CityRiSE: Reasoning Urban Socio-Economic Status in Large Vision-Language Models via Reinforcement Learning

Tianhui Liu
Information Hub,
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China
Tsinghua University
Beijing, China
tianhui.liu@connect.hkust-gz.edu.cn

Hetian Pang
Department of Electronic Engineering, BNRist,
Tsinghua University
Beijing, China
pht23@mails.tsinghua.edu.cn

Xin Zhang
Department of Electronic Engineering, BNRist,
Tsinghua University
Beijing, China
zhangxin4087@163.com

Jie Feng†
Zhongguancun Academy
Beijing, China
Tsinghua University
Beijing, China
fengjie@bza.edu.cn

Pan Hui†
Information Hub,
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China
panhui@ust.hk

Yong Li†
Department of Electronic Engineering, BNRist,
Tsinghua University
Beijing, China
liyong07@tsinghua.edu.cn

Abstract

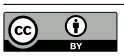
Urban socio-economic sensing plays a vital role in advancing global sustainable development goals. With the advent of Large Vision-Language Models (LVLMs), new opportunities have emerged to address this challenge by framing it as a multi-modal perception and reasoning task. However, recent studies show that LVLMs still struggle to make accurate and interpretable socio-economic predictions from visual data. To overcome these limitations and fully exploit the potential of LVLMs, we propose CityRiSE, a novel framework for Reasoning urban Socio-Economic status in LVLMs via reinforcement learning (RL). With carefully curated multi-modal dataset and verifiable reward design, our approach guides the LVLM to focus on semantically meaningful visual cues, enabling structured and goal-oriented reasoning for generalist socio-economic status prediction. Experiments demonstrate that CityRiSE, equipped with emergent reasoning, significantly outperforms existing baselines, improving both prediction accuracy and generalization across diverse urban contexts, especially on unseen cities and unseen indicators. This work highlights the promise of combining RL and LVLMs for interpretable and generalist urban socio-economic sensing.

CCS Concepts

• **Computing methodologies** → **Scene understanding**.

Keywords

Large Vision-Language Model, Socioeconomic Prediction, Urban Imagery, Reinforcement Learning



MM '26, Rio de Janeiro, Brazil

ACM ISBN 979-8-4007-2213-4/2026/11
https://doi.org/10.1145/3767308.3835986

ACM Reference Format:
Tianhui Liu, Hetian Pang, Xin Zhang, Jie Feng†, Pan Hui†, and Yong Li†. 2026. CityRiSE: Reasoning Urban Socio-Economic Status in Large Vision-Language Models via Reinforcement Learning. In *Proceedings of the 34th ACM International Conference on Multimedia (MM '26), November 10–14, 2026, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3767308.3835986

1 Introduction

Socio-economic indicators, such as GDP, population density, education level, and healthcare access, are critical for assessing the development, equity, and sustainability of urban regions. These metrics play a foundational role in guiding urban planning, policy-making, and international initiatives such as the United Nations Sustainable Development Goals (UN SDGs), which emphasize the importance of data-driven insights for fostering sustainable and inclusive growth. In recent years, there has been growing interest in predicting these indicators using externally observable data sources, in order to provide timely and spatially detailed insights that go beyond the limitations of traditional census-based methods. However, this task is inherently challenging. Socio-economic indicators are abstract, high-level targets that must be inferred from complex, heterogeneous, and often noisy data sources, with satellite imagery and street view images being among the most widely used modalities. Extracting informative features from such multi-modal inputs and effectively mapping them to socio-economic outcomes remains a non-trivial problem.

In response to these challenges, a number of recent studies have begun to explore learning-based approaches for socio-economic prediction from urban visual data. Early approaches, such as SVF [5], uses a two-stage pipeline where visual features are extracted from street view images using a computer vision model and then fed into regression model for socio-economic prediction. More recent efforts explore more integrated approaches based on knowledge graphs

†Corresponding authors.

Table 1: Comparison of Methods on City Transferability, Indicator Generalization, Data Scale, and Reasoning Ability.

Method	City Trans.	Indicator Gen.	Data Scale	Reasoning Ability
SVF	✓	✗	10^4	✗
KnowCL	✓	✗	10^3	✗
UrbanCLIP	✓	✗	10^4	✗
GeoLLM	✓	✗	10^4	✗
UrbanMLLM	✗	✗	10^5	✗
UI-CoT	✗	✗	10^5	✗
CityRiSE	✓	✓	10^3	✓

and contrastive learning frameworks to better model complex urban patterns [1, 21, 45]. With the emergence of large vision-language models [6, 7], these powerful models have also started to be applied to socio-economic indicator prediction tasks [12, 13, 25, 39, 42]. However, as summarized in Table 1, existing approaches face several limitations: First, while many existing models achieve strong performance on indicators or cities seen during training, their transferability to unseen regions or targets remains limited. Some methods explore cross-city prediction, but with poor performance, and few address the more challenging task of cross-indicator generalization. Second, existing approaches incur high data costs. Classic methods require large amounts of training data and often rely on external resources such as knowledge graphs. LVLM-based methods depend on supervised fine-tuning (SFT) with tens of thousands of labeled samples, limiting their scalability in low-resource settings. Third, most models lack interpretability. They output numerical predictions without providing reasoning steps or explanation, making it difficult to analyze or trust the model's decision-making process.

To address these limitations, we propose *CityRiSE*, a reinforcement learning framework built on top of the LVLM for generalist urban socio-economic status prediction. CityRiSE first enables generalization to unseen cities and indicators by normalizing diverse socio-economic targets into a unified prediction format and learning transferable reasoning patterns that connect visual input to high-level socio-economic concepts during training. In addition, we adopt Group Relative Policy Optimization (GRPO) as our training strategy and introduce two auxiliary datasets, the Perceptual Urban Reasoning Data and the General Visual Reasoning Data, that cultivate complementary reasoning skills and help the LVLM develop transferable cognitive abilities. This allows CityRiSE to achieve strong performance using only 5,109 training samples, without relying on large-scale supervised data. Finally, to support the emergence of interpretable reasoning, we design a verifiable reward mechanism consisting of a regression reward that enforces numerical correctness and a keyword reward that implicitly encourages the model to generate coherent, goal-directed reasoning chains. The main contributions of this paper are summarized as follows:

- We are the first, to our knowledge, to use reinforcement learning with LVLMs to enable emergent visual reasoning for high-level urban socio-economic status prediction.
- By learning to predict normalized indicator status, CityRiSE acquires transferable reasoning patterns that support strong generalization across cities and indicators, with cross-indicator generalization demonstrated for the first time in this domain.
- By adopting GRPO and constructing auxiliary datasets that target transferable perceptual and reasoning skills, CityRiSE achieves strong performance using only 5k training samples, surpassing even proprietary commercial LVLMs despite the significantly reduced data scale.
- Through verifiable reward design, CityRiSE induces emergent, coherent, and interpretable reasoning chains for indicator prediction, offering transparency and explanatory insights beyond black-box outputs.
- Extensive experiments demonstrate that CityRiSE outperforms existing SOTA baselines in 10 urban socio-economic status prediction tasks, particularly for prediction in unseen cities and unseen indicators. Our codes and datasets are open-sourced via https://github.com/tsinghua-fib-lab/CityRiSE.

