

Reasoning Is All You Need for Urban Planning AI

Sijie Yang¹, Jiatong Li^{1,2}, Filip Biljecki^{1,3,*}

¹Department of Architecture, National University of Singapore

²School of Architecture, Tsinghua University

³Department of Real Estate, National University of Singapore

*Corresponding author: filip@nus.edu.sg

Abstract

AI has proven highly successful at urban planning *analysis*—learning patterns from data to predict future conditions. The next frontier is AI-assisted *decision-making*: agents that recommend sites, allocate resources, and evaluate trade-offs while reasoning transparently about constraints and stakeholder values. Recent breakthroughs in reasoning AI—CoT¹ prompting, ReAct², and multi-agent collaboration frameworks—now make this vision achievable.

This position paper presents the Agentic Urban Planning AI Framework for reasoning-capable planning agents that integrates three cognitive layers (Perception, Foundation, Reasoning) with six logic components (Analysis, Generation, Verification, Evaluation, Collaboration, Decision) through a multi-agents collaboration framework. We demonstrate why planning decisions require explicit reasoning capabilities that are value-based (applying normative principles), rule-grounded (guaranteeing constraint satisfaction), and explainable (generating transparent justifications)—requirements that statistical learning alone cannot fulfill. We compare reasoning agents with statistical learning, present a comprehensive architecture with benchmark evaluation metrics, and outline critical research challenges. This framework shows how AI agents can augment human planners by systematically exploring solution spaces, verifying regulatory compliance, and deliberating over trade-offs transparently—not replacing human judgment but amplifying it with computational reasoning capabilities.

By 2050, 68% of humanity will live in cities (UN 2019). AI has transformed urban planning analytics (Jha et al. 2021; Sanchez et al. 2023; Peng et al. 2024; Wang et al. 2025b), achieving unprecedented accuracy in prediction tasks. Yet a critical question emerges as AI capabilities advance: *Can statistical learning alone support planning decisions, or do we need explicit reasoning capabilities?* We argue that planning decisions demand reasoning agents that are **value-based**, **rule-grounded**, and **explainable**—capabilities that statistical pattern learning alone cannot provide.

¹Chain-of-Thought: A prompting technique that encourages large language models to break down complex reasoning into intermediate steps (Wei et al. 2023)

²Reasoning and Acting: A framework that interleaves reasoning traces with task-specific actions in language models (Yao et al. 2023b)

Our contributions: (1) We compare reasoning agents with statistical learning, demonstrating why explicit reasoning is foundational for planning decisions; (2) We present the *Agentic Urban Planning AI Framework*—a three-layer cognitive architecture (Perception, Foundation, Reasoning) integrating six logic components (Analysis, Generation, Verification, Evaluation, Collaboration, Decision) through a multi-agents collaboration framework, formalised with algorithms and evaluation metrics; (3) We outline five critical research challenges and a path forward for building reasoning-capable planning agents that augment human judgment with computational reasoning capabilities.

AI for UP: From Analytics to Decision Support

AI’s success in urban planning analytics is undeniable—predicting traffic with RNNs³ (Lv et al. 2018), classifying land uses with CNNs⁴ (Zhang et al. 2018), forecasting building carbon emissions in cities with GNNs⁵ (Yap et al. 2025), assessing urban heat islands (Xu et al. 2022), evaluating thermal comfort (Yang et al. 2025b), analysing urban morphology and street environment impacts (Qiu et al. 2022; Yang et al. 2023), identifying street activity patterns (Li, Ma, and Lai 2025), and developing liveability indices (Lei et al. 2025). These *pattern-based*, *data-driven* GeoAI approaches (Liu and Biljecki 2022) excel at learning complex correlations from historical urban data—predicting *what will happen*.

The frontier now is AI-assisted *decision-making*: recommending sites, allocating resources, evaluating trade-offs (Zhu, Chen, and Zhang 2025; Liu et al. 2025). Recent breakthroughs in reasoning AI make this achievable. CoT prompting (Wei et al. 2023) generates intermediate reasoning steps; ReAct (Yao et al. 2023b) interleaves reasoning with tool-augmented actions; multi-agent collaboration frameworks like AutoGen (Wu et al. 2024a) enable coordinated deliberation among specialized agents. These techniques enable AI agents to deliberate transparently, guarantee regulatory compliance, and generate explainable justifications—capabilities planning decisions demand.

We distinguish two paradigms by *how decisions are made*

³Recurrent Neural Networks

⁴Convolutional Neural Networks

⁵Graph Neural Networks

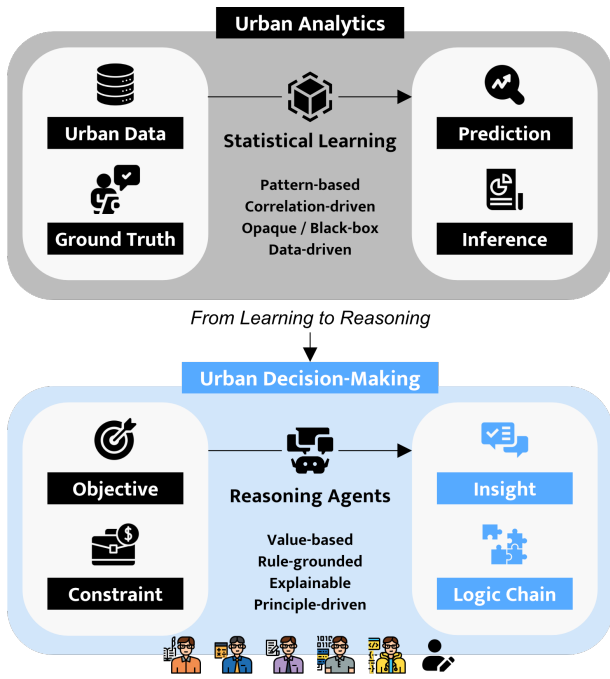


Figure 1: AI’s dual role in urban planning: analysis (prediction tasks) and decision support (recommendation tasks with explicit reasoning).

(Figure 1): **Statistical learning** refers to systems which learn decision patterns from historical data through statistical correlations. These *pattern-based, data-driven* approaches excel at prediction but have opaque decision processes. They can replicate historical allocations, detect likely violations, and recommend solutions—but struggle to apply normative principles, guarantee constraint satisfaction, or explain reasoning chains. **Reasoning Agents** refer to systems which generate explicit reasoning traces to reach decisions, using LLM-based reasoning (CoT, ReAct). These agents can challenge unjust patterns, resolve contradictory rules, and explain counterfactual logic—capabilities statistical learning lacks.

