case acceptability, mean acceptability, and dispersion).

Figure 5 summarizes the evidence-restricted evaluation interface: an evidence-only plan-summary JSON is provided to the evaluator, persona-conditioned scores are returned in structured JSON, and the outputs are aggregated into acceptability and disagreement indicators suitable for audit and reproducibility.

Ethics and privacy note. The experiments use only public infrastructure records and fully synthetic citizen agents; no personally identifiable mobility traces or human subject data are collected. The LLM is used as an evaluator conditioned on structured evidence summaries, and limitations and potential biases are discussed in the main paper.

S2.3. Discussion and Limitations

Figure 6 illustrates how Virtual Citizens and LLM-based evaluation expose decision-legitimacy trade-offs that may be hidden under aggregate engineering objectives—e.g., low average disruption can still correspond to high tail burden/inequality, and veto-robust acceptability can vary substantially across Pareto-efficient plans. In practice, these signals help surface "solutions exist but do not pass" failure modes and support deliberation by making equity and veto risks explicit alongside traditional cost–risk trade-offs.

Robustness We treat the LLM evaluator as a decision-support proxy rather than ground truth. To assess stability, we re-evaluate a subset of solutions under (i) small perturbations in persona wording and (ii) repeated scoring with identical evidence (enabled by caching and deterministic decoding), and report variance of persona scores and CFR_θ (Supplementary). *To assess robustness to the lightweight simulator assumptions, we also perform a parameter perturbation sensitivity analysis (e.g., deterioration rates and intervention effects) and verify that the relative CFR_θ improvements of the proposed variants are preserved (Supplementary).* We also report computational overhead: Virtual Citizens cost scales with M and sparsity k, while LLM calls dominate; caching and the evidence-only JSON interface mitigate both cost and reproducibility concerns.

Decision support, not decision replacement. A central design choice in our framework is to preserve hard operational constraints and auditable engineering signals, while using generative AI only to *evaluate* the social feasibility of candidate plans. Accordingly, the LLM outputs should be interpreted as decision-support indicators (e.g., potential veto risk, disagreement, or perceived unfairness), not as normative ground truth. Final decisions remain with agencies and stakeholder processes, which can contest, override, or refine any plan.

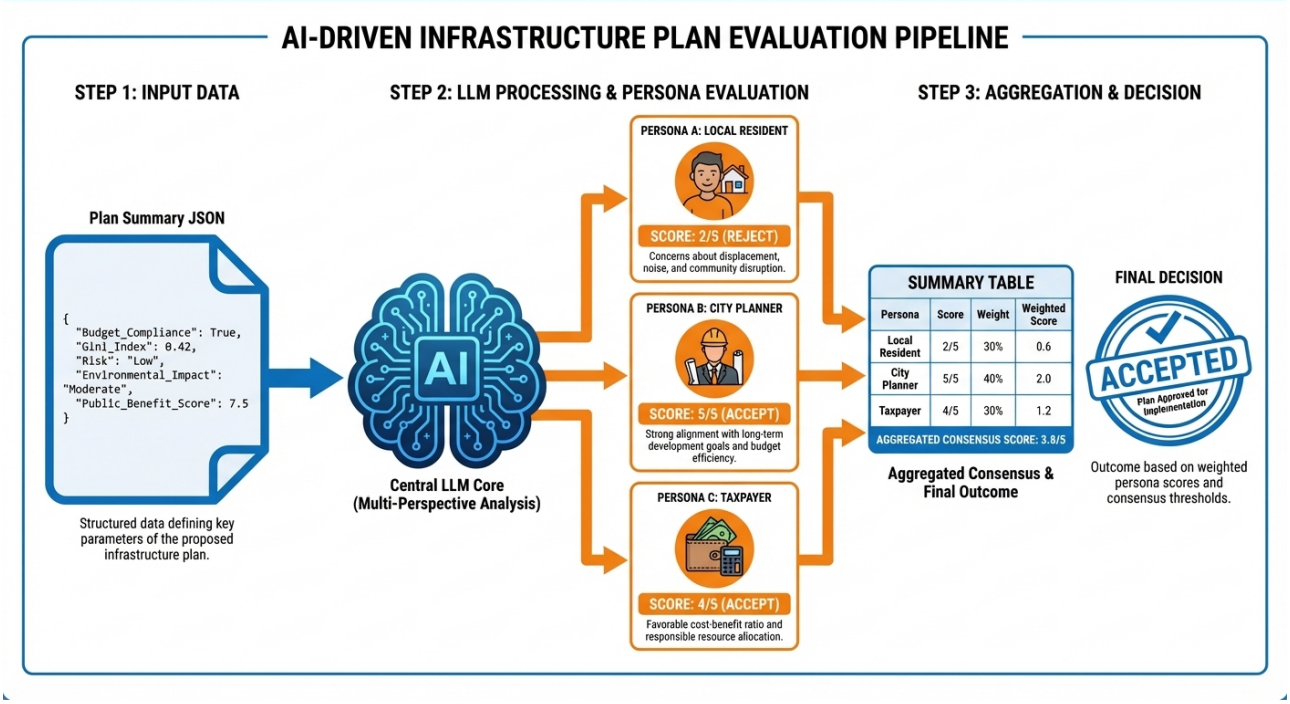

Figure 5. LLM-based multi-stakeholder evaluation: evidence-only plan summary JSON, persona-conditioned scoring, and aggregation.