2 Methods

2.1 Transferable Datasets for Cross-City and Cross-Indicator Generalization

To enable reasoning that generalizes across cities and indicators, we construct a dataset suite comprising a primary dataset for socio-economic prediction as well as two auxiliary datasets, which promote transferable cognitive patterns by targeting key perceptual and logical reasoning skills relevant to urban understanding.

2.1.1 Socio-Economic Indicator Data. We use the CityLens dataset, a multi-modal urban dataset spanning 17 cities worldwide and providing ground-truth annotations for socio-economic indicators across six key urban domains [20]. To normalize the heterogeneous value scales across indicators, we follow GeoLLM [23] and discretize each indicator into 10 bins, with values ranging from 1 to 10. Each input consists of one satellite image and ten street view images from a given region, and the model is required to predict a target socio-economic indicator for that region. To evaluate generalization, we consider two complementary settings: (1) spatial transferability, where 10 cities are used for training and evaluation while the remaining 7 are held out for testing, with both splits covering the Global South and Global North to ensure geographic diversity and mitigate potential regional bias; and (2) indicator generalization, where 5 indicators are used during training and the remaining 5 are used only for testing.

2.1.2 Perceptual Urban Reasoning Data. To enhance the model's ability to understand and reason about urban environments from street view images, we design three perceptual tasks that capture spatial, geographic, and socio-economic aspects of visual reasoning for training. The images and indicator annotations used to construct this dataset are drawn exclusively from the cities and indicators in the training split, without involving any city-indicator combinations from the test split.

Spatial Reasoning. In this task, each input sample contains three street view images, and the model must determine which image is spatially furthest from the other two. This encourages the model to infer spatial proximity using only visual cues. We introduce two settings: (1) two images are sampled from the same city, while the third comes from a different city; (2) two images are taken from the same neighborhood or block, while the third comes from a different neighborhood within the same city. These settings require

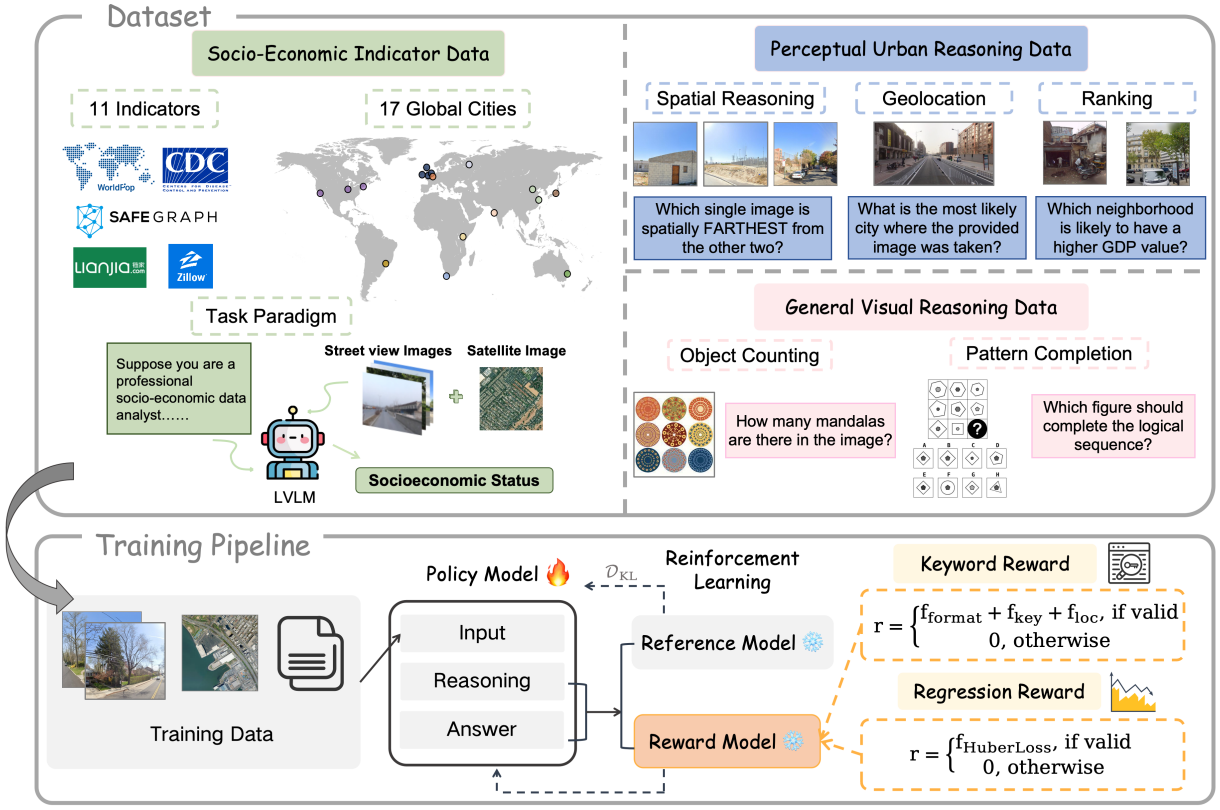


Figure 1: Framework of CityRiSE, which integrates three types of training data with a GRPO training pipeline. The GRPO algorithm incorporates both Keyword Reward and Regression Reward to guide learning.

the model to develop a fine-grained perception of urban layouts, building density, architectural styles, and environmental continuity.

Geolocation Task. This task draws on a well-established tradition in urban geospatial research, where inferring a city's identity from visual data has long served as a benchmark for evaluating spatial understanding [17, 18]. Given a single street view image, the model is required to predict its city of origin. This setting not only tests the model's ability to extract high-level geographic and architectural signals, like signage, vegetation, or road infrastructure, but also provides valuable priors for socio-economic status prediction. By grounding visual inputs in specific geographic contexts, the model can leverage its internally stored world knowledge, such as the average socio-economic profile of the predicted city, to make more informed and context-aware predictions.

Socio-Economic Ranking. Inspired by prior work in image quality assessment that formulates visual comparison as a learning-to-rank problem [34, 41], we design a perceptual reasoning task in which the model compares two street view images and determines which region has a higher value for a given socio-economic indicator, relying solely on visual cues. This pairwise comparison not only helps the model learn to associate visual patterns with latent urban conditions, but also mirrors the underlying structure of the final prediction objective, which involves distinguishing subtle differences in socio-economic status across regions.

2.1.3 General Visual Reasoning Data. To support the model's ability in urban socio-economic sensing and abstraction, we incorporate two types of general visual reasoning tasks to build fundamental perceptual skills essential for urban understanding. The first is an object counting task [26], where the model must determine the number of instances of a given object type within an image. This requires the model to localize, distinguish, and quantify multiple objects under diverse spatial arrangements. The second is a pattern completion task, inspired by classic geometric reasoning problems [16]. In this setting, the model is presented with a visual sequence or matrix with a missing element and is required to infer the underlying rule to select the correct option. These tasks encourage the model to recognize abstract visual patterns, including regularity, symmetry, analogy, and transformation, which are transferable to urban reasoning challenges, such as identifying socio-spatial structures or comparing urban forms across cities.