Why do planning decisions specifically require reasoning agents? Table 1 compares both paradigms across nine decision tasks. While statistical learning handles many tasks effectively, reasoning agents provide three critical capabilities planning demands:

Value-Based & Principle-Driven: Planning decisions are *normative*—reflecting values, principles, and long-term visions rather than learned patterns. Consider equity-driven resource allocation: statistical learning replicates historical allocations, but reasoning agents apply equity principles and challenge unjust patterns embedded in historical data. For novel contexts without precedent—climate adaptation, emerging technologies—reasoning agents apply first principles when historical patterns provide insufficient guidance. When competing values conflict, reasoning agents deliberate on normative priorities rather than merely learning stakeholder preferences from past data.

Planning Decision Task	Statistical Learning	Reasoning Agents
Value-Based & Principle-Driven		
<i>Equity-driven resource allocation</i>		
Replicate historical allocation	•	•
Apply equity principles	○	•
Challenge unjust patterns	–	•
<i>Novel urban planning context</i>		
Transfer similar patterns	•	•
Reason from first principles	–	•
<i>Competing value prioritisation</i>		
Learn stakeholder preferences	•	•
Deliberate on normative priorities	–	•
Rule-Grounded		
<i>Zoning regulation compliance</i>		
Detect likely violations	•	•
Guarantee zero violations	–	•
<i>Multi-constraint optimization</i>		
Find feasible solutions	•	•
Verify all constraints satisfied	○	•
<i>Contradictory rule resolution</i>		
Flag conflicting requirements	○	•
Resolve using legal reasoning	–	•
Explainable		
<i>Decision justification to public</i>		
Provide recommendations	•	•
Generate readable rationale	○	•
<i>Causal impact chain explanation</i>		
Predict outcomes	•	•
Trace cause-effect reasoning	–	•
<i>“What-if” scenario analysis</i>		
Simulate alternative outcomes	•	•
Explain counterfactual logic	–	•

Table 1: Paradigm comparison on planning decision tasks. Legend: • = well-supported; ○ = limited; – = not supported. Both paradigms can address planning tasks, but reasoning agents provide value-based deliberation, rule-grounded verification, and explainable justification.

Rule-Grounded: Planning operates under hard constraints that must be satisfied with certainty. For zoning regulation compliance, statistical learning detects likely violations, but reasoning agents guarantee zero violations through formal verification. Multi-constraint optimization requires verifying that *all* constraints are satisfied simultaneously, not just finding feasible solutions. When contradictory rules arise, reasoning agents resolve conflicts using legal reasoning rather than flagging inconsistencies without resolution.

Explainable: Planning decisions require transparent justifications for legal review and public scrutiny. For decision justification to the public, reasoning agents generate readable rationales explaining *why* recommendations were made. Causal impact chain explanation demands tracing cause-effect reasoning, not just predicting outcomes. “What-if” scenario analysis requires explaining counterfactual logic to support deliberative decision processes. The Sidewalk Labs Toronto smart city proposal’s opacity challenges underscore the necessity of explainable AI in planning contexts.

Reasoning-Capable Urban Planning Agents

We now present a comprehensive architecture that integrates contemporary reasoning techniques with symbolic constraint solving to enable transparent, verifiable, and collaborative planning decision support. As illustrated in Figure 2, our framework comprises three cognitive layers and six logic components that operate through human-AI collaborative workflows (Figure 3).

Agentic Urban Planning AI Framework

Table 2 presents a comprehensive agentic urban planning AI framework organised as a three-layer cognitive architecture: **Perception** → **Foundation** → **Reasoning (Agentic AI)**. This framework applies urban planning reasoning with agentic AI—LLM⁶-based autonomous systems capable of perception-grounded data representation, external tool-augmented analysis, and value-aligned decision-making. While perception and foundation layers capture and organise urban knowledge, the reasoning layer embodies goal-directed agents that deliberate, verify, and act upon urban problems. Critically, we position reinforcement learning dually: in the Foundation layer as environment modelling for simulation and behaviour learning, and in the Reasoning layer as policy optimisation for strategic decision-making. Similarly, RAG serves as the “memory interface” of the Foundation layer, providing retrievable urban knowledge to agentic systems, while LLMs constitute the core reasoning engine built upon foundational pretraining. This architecture clarifies that Agentic AI = LLM (Core) + RAG (Memory) + Tools (Action) + RL (Feedback) + Values (Constraint).

Beyond these three cognitive layers, the framework (illustrated in Figure 2) integrates six specialized logic components that orchestrate planning deliberation.

Three Cognitive Layers: The framework progresses through three complementary layers aligned with urban planning stages:

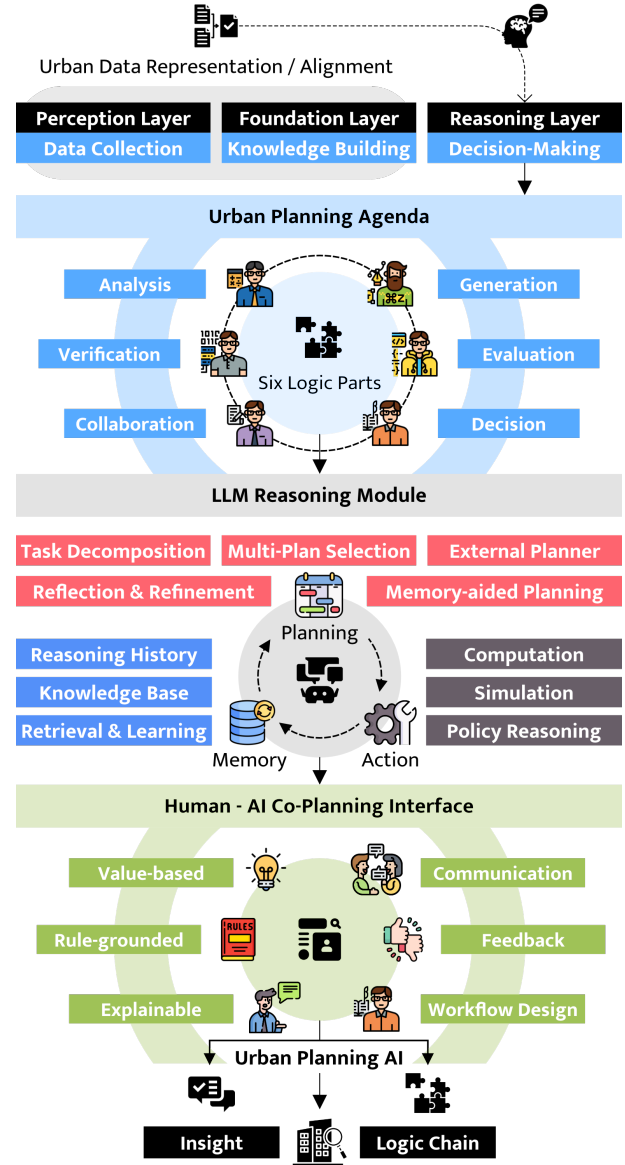


Figure 2: Agentic urban planning AI framework for reasoning-capable urban planning. The architecture comprises three cognitive layers (Perception, Foundation, Reasoning) and six logic components (Analysis, Generation, Verification, Evaluation, Collaboration, Decision) integrated through a human-AI co-planning interface supporting value-based, rule-grounded, and explainable decision-making.