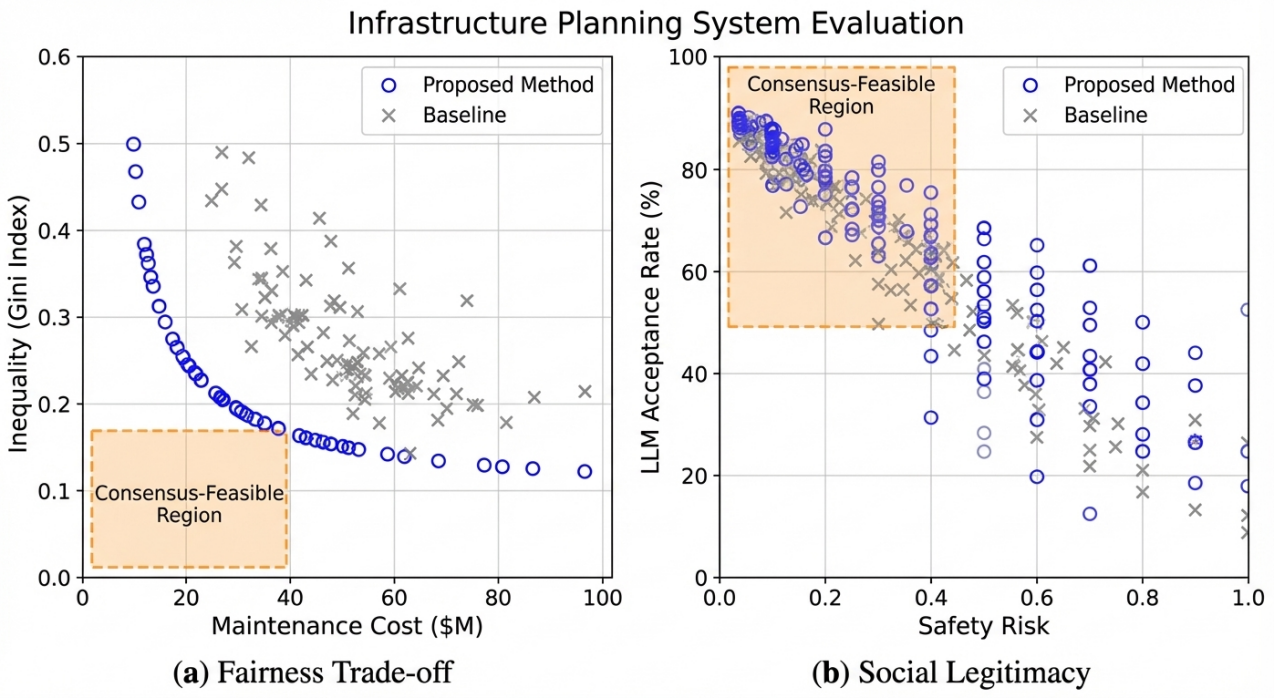

Figure 6. Decision-legitimacy trade-offs enabled by Virtual Citizens and LLM evaluation.

Ethical considerations and bias risks. LLM-based evaluation can introduce biases stemming from training data, persona specification, and prompt framing. If left unmanaged, such biases could systematically favor certain communities or narratives, thereby affecting resource allocation in ways that conflict with public values. We mitigate these risks by (i) restricting the LLM to an *evidence-only plan-summary JSON* to reduce hallucinations and hidden assumptions, (ii) explicitly

representing value pluralism through multiple personas, and (iii) exposing disagreement as a measurable output rather than collapsing to a single score. Nevertheless, careful governance is required in deployment: persona definitions should be documented and reviewed, sensitive attributes must be avoided or handled via approved fairness policies, and auditing should include stress tests under alternative persona sets and prompts.

Limitations. First, the digital-twin simulator is intentionally lightweight to enable thousands of evaluations inside NSGA-II. While this supports scalable Pareto search and transparent traces, it simplifies deterioration dynamics, work-zone interactions, and network traffic assignment; thus, absolute performance values should not be interpreted as agency-calibrated forecasts. Second, the disruption proxy relies on localized OSM-based detours (with conservative caps) and does not model system-wide rerouting equilibrium. Third, despite deterministic decoding and caching, LLM judgments may remain sensitive to prompt wording, persona templates, and summary-field choices. Finally, the current optimization encoding assumes at most one intervention per bridge; richer life-cycle policies (e.g., multiple actions, bundling, or crew/resource constraints) are not yet represented.

Future work. Promising directions include substituting calibrated semi-Markov deterioration models and validated traffic/assignment modules, adding explicit operational constraints (e.g., bundling, work-zone capacity, equity constraints by region), and improving LLM robustness via repeated scoring, uncertainty estimates, and human-in-the-loop validation on a small set of plans. Algorithmically, staged evaluation and surrogate modeling can reduce LLM calls by scoring only filtered frontier candidates, enabling larger portfolios.

S2.4. Ethics Statement.

Privacy. Our experiments use only public infrastructure records and fully simulated *Virtual Citizens*; we do not collect, infer, or process any personally identifiable information, individual mobility traces, or sensitive attributes. The Virtual Citizens model is used solely to estimate distributional burden (e.g., tail delay and inequality) without tracking real individuals.

Bias and limitations. The LLM-based evaluation may reflect biases from model pretraining, persona specification, and prompt framing. Accordingly, acceptability scores should be interpreted as decision-support signals (e.g., potential veto risk), not as normative ground truth.

Mitigation and governance. We mitigate risks by restricting the LLM to an *evidence-only plan-summary JSON* and requiring structured score outputs, which improves auditability and reduces hallucination pathways. We further promote reproducibility via deterministic decoding and caching, treat personas as controlled experimental variables, and report disagreement/dispersion metrics alongside aggregate acceptability to avoid collapsing value pluralism into a single score. In deployment, persona sets and evaluation templates should be documented, reviewed by domain stakeholders, and stress-tested under alternative persona/prompt variants, with human decision makers retaining final authority.