2.2 Guided Reward Design for Interpretable Socio-Economic Reasoning

We train the LVLM using the GRPO algorithm, and design a dual-component reward mechanism to guide the optimization process. Table 2 summarizes the reward types used for different training data. In particular, we introduce two novel reward functions tailored for the socio-economic indicator data. The Keyword Reward guides the LVLM toward more goal-directed perception and generation, thereby facilitating interpretable reasoning processes. The Regression Reward, in turn, promotes outputs that are numerically aligned with the ground-truth values and can be objectively evaluated.

Table 2: Training-task reward schemes and instance statistics.

Data Type	Format Reward	Accuracy Reward	Instances
Socio–Economic Indicator	Keyword	Regression	2,828
Spatial Reasoning	Standard	Standard	632
Geolocation Task	Standard	Standard	350
Socio-Economic Ranking	Standard	Standard	699
Object Counting	Standard	Regression	300
Pattern Completion	Standard	Standard	300

2.2.1 Keyword Reward. To guide the model toward producing responses that reflect meaningful urban visual features, we design a keyword-based reward component. This module is integrated into the format reward of GRPO, serving as an auxiliary signal that provides a lightweight perceptual prior during training. Rather than prescribing a fixed reasoning template, it encourages the model to attend to and cover task-relevant visual evidence that is potentially informative for socio-economic prediction. Concretely, we define a list of six urban-perceptual keywords: "person", "vehicle", "greenery", "road infrastructure", "street furniture", and "building". These keywords are derived from prior work on urban visual intelligence [5], which extracts 13 visual features using deep learning models for socio-economic prediction. We cluster these 13 features into six categories from the perceptual perspective of LVLMs. These keywords encompass nearly all visual cues that are visible in street view images and relevant to socio-economic prediction. Guided by these keywords, the model is encouraged to focus on informative visual content while retaining the flexibility to discover and integrate additional cues beyond the predefined keyword set. In addition, we assign a special reward to the mention of "location", which indicates that the model attempts to anchor the visual input to a specific city. This grounding process enables the model to retrieve its internally stored prior knowledge about the city, such as its average socio-economic levels, and integrate it into its reasoning. This step is crucial for models that aim to generalize across diverse urban environments.

$$R_{\text{keyword}}(r) = \lambda_{\text{base}} \cdot \mathbf{1}_{\text{format}(r)} + \sum_{k \in \mathcal{K}} \lambda_k \cdot \mathbf{1}_{k \in r} + \lambda_{\text{loc}} \cdot \mathbf{1}_{\text{loc} \in r}. \quad (1)$$

$$\mathbf{1}_{\text{condition}} = \begin{cases} 1, & \text{if the condition holds,} \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

2.2.2 Regression Reward. To better reflect the nature of our task as a regression problem, we design a reward function sensitive to the magnitude of prediction errors. In socio-economic indicator prediction, the model outputs a continuous score, and small deviations from the ground truth should not be penalized as harshly as large ones. For example, if the ground truth is 8, predicting 7 should be rewarded more than predicting 1, a subtlety that binary correctness-based metrics fail to capture. To address this, we introduce a regression reward based on the Huber loss:

$$\text{error} = y_{\text{pred}} - y_{\text{true}}, \quad |\text{error}| = \left| y_{\text{pred}} - y_{\text{true}} \right|, \quad (3)$$

$$L_{\text{Huber}}(\text{error}) = \begin{cases} \frac{1}{2}\,\text{error}^2, & \text{if } |\text{error}| \le \delta \\ \delta\left(|\text{error}| - \frac{1}{2}\delta\right), & \text{otherwise} \end{cases}, \quad (4)$$

$$R_{\text{regression}} = \exp\left(-\alpha\, L_{\text{Huber}}(\text{error})\right). \quad (5)$$

This formulation yields a reward that decays smoothly and nonlinearly with prediction error. The use of Huber loss ensures stability near the correct value due to its quadratic form, while reducing sensitivity to outliers through its linear behavior for larger errors. The exponential transformation ensures that predictions closer to the true value receive rewards close to 1, while distant predictions are penalized more strongly. Overall, this reward better aligns with the continuous, fine-grained nature of the prediction task, and provides a more informative learning signal during training. The same reward formulation is used for Object Counting Data, which also involves count-valued outputs.

2.3 GRPO Pipeline for Data-Efficient Learning

We train the LVLM using GRPO on the transferable dataset described in Section 2.1.1. As Algorithm 1 shows, GRPO operates by generating multiple candidate responses for each data point and calculating the reward for each candidate based on the reward design outlined in Section 2.2. To encourage the model to favor higher-reward completions within each group, GRPO computes a group-normalized advantage for each response. This advantage reflects how well each response performs relative to other candidates in the same group. Finally, the model is updated by minimizing a loss function that incorporates both the advantage and the KL divergence, ensuring the model learns to improve its performance on the dataset.

Algorithm 1 Training Algorithm for CityRiSE using GRPO

Input: Pretrained policy π_θ, reference policy π_{ref}, reward model R, training dataset $\mathcal{D}$, KL coeff β, clip coeff ϵ.
Output: Optimized policy π_θ.
for each $(x_i, y_i) \in \mathcal{D}$ **do**
  Generate N responses: $\{o_i^{(j)}\}_{j=1}^{N} \sim \pi_\theta(\cdot \mid x_i)$
  Compute rewards: $r_i^{(j)} = R(o_i^{(j)})$
  Compute advantages: $A_i^{(j)} = \frac{r_i^{(j)} - \text{mean}(\{r_i^{(1)}, \ldots, r_i^{(N)}\})}{\text{std}(\{r_i^{(1)}, \ldots, r_i^{(N)}\})}$
  Compute ratios: $s_i^{(j)} = \frac{\pi_\theta(o_i^{(j)} \mid x_i)}{\pi_{\theta_{\text{old}}}(o_i^{(j)} \mid x_i)}$
  Compute the following objective:
  $\mathcal{J}_i = \mathbb{E}[\{o_i^{(j)}\}_{j=1}^{N} \sim \pi_{\theta_{\text{old}}}(\cdot \mid x_i)] \frac{1}{N} \sum_{j=1}^{N} [\min\left(s_i^{(j)} A_i^{(j)},\right.$
  $\left.\text{clip}(s_i^{(j)}, 1-\epsilon, 1+\epsilon) A_i^{(j)}\right) - \beta \mathcal{D}_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}})]$
  Update π_θ by maximizing $\mathcal{J}_i$
end for
return π_θ

3 Experiments

To evaluate the effectiveness, generalization ability, and interpretability of CityRiSE, we organize our experiments around the following three research questions:

- **RQ1:** How does CityRiSE perform compared to proprietary LVLMs and SFT-based LVLMs, across in-domain settings, cross-city transferability, and cross-indicator generalization?
- **RQ2:** How does the design of the reward mechanism and the composition of training data influence the performance of CityRiSE?

Table 3: Main results on Socio-Economic Indicator Data-Test. The evaluation metric is the coefficient of determination (R^2). In each row, bold indicates the best result, and underline denotes the second-best.