⁶Large Language Model

Cognitive Layer	Sub-Module	Representative AI	Urban Planning Function	Planning Stage
Perception Layer	Visual Perception	SAM (Kirillov et al. 2023) ViT (Dosovitskiy 2020)	Urban imagery segmentation Urban imagery embedding	Data Collection
	Cross-Modal Fusion	CLIP (Radford et al. 2021) BLIP-2 (Li et al. 2023)	Vision-language alignment Multi-modal understanding	
	3D Reconstruction	NeRF (Mildenhall et al. 2020) 3DGS (Kerbl et al. 2023)	3D urban scene reconstruction Urban geometry modeling	
Foundation Layer	Statistical Learning	XGBoost (Chen and Guestrin 2016) SHAP (Lundberg and Lee 2017)	Tabular data prediction Model interpretation	Knowledge
	Large Language Models	Qwen 3 (Yang et al. 2025a) Llama 3 (Grattafiori et al. 2024)	Planning document analysis Policy semantic understanding	
	RAG	RAG (Lewis et al. 2020) LangChain (Chase 2022)	Knowledge retrieval Regulation query system	
	Simulation & RL	DQN (Mnih et al. 2015) PPO (Schulman et al. 2017)	Policy learning Multi-objective optimization	
Reasoning Layer	Cognitive Reasoning	CoT (Wei et al. 2023) ToT (Yao et al. 2023a)	Step-by-step reasoning Alternative exploration	All Stages
	Goal-Oriented Planning	LATS (Zhou et al. 2024a) Voyager (Wang et al. 2023)	Autonomous task planning Embodied learning	Analysis, Generation
	Tool-Augmented	ReAct (Yao et al. 2023b) Toolformer (Schick et al. 2023)	External tool invocation API-based computation	Generation, Verification
	Normative Reasoning	Const AI (Bai et al. 2022) RLHF (Christiano et al. 2017)	Value-aligned reasoning Preference learning	Evaluation, Decision
	Multi-Agent System	AutoGen (Wu et al. 2024b) MetaGPT (Hong et al. 2023)	Multi-agent collaboration Role-based coordination	Collaboration, Decision
	Test-Time Reasoning	GPT-o1 (OpenAI 2024) DeepSeek-R1 (Guo et al. 2025)	Advanced reasoning RL-based reasoning	Verification, Decision

Table 2: Agentic urban planning AI framework: A three-layer cognitive architecture for planning. The framework progresses through **Perception** (data collection), **Foundation** (knowledge building with statistical learning, LLMs, RAG, and simulation), and **Reasoning** (agentic AI for urban planning tasks). The reasoning layer comprises six functional stages: analysis, generation, verification, evaluation, collaboration, and decision. Representative AI models are mapped to their corresponding urban planning functions and stages.

Perception Layer (Data Collection): This foundation layer collects and processes multi-modal urban data (Yang, Lei, and Biljecki 2025). Computer vision models (SAM, ViT) extract spatial information from satellite imagery and street-level photos (Liang et al. 2025). Multi-modal models (CLIP, BLIP-2) link visual data with textual descriptions. 3D reconstruction techniques (NeRF, 3DGS) create detailed spatial representations. This layer transforms raw urban data into structured, machine-interpretable formats that inform downstream reasoning.

Foundation Layer (Knowledge Building): Statistical learning models build predictive knowledge from historical data. XGBoost and SHAP provide interpretable predictions and feature importance for traffic patterns, development impacts, and cost estimation. Large language models (Qwen, Llama 3) synthesize planning knowledge from regulatory documents, guidelines, and precedents (Hou et al. 2025; Zheng et al. 2025). Recent unified multimodal models (Xie et al. 2024; Zhou et al. 2024b) enable integrated understanding and generation across modalities. RAG and LangChain enable retrieval of relevant planning knowledge to support reasoning. This layer constructs the knowledge base that reasoning agents query during decision-making.

Reasoning Layer (Decision-Making): This layer performs explicit logical reasoning for planning decisions. Cognitive reasoning agents (DQN, PPO) learn strategic decision policies. Chain-of-Thought and Tree-of-Thought reasoning decompose complex planning problems into verifiable steps. Tool-using agents (ReAct, Toolformer) invoke constraint solvers and simulation tools. Multi-agent systems (AutoGen, MetaGPT) coordinate specialized reasoning agents (Wang et al. 2025a). Value-aligned reasoning (Constitutional AI, RLHF) ensures decisions reflect planning principles. Advanced reasoning models (GPT-o1, DeepSeek-R1) provide process supervision for reasoning quality.

Six Logic Components: Building on these three layers, the reasoning architecture orchestrates six specialized components that correspond to distinct planning deliberation stages (see Figure 2 and Table 1): *Analysis* conducts spatial and multi-criteria analysis; *Generation* produces planning alternatives through constrained search; *Verification* formally verifies regulatory compliance using symbolic solvers; *Evaluation* assesses proposals against normative criteria (sustainability, equity, resilience); *Collaboration* facilitates multi-stakeholder dialogue and consensus-building (Qian et al. 2023); and *Decision* synthesizes reasoning chains into actionable recommendations with explicit trade-offs. These components operate iteratively, enabling planners to critique reasoning, adjust priorities, and refine proposals through bidirectional human-AI interaction.