Domain	Overall	Unseen Indicators					Unseen Cities					In-domain				
Tasks		DR	BH	VC	LE	HP	GDP	Pop.	MH	PT	BR	GDP	Pop.	MH	PT	BR
Random	-0.929	-0.916	-0.734	-1.106	-0.941	-0.951	-2.742	-0.874	-2.902	-4.694	-2.085	-1.293	-1.220	-0.798	-0.562	-1.191
SVF-GT	/	/	/	/	/	/	-1.046	0.091	-1.964	-1.358	-0.517	0.329	0.268	0.087	0.417	0.210
SVF-Rank	/	/	/	/	/	/	-0.560	0.172	-2.154	-1.581	-0.720	0.366	0.267	0.162	0.358	0.287
UrbanCLIP	/	/	/	/	/	/	-0.595	0.161	-1.537	-3.137	-0.622	0.176	0.015	-0.015	-0.025	0.008
UrbanVLP	/	/	/	/	/	/	-0.279	0.146	-0.174	-0.452	-0.265	0.609	0.500	0.260	0.568	0.510
Qwen2.5-VL-7B	-0.116	-0.130	0.178	-0.251	0.083	-0.026	-0.143	-0.209	-2.172	-0.118	-0.918	-0.280	-0.330	-0.545	-0.013	-0.094
Qwen3-VL-235B	-0.572	-1.534	0.290	-1.150	-0.874	0.056	0.114	-0.405	-5.924	0.004	-0.169	-0.752	-0.944	-1.288	-0.621	0.173
GPT-4.1-Mini	-0.233	0.252	0.285	-0.908	-0.023	0.111	**0.444**	-0.183	-4.943	**0.413**	-1.386	-0.323	-0.222	-0.989	0.342	0.239
Gemini2.5-Flash	-0.718	-0.903	0.395	-1.435	-0.525	-0.211	-0.260	-0.529	-6.709	0.098	-1.801	-0.825	-0.918	-1.078	0.529	-0.220
UrbanMLLM (SFT)	-0.007	-2.249	0.314	-0.087	-1.400	-0.209	0.182	0.274	**0.329**	0.067	-0.946	**0.691**	**0.574**	-0.005	**0.607**	0.534
UI-CoT (SFT-CoT)	0.176	-0.939	**0.398**	0.103	-0.493	-0.105	-0.419	-0.055	0.264	-0.073	0.063	0.546	0.499	0.053	0.446	**0.616**
CityRiSE (only RL)	**0.421**	**0.289**	0.366	**0.305**	**0.200**	**0.218**	0.391	**0.365**	0.259	-0.635	**0.286**	0.385	0.334	**0.399**	0.594	0.603

- **RQ3:** How well are CityRiSE's specific design choices justified, and what complementary roles do street-view and satellite imagery play at inference time?
- **RQ4:** What role does RL-induced reasoning play in shaping CityRiSE's predictive capabilities?

3.1 Baselines

To comprehensively evaluate our approach, we compare it against 8 baselines in three categories: (1) classic methods that do not involve LVLMs, including SVF variants [5], UrbanCLIP [39] and UrbanVLP [12]; (2) general LVLMs, and (3) LVLMs supervised fine-tuned on domain-specific data, consisting of UrbanMLLM [42], trained on the Socio-Economic Indicator Data-Train dataset, and UI-CoT [43], trained on a Chain-of-Thought (CoT) augmented version of the same dataset.

3.2 Overall Performance (RQ1)

Table 3 presents the main results on the Socio-Economic Indicator Data-Test set. The evaluation tasks are broadly categorized into three types based on their relationship to the training distribution:

- Unseen Indicators. These tasks involve indicators that do not appear in the training set. Except for the Life Expectancy task, which involves both unseen cities and unseen indicators, all other tasks in this group share cities with the training set but introduce new indicators.
- Unseen Cities. The indicators remain the same as in training, but the evaluation is conducted on cities that were not present in the training set.
- In-domain. Both the cities and indicators in these tasks are consistent with those seen during training. Importantly, the test set is constructed with entirely disjoint samples to prevent any data leakage and ensure a fair evaluation.

Generalization to Unseen Indicators. In the unseen indicators setting, CityRiSE achieves the best overall performance, outperforming all baselines across most of the novel socio-economic indicators. This highlights its strong ability to generalize not only to new visual contexts, but also to entirely unseen prediction targets. The model maintains positive R^2 scores on all unseen indicators, including challenging ones such as Life Expectancy (0.200) and House Price (0.218), where other models struggle or fail to generalize. Notably, traditional deep learning baselines such as SVF, UrbanCLIP and UrbanVLP are absent from this evaluation. This is because such models require a separate predictor for each target variable, making them inherently limited in handling unseen indicators. This limitation underscores a key advantage of LVLMs: their ability to serve as universal predictors for diverse tasks through language-based prompts.

Transferability to Unseen Cities. For unseen cities setting, CityRiSE demonstrates strong spatial transferability. Notably, it achieves the highest performance on Bachelor Ratio with an R^2 of 0.286, outperforming SFT-based models, general LVLMs and classic methods. Although UrbanVLP performs competitively in the in-domain setting, its performance drops substantially on unseen cities, indicating limited cross-city transferability. While CityRiSE shows slightly weaker performance on Public Transport, we observe that many test images lack clear transportation-related visual cues, and PT values in these cities are relatively low compared to the training set. These factors may partially limit generalization, though the overall trend remains consistent across tasks. GPT-4.1-Mini achieves unexpectedly strong results on PT, which we attribute to its conservative predictions that happen to align with the test distribution.

In-domain Analysis. Under the in-domain setting, CityRiSE significantly outperforms its base model Qwen2.5-VL-7B across all five indicators, with R^2 scores improving from -0.330 to 0.334 on Population and from -0.013 to 0.594 on Public Transport. These results demonstrate the strong predictive capability of CityRiSE. Qwen3-VL-235B A22B Thinking, GPT-4.1-Mini, and Gemini2.5-Flash are all strong general large vision-language models, while their zero-shot performance on socio-economic status prediction remains suboptimal, further highlighting the challenge of this task. Interestingly, SFT-based models, including UrbanMLLM and UI-CoT, achieve the highest performance on some indicators, particularly GDP and Population. This aligns with the insight proposed in Chu et al. [3] that "SFT Memorizes, RL Generalizes". Supervised fine-tuning is highly effective for in-domain memorization, whereas reinforcement learning tends to improve generalization beyond training

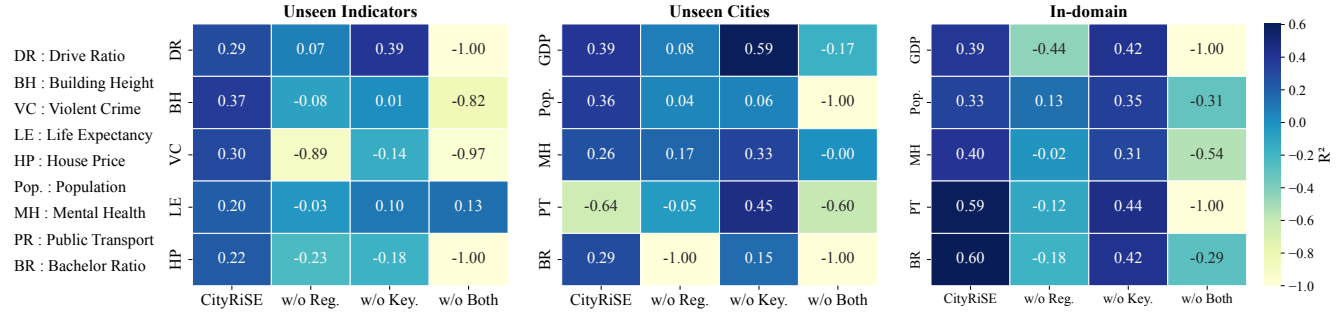


Figure 2: Results of Reward Ablation, where 'Reg.' refers to the Regression Reward, 'Key.' refers to the Keyword Reward, and 'w/o Both' indicates the setting where both rewards are removed.

data. The comparable performance of SVF-GT and SVF-Rank across in-domain indicators provides empirical support for the validity of our normalization strategy. While not implying full equivalence, the results suggest that normalization preserves predictive accuracy. CityRiSE further surpasses both baselines, demonstrating the benefits of combining reinforcement learning with LVLMs.

3.3 Ablation Study (RQ2)

Reward Ablation. Figure 2 shows the results of a reward ablation study comparing the full CityRiSE model with three variants: without Regression Reward, without Keyword Reward, and without both. Removed rewards are replaced by the standard reward. For clarity, R^2 values below -1 are clipped in the figure. Removing either reward consistently degrades performance, while removing both leads to severe failure across most indicators. This effect is evident in the in-domain setting, where the R^2 score on GDP drops from 0.39 to below -1.00, and Public Transport drops from 0.59 to below -1.00, indicating insufficient learning signals for effective prediction. The Keyword Reward shows particular importance for specific indicators, such as the drop on Violent Crime from 0.30 to -0.14, suggesting its role in enhancing semantic grounding. The Regression Reward, on the other hand, contributes more significantly to numerical precision, as reflected in the drop on Population in unseen cities from 0.36 to 0.04. Overall, the Regression Reward has a greater impact than the Keyword Reward, likely because rewarded keywords are already present in the prompts, reducing the marginal benefit of additional keyword supervision. Together, the two rewards offer complementary signals: the Regression Reward supports quantitative alignment, and the Keyword Reward strengthens interpretability and semantic relevance.

Data Ablation. We incorporate the Perceptual Urban Reasoning Data and General Visual Reasoning Data to enhance the model's reasoning capabilities relevant to socio-economic prediction, particularly its generalization to out-of-domain settings. Figure 3 shows the performance of models trained with different data ablations on the unseen indicators and unseen cities subsets. For clarity, tasks where all configurations yield negative R^2 scores are excluded, as they offer limited analytical value. As shown in Figure 3, removing either auxiliary dataset consistently degrades performance across most indicators, confirming the effectiveness of both Perceptual and General Visual Reasoning data. The General Visual Reasoning Data has broad influence, with its removal causing noticeable drops on nearly all indicators, especially Mental Health, where the R^2 score falls from 0.26 to -0.71. In contrast, the Perceptual Urban Reasoning Data plays a more targeted role; its removal notably impacts Life Expectancy and House Price, suggesting that spatial comparison and urban identity reasoning support accurate estimation for region-grounded indicators. Removing both datasets leads to severe degradation across nearly all tasks, including complete collapse on Mental Health, highlighting their complementary roles in enabling robust visual reasoning and generalization.

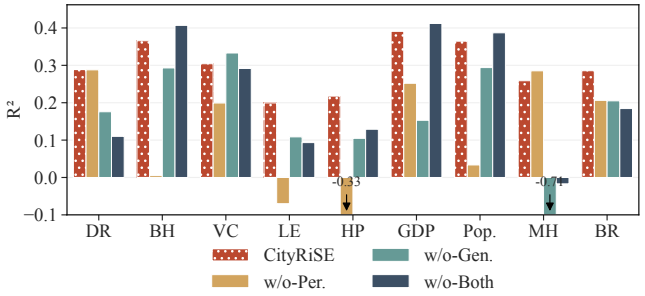


Figure 3: Results of Data Ablation, where 'Per.' refers to the Perceptual Urban Reasoning Data, 'Gen.' refers to the General Visual Reasoning Data.

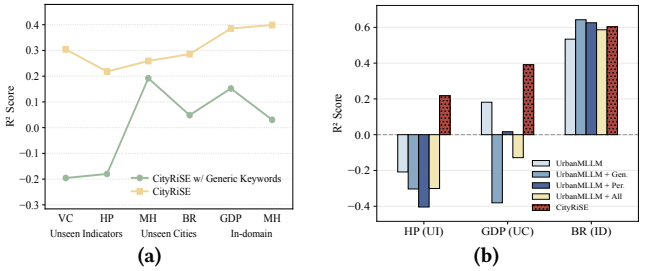


Figure 4: (a) Comparison between CityRiSE and CityRiSE w/ Generic Keywords. (b) Performance of SFT-LVLMs trained with different data compositions.

3.4 Training Design Validation and Input Modality Analysis (RQ3)

Task-Aligned vs. Generic Keyword Reward. To further validate the design of the keyword reward, we compare our task-aligned visual keyword set with an alternative keyword design based on generic reasoning-oriented tokens. Specifically, we replace the six visual

Table 4: Performance of CityRiSE under different test-time input-modality settings, compared with its base model.

Task	UI		UC		ID	
	DR	HP	Pop.	MH	PT	BR
CityRiSE	**0.289**	**0.218**	**0.365**	**0.259**	**0.594**	**0.603**
w/o SAT	0.270	0.154	0.328	0.152	0.579	0.583
w/o STV	0.011	−0.277	0.147	−0.658	−0.202	0.157

keywords in CityRiSE with generic thinking tokens, motivated by prior work [27] suggesting that expressions such as "Hmm", "Wait", and "Therefore" are crucial for the reasoning performance of large reasoning models. We denote the resulting variant as CityRiSE w/ Generic Keywords. As shown in Figure 4a, CityRiSE consistently outperforms this variant on all 6 tasks across three settings. These results indicate that the effectiveness of the keyword reward does not simply come from encouraging deliberative language, but from rewarding task-aligned visual concepts that are directly relevant to urban socio-economic reasoning. This further supports the validity of our reward design.

SFT with Different Data Compositions. We further study how different training data compositions affect SFT performance. Specifically, we conduct supervised fine-tuning on Qwen2.5-VL-7B using the socio-economic indicator data alone and variants further augmented with the auxiliary datasets. Figure 4b presents the results on 3 tasks spanning the in-domain, unseen cities, and unseen indicators settings. As auxiliary datasets are added, UrbanMLLM shows modest gains in the in-domain setting, whereas performance on most out-of-domain tasks declines, with only a few individual tasks improving. This pattern further confirms the strong in-domain nature of SFT, and suggests that auxiliary data alone is insufficient to yield the transferable gains achieved by CityRiSE.

Test-Time Modality Contribution. To examine how CityRiSE utilizes different visual sources, we remove either satellite (SAT) or street-view (STV) imagery at inference time without retraining the model. As shown in Table 4, removing SAT leads to consistent but moderate degradation across all tasks, with the largest drops observed on MH (0.259 → 0.152) and HP (0.218 → 0.154). In contrast, removing STV causes substantially larger declines, producing negative R^2 scores on HP, MH, and PT. These results suggest that, under our input configuration, street-view imagery provides the primary fine-grained visual evidence for socio-economic estimation, while satellite imagery supplies complementary regional-scale context. The consistently superior results of the complete model further demonstrate the benefit of combining both modalities.