Formal Problem Definition. We formalize the urban planning decision problem as a constrained multi-objective optimization with explicit reasoning requirements. Given:

- A planning context $\mathcal{C} = \langle \mathcal{D}, \mathcal{K}, \mathcal{S} \rangle$ comprising spatial data $\mathcal{D}$, planning knowledge $\mathcal{K}$, and stakeholder input $\mathcal{S}$
- A set of hard constraints $\mathcal{H} = \{h_1, h_2, \dots, h_m\}$ representing regulatory requirements (zoning codes, environmental standards)

- A set of soft objectives $\mathcal{O} = \{o_1, o_2, \dots, o_n\}$ capturing normative criteria (equity, sustainability, liveability)

The reasoning-capable planning agent seeks to generate a proposal $p \in \mathcal{P}$ along with an explicit reasoning chain $r \in \mathcal{R}$ such that:

$$p^*, r^* = \arg \max_{p \in \mathcal{P}, r \in \mathcal{R}} \left[\sum_{i=1}^n w_i \cdot o_i(p) \right] \quad (1)$$

subject to $\forall h_j \in \mathcal{H} : h_j(p) = \text{True}$

where w_i are stakeholder-specified objective weights, and crucially, the reasoning chain r must satisfy:

$$\text{Valid}(r) \wedge \text{Complete}(r) \wedge \text{Traceable}(r, p, \mathcal{C}) \quad (2)$$

This formulation distinguishes reasoning agents from statistical learning: reasoning chains r provide explicit, verifiable justifications linking context $\mathcal{C}$ to proposal p , enabling human inspection and critique.

Multi-Agents Collaboration Framework

As illustrated in Figure 3, the Collaboration component of the Agentic Urban Planning AI Framework is implemented through a multi-agents collaboration framework that integrates the six logic parts (Analysis, Generation, Verification, Evaluation, Collaboration, Decision) operating across the three cognitive layers (Perception, Foundation, Reasoning). The framework supports two complementary collaboration methods tailored to different planning contexts:

Method 1: Linear Individual Review. In this workflow, individual planners sequentially review AI-generated planning proposals. Human planners independently rate proposals, provide comments, and suggest revisions. The AI system processes individual feedback to refine recommendations, enabling focused expert input without requiring group coordination. This method suits contexts where specialized expertise is needed (e.g., transportation engineers reviewing transit proposals) or when scheduling conflicts prevent simultaneous participation.

Method 2: Group Discussion. This workflow facilitates collective deliberation among multiple stakeholders. Planners engage in structured group discussions to evaluate AI recommendations collaboratively, surface conflicting priorities, negotiate trade-offs, and build consensus. The AI system captures group feedback, identifies areas of agreement and disagreement, and generates revised proposals that balance competing interests. This method is essential for contentious decisions requiring public input or multi-stakeholder negotiation (e.g., affordable housing siting, community facility allocation).

Both methods operate through a consistent *Generation-Verification-Evaluation* pipeline anchored by the *Human-AI Interface*. The multi-agents system generates planning alternatives, formally verifies regulatory compliance, and evaluates proposals against normative criteria. Human planners provide *rating* (quantitative assessment of proposals), *commenting* (qualitative feedback explaining concerns or suggestions), and request *revision* (iterative refinement incorporating human input). This bidirectional interaction ensures

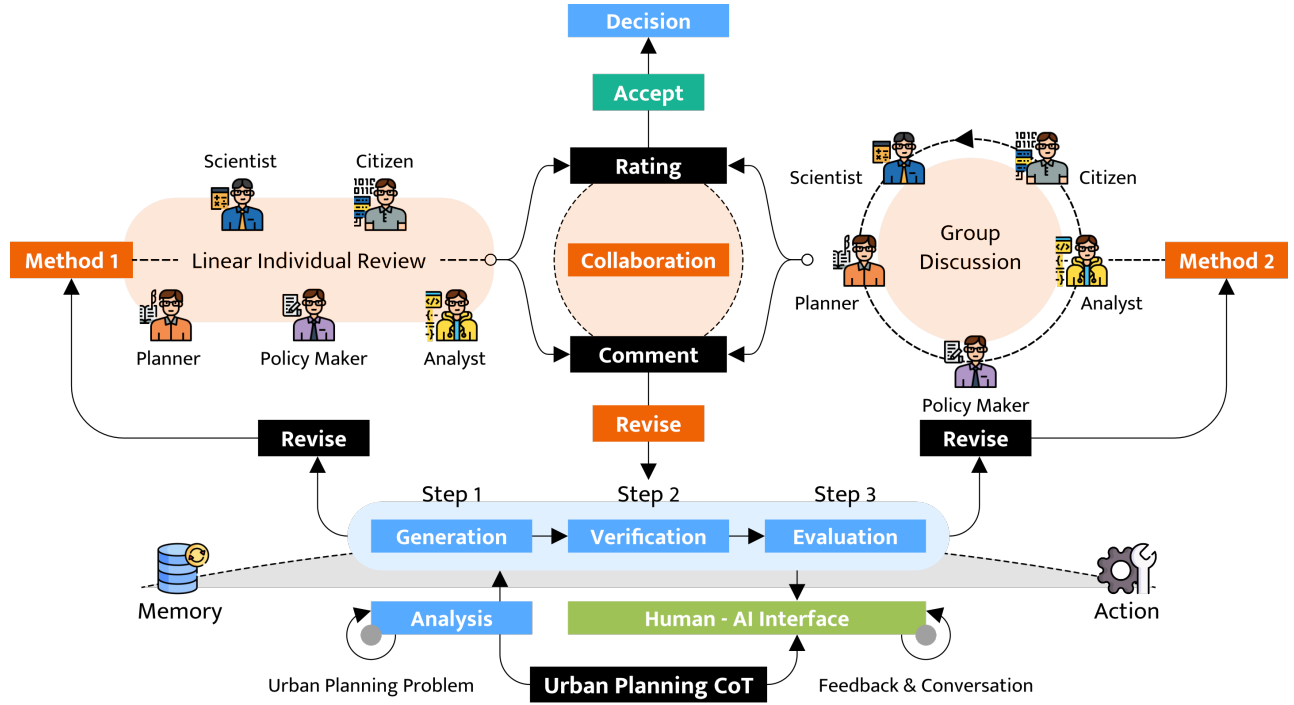


Figure 3: Multi-agents collaboration framework implementing the Collaboration component of the agentic urban planning AI framework. The framework supports two collaboration methods: linear individual review (Method 1) and group discussion (Method 2). Six logic parts (Analysis, Generation, Verification, Evaluation, Collaboration, Decision) operate through the human-AI interface, enabling iterative refinement via rating, commenting, and revision across three cognitive layers.

that AI agents augment rather than automate planning decisions.

Critical to this multi-agents collaboration framework is the *Analysis* component at the foundation, which synthesizes urban planning data from the Perception and Foundation layers to inform generation, and the *Collaboration* component, which structures stakeholder input into machine-interpretable constraints and coordinates agent interactions. The final *Decision* component integrates reasoning chains from multiple agents with human judgment, presenting *Accept/Revise* options that preserve human agency in decision-making.

Algorithm 1 formalizes the core reasoning-verification-collaboration pipeline that operationalizes the Agentic Urban Planning AI Framework.

Conclusion and Research Agenda

This position paper has presented the Agentic Urban Planning AI Framework for reasoning-capable urban planning agents that integrates three cognitive layers (Perception, Foundation, Reasoning) with six logic components (Analysis, Generation, Verification, Evaluation, Collaboration, Decision) through a multi-agents collaboration framework. We have shown why urban planning decisions require explicit reasoning capabilities—multi-constraint satisfaction, transparent justification, normative deliberation—and demonstrated how this architecture addresses these requirements through value-based, rule-grounded, and ex-

plainable decision-making.

Open Research Challenges. Realising this vision requires addressing five critical challenges:

(1) *Constraint Knowledge Formalisation:* How do we encode urban planning knowledge—zoning codes, environmental regulations, design guidelines—in machine-interpretable form while maintaining flexibility? Key questions: formal languages for spatial/temporal/normative constraints; automatic extraction from regulatory documents; handling ambiguous or conflicting regulations.

(2) *Reasoning Quality and Verification:* How do we ensure reasoning chains are correct and complete, especially when LLMs can generate plausible but invalid reasoning? Key questions: verifiers detecting constraint violations or logical errors; formal methods ensuring completeness; benchmarking reasoning quality when ground truth is normative.

(3) *Scalability and Efficiency:* Real-world planning involves thousands of constraints and iterative refinement. Key questions: efficient search algorithms pruning invalid paths early; balancing reasoning depth with inference speed; caching/modular strategies enabling real-time interaction.

(4) *Learning-Reasoning Integration:* What is the optimal division of labor between learning and reasoning components? Key questions: which tasks benefit most from learning (prediction) vs. reasoning (constraint satisfaction); how learned models provide probabilistic inputs to reasoning; detecting when learned estimates require verification.

Algorithm 1: Agentic Urban Planning CoT Pipeline with Human-AI Interface

```
1: Input: Context  $\mathcal{C} = \langle \mathcal{D}, \mathcal{K}, \mathcal{S} \rangle$ , constraints  $\mathcal{H}$ , objectives  $\mathcal{O}$ , method  $M$ 
2: Output: Proposal  $p^*$  with complete reasoning chain  $r^*$ 
3:
4: // Phase 1: Analysis - Understand context and formulate strategies
5:  $\mathcal{H}, \mathcal{O} \leftarrow \text{HumanAI.InitRequirements}()$  // Collect human requirements
6:  $features \leftarrow \text{Extract}(\mathcal{D})$  // Extract spatial, demographic, infrastructure features
7:  $knowledge \leftarrow \text{RAG.RetrieveCases}(\mathcal{C}, \mathcal{K})$  // Query similar historical cases
8:  $r_{ana} \leftarrow \text{Analyze}(features, knowledge)$  // CoT: diagnose issues, identify opportunities
9:  $strategies \leftarrow \text{ProposeStrategies}(r_{ana}, \mathcal{H}, \mathcal{O})$  // Generate improvement strategies
10:  $\text{HumanAI.Present}(r_{ana}, strategies)$  // Display analysis reasoning to user
11:
12: // Phase 2: Generation - Create diverse proposals from strategies
13:  $\mathcal{P} \leftarrow \emptyset, \mathcal{R} \leftarrow \emptyset$  // Initialize proposal and reasoning sets
14:  $regs \leftarrow \text{RAG.RetrieveRegs}(\mathcal{C})$  // Query relevant zoning codes and regulations
15: for all  $s_j \in strategies$  do
16:    $r_{gen} \leftarrow \text{ReasonStrategy}(s_j)$  // CoT: site selection, allocation logic
17:    $(p_j, r_{des}) \leftarrow \text{Design}(s_j, regs, \mathcal{H})$  // Design proposal following regulations
18:    $r_j \leftarrow r_{ana} \oplus r_{gen} \oplus r_{des}$  // Chain reasoning: analysis + generation + design
19:    $\mathcal{P} \leftarrow \mathcal{P} \cup \{p_j\}, \mathcal{R} \leftarrow \mathcal{R} \cup \{r_j\}$  // Add to candidate sets
20:    $\text{HumanAI.Display}(p_j, r_j)$  // Real-time visualization with reasoning
21: end for
22:
23: // Phase 3: Verification - Check constraint satisfaction symbolically
24: for all  $(p, r) \in \mathcal{P} \times \mathcal{R}$  do
25:    $(valid, viols) \leftarrow \text{Check}(p, \mathcal{H})$  // Symbolic constraint checking (e.g., CSP solver)
26:    $r_{ver} \leftarrow \text{Explain}(viols)$  // Generate reasoning explaining verification result
27:   if  $\neg valid$  then
28:      $\text{HumanAI.LogViol}(viols); \text{Remove}(p, r)$  // Log violations, filter invalid
29:   else
30:      $r \leftarrow r \oplus r_{ver}$  // Append verification reasoning to chain
31:   end if
32: end for
33:
34: // Phase 4: Evaluation - Assess impacts and score proposals
35: for all  $(p, r) \in \mathcal{P} \times \mathcal{R}$  do
36:    $imp \leftarrow \text{AssessImpacts}(p, \mathcal{O})$  // Assess equity, sustainability, economic impacts
37:    $reva \leftarrow \text{ReasonImpacts}(imp)$  // CoT: explain impact assessment and trade-offs
38:    $score \leftarrow \text{Score}(imp, \mathcal{O})$  // Compute value alignment score (VAS)
39:    $r \leftarrow r \oplus reva, \text{store}(p, r, score)$  // Append evaluation reasoning, store
40: end for
41:  $ranked \leftarrow \text{Rank}(\mathcal{P}, scores)$  // Rank proposals by score
42:  $\text{HumanAI.DisplayRanked}(ranked)$  // Show ranked alternatives to stakeholders
43:
44: // Phase 5: Collaboration - Multi-role review and feedback collection
45:  $\mathcal{A} \leftarrow \{\text{Planner, Scientist, Citizen, Analyst, PolicyMaker}\}$  // Define stakeholder roles
46:  $(p_{top}, r_{top}) \leftarrow \text{SelectTop}(\mathcal{P}, scores)$  // Select top-ranked proposal
47:  $\text{HumanAI.Present}(p_{top}, r_{top})$  // Present proposal with reasoning to stakeholders
48:  $fb \leftarrow \text{CollectFeedback}(\mathcal{A}, p_{top}, M)$  // Method 1: Linear review; Method 2: Group discussion
49: for all  $a \in \mathcal{A}$  do
50:    $r_{rev}^a \leftarrow a.\text{Review}(p_{top}, r_{top})$  // Role-specific CoT (e.g., environmental concerns)
51:    $r_{top} \leftarrow r_{top} \oplus r_{rev}^a$  // Append all review reasoning to chain
52: end for
53:
54: // Phase 6: Decision - Synthesize feedback and finalize recommendation
55:  $conf \leftarrow \text{FindConflicts}(fb)$  // Identify conflicting opinions among stakeholders
56:  $r_{con} \leftarrow \text{AnalyzeConf}(conf)$  // CoT: analyze nature and severity of conflicts
57: if  $conf \neq \emptyset$  then
58:    $r_{res} \leftarrow \text{Resolve}(conf, fb)$  // CoT: explain conflict resolution strategy
59:    $(p_{rev}, r_{ref}) \leftarrow \text{Refine}(p_{top}, r_{res})$  // Refine proposal to address conflicts
60:    $r^* \leftarrow r_{top} \oplus r_{con} \oplus r_{res} \oplus r_{ref}$  // Complete reasoning chain
61:    $\text{HumanAI.Present}(p_{rev}, r^*)$  // Show final decision with full justification
62:   return  $(p_{rev}, r^*)$  // Return revised proposal with reasoning
63: else
64:    $r_{dec} \leftarrow \text{Justify}(p_{top}, fb)$  // CoT: explain final decision rationale
65:    $r^* \leftarrow r_{top} \oplus r_{dec}$  // Complete reasoning chain with decision justification
66:    $\text{HumanAI.Present}(p_{top}, r^*)$  // Show final decision with full justification
67:   return  $(p_{top}, r^*)$  // Return accepted proposal with reasoning
68: end if
```