3.5 Reasoning Analysis and Comparison (RQ4)

Reasoning Impact. To isolate the contribution of reasoning, we construct a no-reasoning variant that keeps the GRPO training procedure, regression reward, and auxiliary data unchanged, but directly generates the final prediction without intermediate reasoning. As shown in Figure 6a, this variant outperforms the base model but remains consistently below CityRiSE across all three evaluation settings. These results distinguish the benefit of RL training from that of reasoning and demonstrate that the reasoning process itself contributes to more accurate socio-economic prediction.

Reasoning Integration. Figure 6b compares models trained with three different strategies on 3 representative tasks spanning three evaluation settings. The three models differ in how reasoning is introduced during training. UI-CoT is fine-tuned on manually constructed CoT annotations; UI-CoT + RL adds reinforcement learning on top of it. In contrast, CityRiSE is trained purely via RL, without access to human-crafted reasoning chains, relying instead on reward design to induce emergent reasoning behavior. UI-CoT tends to generate rigid, template-like reasoning that works in-domain but generalizes poorly. RL fine-tuning improves its in-domain results but brings limited out-of-domain gains, likely due to the fixed reasoning style inherited from supervised training, which remains largely unchanged. In contrast, CityRiSE develops reasoning organically during RL training, guided by verifiable rewards. This results in more flexible and transferable inference patterns, with strong performance on both unseen cities and indicators. These findings support the hypothesis that emergent, reward-induced reasoning is more adaptable to novel contexts than explicitly scripted approaches. We further apply RL to UrbanMLLM, a SFT-only model. Despite prompting it to "Please reason step-by-step and write your reasoning inside the <think> </think> tags" during RL, the model consistently outputs only the final answer. This suggests that if reasoning is not introduced during the early SFT stage, the model may become too rigid for RL to induce reasoning capability later.

Reasoning Case. Figure 5 presents a qualitative comparison of partial reasoning chains generated by three models for the same socio-economic prediction task. Although CityRiSE is trained via reinforcement learning without access to answer labels, it develops a clear and coherent reasoning structure. This emerges from the combined effect of the Keyword Reward and prompt design, which offer implicit and explicit guidance toward goal-directed reasoning. Compared to its base model (Qwen2.5-VL-7B), CityRiSE demonstrates a more holistic and higher-level perception across multiple images, rather than producing low-information summaries of each image in isolation. Moreover, CityRiSE is able to more accurately perceive and reason about the street view images. It successfully infers the likely location as "East Africa", aligning with the output of GPT-4.1-Mini, a much larger proprietary model. This geographic inference provides valuable prior knowledge for socio-economic status prediction, as regional context can help calibrate the model's expectations for certain indicators. These results demonstrate not only robust visual grounding, but also a natural and uninterrupted reasoning process, emerging organically through RL training.

4 Related Work

Socio-Economic Prediction from Urban Imagery. In general, Urban socio-economic sensing [15] aims to predict high-level indicators based on visual data from urban environments, particularly satellite and street view images. The core challenge is effectively extracting useful multi-source information from these multi-modal data for accurate assessment. To address this challenge, researchers have developed various modeling approaches that leverage the diverse, multi-modal urban data, ranging from classic regression methods [5] to modern deep learning [21, 45] and representation learning [1, 40] techniques. With the advent of LLMs, a promising new direction involves leveraging their embedded commonsense and world knowledge to enrich urban visual data with contextual

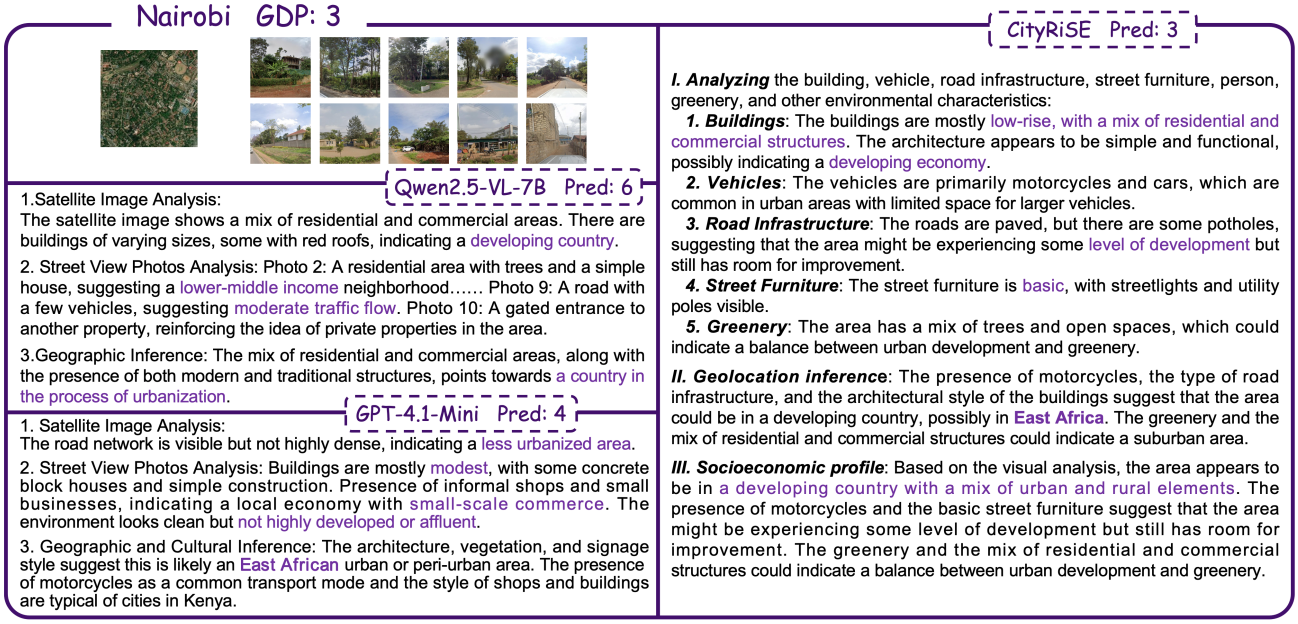


Figure 5: Qualitative comparison of reasoning patterns produced by different models. CityRiSE exhibits more interpretable and goal-directed reasoning behavior, while baseline models often generate less informative responses.

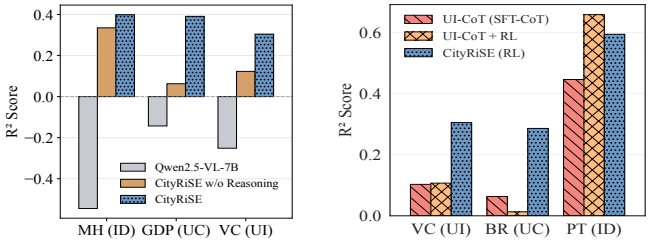


Figure 6: (a) Comparison of CityRiSE, its no-reasoning RL variant, and the base model across representative tasks. (b) Performance of models trained with different methods on three types of tasks.

information [8, 19, 20]. The typical paradigms for this integration fall into two categories: using LLMs as a supplementary tool for extra information to assist the aforementioned classical methods [12, 39], and directly employing LLMs as the predictor by leveraging their rich internal knowledge and modeling capabilities. The second approach, which more tightly integrates the LLM's inherent knowledge with indicator prediction, has garnered more attention [11, 22, 44]. GeoLLM [23] links the LLM's internal knowledge to prediction indicators via address information, and UrbanMLLM [42] completes multi-type and multi-task indicator prediction directly through large-scale data pre-training and supervised learning.