(5) *Fairness, Equity, and Value Alignment*: How do we ensure reasoning systems question historical biases and align with diverse stakeholder values? Key questions: explicit equity evaluation; transparent value elicitation and multi-stakeholder deliberation; auditing reasoning chains for hidden assumptions.

Beyond technical challenges, deployment raises policy questions about liability, transparency requirements, and democratic participation. Addressing these demands interdisciplinary collaboration among AI researchers, planners, legal scholars, and community stakeholders.

The Path Forward. The path forward requires collaboration between AI researchers, urban planners, and policymakers. Key priorities include: developing machine-readable planning knowledge bases (zoning codes, environmental standards); creating benchmarks for reasoning quality and constraint compliance; building agent architectures that integrate neuro-symbolic reasoning with test-time search; establishing verification and auditing protocols for AI planning assistants; and fostering interdisciplinary research on human-agent collaboration.

Urban planning decisions shape climate resilience, equity, opportunity, and health for generations. AI agents that reason transparently, verify constraints formally, and collaborate with human planners can help address unprecedented challenges—climate adaptation, housing crises, sustainable development—that demand both computational power and human wisdom. Reasoning is not merely beneficial—it is *foundational* for AI systems that augment, rather than undermine, trustworthy planning.

The technical capabilities now exist. The challenge is to build systems worthy of the decisions they will help shape. The opportunity is immense. The time is now.

Evaluation Benchmark Metrics

To assess reasoning-capable planning agents, we propose a comprehensive evaluation framework aligned with the three core requirements (value-based, rule-grounded, explainable) and six logic components of the Agentic Urban Planning AI Framework. Table 3 presents benchmark metrics organized by evaluation dimensions.

Formal Evaluation Metrics. We define key metrics corresponding to the six logic components in Algorithm 1:

Constraint Satisfaction Rate (Verification): Measures the proportion of hard constraints satisfied by generated proposals.

$$CSR(p, \mathcal{H}) = \frac{|\{h \in \mathcal{H} : h(p) = \text{True}\}|}{|\mathcal{H}|} \quad (3)$$

Reasoning Chain Quality (Analysis, Generation): Assesses the logical coherence and completeness of reasoning chains.

$$Q(r) = \alpha \cdot \text{Coherence}(r) + \beta \cdot \text{Completeness}(r) + \gamma \cdot \text{Traceability}(r) \quad (4)$$

where $\alpha + \beta + \gamma = 1$ are weighting parameters.

Value Alignment Score (Evaluation): Quantifies alignment between AI decisions and normative planning principles.

$$VAS(p, \mathcal{O}) = \sum_{i=1}^n w_i \cdot \frac{o_i(p) - o_i^{\min}}{o_i^{\max} - o_i^{\min}} \quad (5)$$

where $o_i^{\min}$ and $o_i^{\max}$ define normalisation bounds for objective o_i .

Human-AI Collaboration Efficiency (Collaboration): Measures interaction cycles required to reach acceptable proposals.

$$HACE = \frac{N_{\text{accepted}}}{N_{\text{iterations}} + \lambda \cdot T_{\text{total}}} \quad (6)$$

where N_{accepted} is the number of accepted proposals, $N_{\text{iterations}}$ is the total interaction cycles, T_{total} is the total time, and λ is a time penalty coefficient.

Decision Quality Score (Decision): Quantifies the overall quality of final decisions considering multiple criteria.

$$DQS(p^*, r^*) = \omega_1 \cdot CSR(p^*, \mathcal{H}) + \omega_2 \cdot Q(r^*) + \omega_3 \cdot VAS(p^*, \mathcal{O}) \quad (7)$$

where $\omega_1 + \omega_2 + \omega_3 = 1$ are component weights.

These metrics enable systematic evaluation of reasoning quality, constraint compliance, and human-AI collaboration effectiveness. Future research should develop standardized benchmark datasets for urban planning tasks with ground-truth constraint annotations, expert reasoning chains, and multi-stakeholder deliberation records.

Acknowledgments

We thank our colleagues at the NUS Urban Analytics Lab for the discussions. This research is supported by NUS Research Scholarship (NUSGS-CDE DO IS AY22&L GR-SUR0600042). This research is part of the project Large-scale 3D Geospatial Data for Urban Analytics, which is supported by the National University of Singapore under the Start Up Grant R-295-000-171-133. This research is part of the project Multi-scale Digital Twins for the Urban Environment: From Heartbeats to Cities, which is supported by the Singapore Ministry of Education Academic Research Fund Tier 1.

References

- Bai, Y.; Kadavath, S.; Kundu, S.; Askeel, A.; et al. 2022. Constitutional AI: Harmlessness from AI Feedback. arXiv:2212.08073.
- Chase, H. 2022. LangChain. <https://github.com/langchain-ai/langchain>.
- Chen, T.; and Guestrin, C. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, 785–794. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-4232-2.
- Christiano, P. F.; Leike, J.; Brown, T.; Martic, M.; Legg, S.; and Amodei, D. 2017. Deep Reinforcement Learning from Human Preferences. *Advances in neural information processing systems*, 30.
- Dosovitskiy, A. 2020. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv preprint arXiv:2010.11929*.
- Grattafiori, A.; Dubey, A.; Jauhri, A.; et al. 2024. The Llama 3 Herd of Models. arXiv:2407.21783.

Evaluation Dimension	Benchmark Metrics	Description	Pipeline Component
Analysis	Feature extraction accuracy	Precision/recall of spatial features and trade-off identification	Phase 1: Analysis
	Contextual understanding	Ability to interpret planning context $\mathcal{C}$ and identify key constraints	Phase 1: Analysis
Generation	Proposal diversity	Number of distinct valid alternatives generated per iteration	Phase 2: Generation
	Reasoning chain coherence	Human evaluation of CoT/ToT logical coherence (1-5 scale)	Phase 2: Generation
	Generation time	Time to produce $N_{samples}$ alternatives with reasoning (seconds)	Phase 2: Generation
Verification	Constraint satisfaction rate (CSR)	% of hard constraints $\mathcal{H}$ satisfied by proposals (Eq. 1)	Phase 3: Verification
	Constraint violation rate	% of proposals violating zoning, environmental, infrastructure constraints	Phase 3: Verification
	Verification latency	Time to verify proposal compliance (seconds per constraint)	Phase 3: Verification
Evaluation	Value alignment score (VAS)	Alignment with normative planning objectives $\mathcal{O}$ (Eq. 3)	Phase 4: Evaluation
	Equity impact assessment	Distributional fairness measures (Gini coefficient, accessibility gaps)	Phase 4: Evaluation
	Principle adherence	% of decisions aligning with sustainability, equity principles	Phase 4: Evaluation
Collaboration	Collaboration efficiency (HACE)	Interaction cycles to reach acceptable proposals (Eq. 4)	Phase 5: Collaboration
	Feedback incorporation rate	% of human critiques successfully integrated into revised proposals	Phase 5: Collaboration
	Stakeholder comprehension	Can non-experts understand AI rationales? (1-5 scale)	Phase 5: Collaboration
Decision	Decision quality score (DQS)	Overall quality combining CSR, Q, VAS (Eq. 5)	Phase 6: Decision
	Explanation completeness	% of decisions with complete justifications citing sources and regulations	Phase 6: Decision
	Decision agreement	Agreement rate between AI recommendations and human planner decisions	Phase 6: Decision

Table 3: Benchmark metrics for evaluating reasoning-capable planning agents organised by the six logic components in Algorithm 1. Each metric is mapped to its corresponding pipeline phase, ensuring alignment between formal evaluation (Equations 1-5) and operational implementation.

- Guo, D.; Yang, D.; Zhang, H.; et al. 2025. DeepSeek-R1 Incentivizes Reasoning in LLMs through Reinforcement Learning. *Nature*, 645(8081): 633–638.
- Hong, S.; Zhuge, M.; Chen, J.; Zheng, X.; Cheng, Y.; Wang, J.; Zhang, C.; Wang, Z.; Yau, S. K. S.; and Lin, Z. 2023. MetaGPT: Meta Programming for a Multi-Agent Collaborative Framework. In *The Twelfth International Conference on Learning Representations*.
- Hou, C.; Zhang, F.; Li, Y.; Li, H.; Mai, G.; Kang, Y.; Yao, L.; Yu, W.; Yao, Y.; and Gao, S. 2025. Urban Sensing in the Era of Large Language Models. *The Innovation*, 6(1).
- Jha, A. K.; Ghimire, A.; Thapa, S.; Jha, A. M.; and Raj, R. 2021. A Review of AI for Urban Planning: Towards Building Sustainable Smart Cities. In *2021 6th International Conference on Inventive Computation Technologies (ICICT)*, 937–944. IEEE.
- Kerbl, B.; Kopanas, G.; Leimkühler, T.; and Drettakis, G. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. arXiv:2308.04079.
- Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; and Lo, W.-Y. 2023. Segment Anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4015–4026.
- Lei, B.; Liu, P.; Liang, X.; Yan, Y.; and Biljecki, F. 2025. Developing the Urban Comfort Index: Advancing Liveability Analytics with a Multidimensional Approach and Explainable Artificial Intelligence. *Sustainable Cities and Society*, 120: 106121.
- Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; and Rocktäschel, T. 2020. Retrieval-Augmented Generation for Knowledge-Intensive Nlp Tasks. *Advances in neural information processing systems*, 33: 9459–9474.
- Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping Language-Image Pre-Training with Frozen Image Encoders and Large Language Models. In *International Conference on Machine Learning*, 19730–19742. PMLR.
- Li, J.; Ma, M.; and Lai, Y. 2025. Identifying Street Multi-Activity Potential (SMAP) and Local Networks with MLLMs and Multi-View Graph Clustering. *Computers, Environment and Urban Systems*, 122: 102350.
- Liang, X.; Xie, J.; Zhao, T.; Stouffs, R.; and Biljecki, F. 2025. OpenFACADES: An Open Framework for Architectural Caption and Attribute Data Enrichment via Street View Imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 230: 918–942.
- Liu, P.; and Biljecki, F. 2022. A Review of Spatially-Explicit GeoAI Applications in Urban Geography. *International Journal of Applied Earth Observation and Geoinformation*, 112: 102936.
- Liu, R.; Zhe, T.; Peng, Z.-R.; Catbas, N.; Ye, X.; Wang, D.; and Fu, Y. 2025. Towards Urban Planning AI Agent in the Age of Agentic AI. arXiv:2507.14730.
- Lundberg, S.; and Lee, S.-I. 2017. A Unified Approach to Interpreting Model Predictions. arXiv:1705.07874.
- Lv, Z.; Xu, J.; Zheng, K.; Yin, H.; Zhao, P.; and Zhou, X. 2018. Lc-Rnn: A Deep Learning Model for Traffic Speed Prediction. In *IJCAI*, volume 2018, 27.
- Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. arXiv:2003.08934.
- Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A. A.; Veness, J.; Bellemare, M. G.; Graves, A.; Riedmiller, M.; Fidjeland, A. K.; Ostrovski, G.; Petersen, S.; Beattie, C.; Sadik, A.; Antonoglou, I.; King, H.; Kumaran, D.; Wierstra, D.; Legg, S.; and Hassabis, D. 2015. Human-Level Control through Deep Reinforcement Learning. *Nature*, 518(7540): 529–533.
- OpenAI. 2024. Learning to reason with LLMs. Technical report, OpenAI. Available at <https://openai.com/index/learning-to-reason-with-llms/>.
- Peng, Z.-R.; Lu, K.-F.; Liu, Y.; and Zhai, W. 2024. The Pathway of Urban Planning AI: From Planning Support to Plan-Making. *Journal of Planning Education and Research*, 44(4): 2263–2279.
- Qian, K.; Mao, L.; Liang, X.; Ding, Y.; Gao, J.; Wei, X.; Guo, Z.; and Li, J. 2023. AI Agent as Urban Planner: Steering Stakeholder Dynamics in Urban Planning via Consensus-based Multi-Agent Reinforcement Learning. arXiv:2310.16772.
- Qiu, W.; Zhang, Z.; Liu, X.; Li, W.; Li, X.; Xu, X.; and Huang, X. 2022. Subjective or Objective Measures of Street Environment, Which Are More Effective in Explaining Housing Prices? *Landscape and Urban Planning*, 221: 104358.
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; and Clark, J. 2021. Learning Transferable Visual Models from Natural Language Supervision. In *International Conference on Machine Learning*, 8748–8763. PmLR.
- Sanchez, T. W.; Shumway, H.; Gordner, T.; and Lim, T. 2023. The Prospects of Artificial Intelligence in Urban Planning. *International Journal of Urban Sciences*, 27(2): 179–194.
- Schick, T.; Dwivedi-Yu, J.; Dessì, R.; Raileanu, R.; Lomeli, M.; Hambro, E.; Zettlemoyer, L.; Cancedda, N.; and Scialom, T. 2023. Toolformer: Language Models Can Teach Themselves to Use Tools. *Advances in Neural Information Processing Systems*, 36: 68539–68551.
- Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal Policy Optimization Algorithms. arXiv:1707.06347.
- UN. 2019. *World Urbanization Prospects: The 2018 Revision*. New York: United Nations. ISBN 978-92-1-148319-2.
- Wang, C.; Kang, Y.; Gong, Z.; Zhao, P.; Feng, Y.; Zhang, W.; and Li, G. 2025a. CartoAgent: A Multimodal Large Language Model-Powered Multi-Agent Cartographic Framework for Map Style Transfer and Evaluation. *International Journal of Geographical Information Science*, 39(9): 1904–1937.