Generalist Reasoning in LVLMs. Prior work primarily guides LVLMs through fixed prompt design for visual reasoning [2, 10, 33, 38]. More recently, autonomous visual reasoning in LVLMs based purely on Reinforcement Learning has become a popular topic. The core goal is to train LVLMs via RL, to independently learn to analyze and reason, thereby boosting performance and crucially generalization ability in typical visual tasks [24, 30–32]. Current work in this space has mainly concentrated on the understanding and reasoning of low-level, intuitive visual elements, such as mathematical geometry problems and natural image comprehension. By utilizing datasets annotated with explicit reasoning processes [4, 14, 29, 37] and carefully designed reward functions [28, 35, 36], similar to the approach of DeepSeek-R1 [9], researchers guide LVLMs toward more general visual reasoning, leading to higher performance and better generalization. Distinct from these efforts focusing on intuitive, low-level visual elements, our work is the first to apply the RL paradigm to implicit, high-level visual reasoning. We achieve the association and mining of intuitive elements with hidden knowledge, thus enabling the generalized prediction of unseen socio-economic indicators for the first time while simultaneously enhancing the generalized prediction capability for unseen cities.

5 Conclusion

In this paper, we propose CityRiSE, the first framework to achieve generalist urban socio-economic status prediction across diverse cities and indicators by leveraging the emergent reasoning capabilities of LVLMs through reinforcement learning. CityRiSE first achieve cross-city and cross-indicator generalization by training the model on normalized prediction targets, a process that encourages learning transferable reasoning patterns connecting visual input to high-level socio-economic concepts. By combining GRPO with two auxiliary datasets that enhance transferable perceptual and logical reasoning skills, CityRiSE achieves strong performance using only 5k training samples, demonstrating high data efficiency. Finally, we design a verifiable multi-component reward mechanism that induces coherent and goal-directed reasoning, thereby enhancing both interpretability and prediction quality. Extensive experimental results demonstrate the superiority of CityRiSE over SOTA baselines for 10 urban socio-economic prediction tasks.

Acknowledgments

This work was supported in part by the National Key Research and Development Program of China under Grant 2024YFC3307602, the Guangdong Provincial Talent Program under Grant No. 2023JC10X009, and the National Key Research and Development Program of China under Grant 2024YFC3307603. It was also sponsored by the Tsinghua-Toyota Joint Research Institute Inter-disciplinary Program (Toyota-Future City Simulator). Additional support was provided the Zhongguancun Academy under Grant No. XTS0074.

References

[1] Meng Chen, Zechen Li, Weiming Huang, Yongshun Gong, and Yilong Yin. 2024. Profiling urban streets: A semi-supervised prediction model based on street view imagery and spatial topology. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 319–328.

[2] An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. 2024. Spatialrgpt: Grounded spatial reasoning in vision-language models. *Advances in Neural Information Processing Systems* 37 (2024), 135062–135093.

[3] Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. 2025. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. *arXiv preprint arXiv:2501.17161* (2025).

[4] Yihe Deng, Hritik Bansal, Fan Yin, Nanyun Peng, Wei Wang, and Kai-Wei Chang. 2025. Openvlthinker: An early exploration to complex vision-language reasoning via iterative self-improvement. *arXiv e-prints* (2025), arXiv–2503.

[5] Zhuangyuan Fan, Fan Zhang, Becky P. Y. Loo, and Carlo Ratti. 2023. Urban visual intelligence: Uncovering hidden city profiles with street view images. *Proceedings of the National Academy of Sciences* 120, 27 (2023), e2220417120. doi:10.1073/pnas.2220417120

[6] Jie Feng, Tianhui Liu, Yuwei Du, Siqi Guo, Yuming Lin, and Yong Li. 2025. Citygpt: Empowering urban spatial cognition of large language models. In *Proceedings of the 31st ACM SIGKDD conference on knowledge discovery and data mining v. 2*. 591–602.

[7] Jie Feng, Shengyuan Wang, Tianhui Liu, Yanxin Xi, and Yong Li. 2025. Urbanllava: A multi-modal large language model for urban intelligence. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 6209–6219.

[8] Jie Feng, Jun Zhang, Tianhui Liu, Xin Zhang, Tianjian Ouyang, Junbo Yan, Yuwei Du, Siqi Guo, and Yong Li. 2025. Citybench: Evaluating the capabilities of large language models for urban tasks. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*. 5413–5424.

[9] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. 2025. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature* 645, 8081 (2025), 633–638.

[10] Qiushan Guo, Shalini De Mello, Hongxu Yin, Wonmin Byeon, Ka Chun Cheung, Yizhou Yu, Ping Luo, and Sifei Liu. 2024. Regiongpt: Towards region understanding vision language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13796–13806.

[11] Sungwon Han, Donghyun Ahn, Seungeon Lee, Minhyuk Song, Sungwon Park, Sangyoon Park, Jihee Kim, and Meeyoung Cha. 2024. Geosee: Regional socio-economic estimation with a large language model. *arXiv preprint arXiv:2406.09799* (2024).

[12] Xixuan Hao, Wei Chen, Yibo Yan, Siru Zhong, Kun Wang, Qingsong Wen, and Yuxuan Liang. 2025. UrbanVLP: Multi-granularity vision-language pretraining for urban socioeconomic indicator prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 28061–28069.

[13] Jun He, Yi Lin, Zilong Huang, Jiacong Yin, Junyan Ye, Yuchuan Zhou, Weijia Li, and Xiang Zhang. 2025. Urbanfeel: A comprehensive benchmark for temporal and perceptual understanding of city scenes through human perspective. *arXiv preprint arXiv:2509.22228* (2025).

[14] Wenxuan Huang, Bohan Jia, Shaosheng Cao, Zheyu Ye, Zhe Xu, Yao Hu, Shaohui Lin, et al. 2026. Vision-r1: Incentivizing reasoning capability in multimodal large language models. In *International Conference on Learning Representations*, Vol. 2026. 63794–63812.

[15] Yuhao Kang, Fan Zhang, Song Gao, Hui Lin, and Yu Liu. 2020. A review of urban physical environment sensing using street view imagery in public health studies. *Annals of GIS* 26, 3 (2020), 261–275.

[16] Sicong Leng, Jing Wang, Jiaxi Li, Hao Zhang, Zhiqiang Hu, Boqiang Zhang, Yuming Jiang, Hang Zhang, Xin Li, Lidong Bing, Deli Zhao, Wei Lu, Yu Rong, Aixin Sun, and Shijian Lu. 2025. MMR1: Enhancing Multimodal Reasoning with Variance-Aware Sampling and Open Resources. arXiv:2509.21268 [cs.CV] https://arxiv.org/abs/2509.21268

[17] Ling Li, Yu Ye, Bingchuan Jiang, and Wei Zeng. 2024. GeoReasoner: Geo-localization with Reasoning in Street Views using a Large Vision-Language Model. In *International Conference on Machine Learning (ICML)*.

[18] Ling Li, Yao Zhou, Yuxuan Liang, Fugee Tsung, and Jiaheng Wei. 2025. Recognition through Reasoning: Reinforcing Image Geo-localization with Large Vision-Language Models. *arXiv preprint arXiv:2506.14674* (2025).