- Wang, D.; Lu, C.-T.; Ye, X.; Yigitcanlar, T.; and Fu, Y. 2025b. Generative AI Meets Future Cities: Towards an Era of Autonomous Urban Intelligence. *arXiv:2304.03892*.
- Wang, G.; Xie, Y.; Jiang, Y.; Mandekar, A.; Xiao, C.; Zhu, Y.; Fan, L.; and Anandkumar, A. 2023. Voyager: An Open-Ended Embodied Agent with Large Language Models. *arXiv:2305.16291*.
- Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q.; and Zhou, D. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *arXiv:2201.11903*.
- Wu, Q.; Bansal, G.; Zhang, J.; Wu, Y.; Li, B.; Zhu, E.; Jiang, L.; Zhang, X.; Zhang, S.; and Liu, J. 2024a. Autogen: Enabling next-Gen LLM Applications via Multi-Agent Conversations. In *First Conference on Language Modeling*.
- Wu, Q.; Bansal, G.; Zhang, J.; Wu, Y.; Li, B.; Zhu, E.; Jiang, L.; Zhang, X.; Zhang, S.; and Liu, J. 2024b. Autogen: Enabling next-Gen LLM Applications via Multi-Agent Conversations. In *First Conference on Language Modeling*.
- Xie, J.; Mao, W.; Bai, Z.; Zhang, D. J.; Wang, W.; Lin, K. Q.; Gu, Y.; Chen, Z.; Yang, Z.; and Shou, M. Z. 2024. Show-o: One Single Transformer to Unify Multimodal Understanding and Generation. *arXiv:2408.12528*.
- Xu, X.; Qiu, W.; Li, W.; Huang, D.; Li, X.; and Yang, S. 2022. Comparing Satellite Image and GIS Data Classified Local Climate Zones to Assess Urban Heat Island: A Case Study of Guangzhou. *Frontiers in Environmental Science*, 10: 1029445.
- Yang, A.; Li, A.; Yang, B.; et al. 2025a. Qwen3 Technical Report. *arXiv:2505.09388*.
- Yang, S.; Chong, A.; Liu, P.; and Biljecki, F. 2025b. Thermal Comfort in Sight: Thermal Affordance and Its Visual Assessment for Sustainable Streetscape Design. *Building and Environment*, 271: 112569.
- Yang, S.; Krenz, K.; Qiu, W.; and Li, W. 2023. The Role of Subjective Perceptions and Objective Measurements of the Urban Environment in Explaining House Prices in Greater London: A Multi-Scale Urban Morphology Analysis. *ISPRS International Journal of Geo-Information*, 12(6): 249.
- Yang, S.; Lei, B.; and Biljecki, F. 2025. Urban Comfort Assessment in the Era of Digital Planning: A Multidimensional, Data-driven, and AI-assisted Framework. *arXiv:2508.16057*.
- Yao, S.; Yu, D.; Zhao, J.; Shafran, I.; Griffiths, T.; Cao, Y.; and Narasimhan, K. 2023a. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. *Advances in neural information processing systems*, 36: 11809–11822.
- Yao, S.; Zhao, J.; Yu, D.; Du, N.; Shafran, I.; Narasimhan, K.; and Cao, Y. 2023b. ReAct: Synergizing Reasoning and Acting in Language Models. *arXiv:2210.03629*.
- Yap, W.; Wu, A. N.; Miller, C.; and Biljecki, F. 2025. Revealing Building Operating Carbon Dynamics for Multiple Cities. *Nature Sustainability*, 1–12.
- Zhang, C.; Sargent, I.; Pan, X.; Li, H.; Gardiner, A.; Hare, J.; and Atkinson, P. M. 2018. An Object-Based Convolutional Neural Network (OCNN) for Urban Land Use Classification. *Remote sensing of environment*, 216: 57–70.
- Zheng, Y.; Xu, F.; Lin, Y.; Santi, P.; Ratti, C.; Wang, Q. R.; and Li, Y. 2025. Urban Planning in the Era of Large Language Models. *Nature Computational Science*, 5(9): 727–736.
- Zhou, A.; Yan, K.; Shlapentokh-Rothman, M.; Wang, H.; and Wang, Y.-X. 2024a. Language Agent Tree Search Unifies Reasoning Acting and Planning in Language Models. *arXiv:2310.04406*.
- Zhou, C.; Yu, L.; Babu, A.; Tirumala, K.; Yasunaga, M.; Shamis, L.; Kahn, J.; Ma, X.; Zettlemoyer, L.; and Levy, O. 2024b. Transfusion: Predict the Next Token and Diffuse Images with One Multi-Modal Model. *arXiv:2408.11039*.
- Zhu, H.; Chen, G.; and Zhang, W. 2025. PlanGPT: Enhancing Urban Planning with a Tailored Agent Framework. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, 764–783.