[19] Zhuoheng Li, Yaochen Wang, Zhixue Song, Yuqi Huang, Rui Bao, Guanjie Zheng, and Zhenhui Jessie Li. 2024. What can LLM tell us about cities? *arXiv preprint arXiv:2411.16791* (2024).

[20] Tianhui Liu, Jie Feng, Hetian Pang, Xin Zhang, Tianjian Ouyang, Zhiyuan Zhang, and Yong Li. 2025. CityLens: Benchmarking Large Language-Vision Models for Urban Socioeconomic Sensing. *arXiv preprint arXiv:2506.00530* (2025).

[21] Yu Liu, Xin Zhang, Jingtao Ding, Yanxin Xi, and Yong Li. 2023. Knowledge-infused contrastive learning for urban imagery-based socioeconomic prediction. In *Proceedings of the ACM web conference 2023*. 4150–4160.

[22] Rohin Manvi, Samar Khanna, Marshall Burke, David Lobell, and Stefano Ermon. 2024. Large language models are geographically biased. *arXiv preprint arXiv:2402.02680* (2024).

[23] Rohin Manvi, Samar Khanna, Gengchen Mai, Marshall Burke, David B. Lobell, and Stefano Ermon. 2024. GeoLLM: Extracting Geospatial Knowledge from Large Language Models. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=TqL2xBwXP3

[24] Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Tiancheng Han, Botian Shi, Wenhai Wang, Junjun He, et al. 2025. Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2503.07365* (2025).

[25] Qirui Mi, Qipeng Yang, Zijun Fan, Wentian Fan, Heyang Ma, Chengdong Ma, Siyu Xia, Bo An, Jun Wang, and Haifeng Zhang. 2025. EconGym: A Scalable AI Testbed with Diverse Economic Tasks. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

[26] Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. 2023. Teaching CLIP to Count to Ten. arXiv:2302.12066 [cs.CV] https://arxiv.org/abs/2302.12066

[27] Chen Qian, Dongrui Liu, Haochen Wen, Zhen Bai, Yong Liu, and Jing Shao. 2025. Demystifying Reasoning Dynamics with Mutual Information: Thinking Tokens are Information Peaks in LLM Reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=E1FrjgaG1J

[28] Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, et al. 2025. Vlm-r1: A stable and generalizable r1-style large vision-language model. *arXiv preprint arXiv:2504.07615* (2025).

[29] Yifan Shen, Yuanzhe Liu, Jingyuan Zhu, Xu Cao, Xiaofeng Zhang, Yixiao He, Wenming Ye, James Matthew Rehg, and Ismini Lourentzou. 2025. Fine-Grained Preference Optimization Improves Spatial Reasoning in VLMs. *arXiv preprint arXiv:2506.21656* (2025).

[30] Zhaochen Su, Peng Xia, Hangyu Guo, Zhenhua Liu, Yan Ma, Xiaoye Qu, Jiaqi Liu, Yanshu Li, Kaide Zeng, Zhengyuan Yang, et al. 2025. Thinking with images for multimodal reasoning: Foundations, methods, and future frontiers. *arXiv preprint arXiv:2506.23918* (2025).

[31] Yaoting Wang, Shengqiong Wu, Yuecheng Zhang, Shuicheng Yan, Ziwei Liu, Jiebo Luo, and Hao Fei. 2025. Multimodal chain-of-thought reasoning: A comprehensive survey. *arXiv preprint arXiv:2503.12605* (2025).

[32] Yana Wei, Liang Zhao, Jianjian Sun, Kangheng Lin, Jingcheng Hu, Yinmin Zhang, En Yu, Haoran Lv, Zejia Weng, Jia Wang, et al. 2026. Open vision reasoner: Transferring linguistic cognitive behavior for visual reasoning. *Advances in Neural Information Processing Systems* 38 (2026), 4463–4494.

[33] Penghao Wu and Saining Xie. 2024. V*: Guided visual search as a core mechanism in multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13084–13094.

[34] Tianhe Wu, Jian Zou, Jie Liang, Lei Zhang, and Kede Ma. 2025. VisualQuality-R1: Reasoning-Induced Image Quality Assessment via Reinforcement Learning to Rank. *arXiv preprint arXiv:2505.14460* (2025).

[35] Jiaer Xia, Yuhang Zang, Peng Gao, Sharon Li, and Kaiyang Zhou. 2025. Visionary-r1: Mitigating shortcuts in visual reasoning with reinforcement learning. *arXiv preprint arXiv:2505.14677* (2025).

[36] Tong Xiao, Xin Xu, Zhenya Huang, Hongyu Gao, Quan Liu, Qi Liu, and Enhong Chen. 2025. Advancing Multimodal Reasoning Capabilities of Multimodal Large Language Models via Visual Perception Reward. *arXiv preprint arXiv:2506.07218* (2025).

[37] Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. 2024. Llava-cot: Let vision language models reason step-by-step. *arXiv preprint arXiv:2411.10440* (2024).

[38] Weiye Xu, Jiahao Wang, Weiyun Wang, Zhe Chen, Wengang Zhou, Aijun Yang, Lewei Lu, Houqiang Li, Xiaohua Wang, Xizhou Zhu, et al. 2026. Visulogic: A benchmark for evaluating visual reasoning in multi-modal large language models. In *International Conference on Learning Representations*, Vol. 2026. 25966–26003.

[39] Yibo Yan, Haomin Wen, Siru Zhong, Wei Chen, Haodong Chen, Qingsong Wen, Roger Zimmermann, and Yuxuan Liang. 2024. Urbanclip: Learning text-enhanced urban region profiling with contrastive language-image pretraining from the web. In *Proceedings of the ACM Web Conference 2024*. 4006–4017.

[40] Xixian Yong and Xiao Zhou. 2024. MuseCL: predicting urban socioeconomic indicators via multi-semantic contrastive learning. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*. 7536–7544.

[41] Weixia Zhang, Kede Ma, Guangtao Zhai, and Xiaokang Yang. 2021. Uncertainty-aware blind image quality assessment in the laboratory and wild. *IEEE Transactions on Image Processing* 30 (Mar. 2021), 3474–3486.

[42] Xin Zhang, Tianjian Ouyang, Yu Shang, Qingmin Liao, and Yong Li. 2025. UrbanMLLM: Joint Learning of Cross-view Imagery for Urban Understanding. https://openreview.net/forum?id=YBht9Vp5vC

[43] Yunke Zhang, Ruolong Ma, Xin Zhang, and Yong Li. 2025. Perceiving Urban Inequality from Imagery Using Visual Language Models with Chain-of-Thought Reasoning. In *Proceedings of the ACM on Web Conference 2025*. Association for Computing Machinery, 5342–5351. https://doi.org/10.1145/3696410.3714536

[44] Zhilun Zhou, Jingyang Fan, Yu Liu, Fengli Xu, Depeng Jin, and Yong Li. 2024. Synergizing llm agents and knowledge graph for socioeconomic prediction in lbsn. *arXiv preprint arXiv:2411.00028* (2024).

[45] Zhilun Zhou, Yu Liu, Jingtao Ding, Depeng Jin, and Yong Li. 2023. Hierarchical knowledge graph learning enabled socioeconomic indicator prediction in location-based social network. In *Proceedings of the ACM web conference 2023*. 122–132.