PlanGPT-VL: Enhancing Urban Planning with Domain-Specific Vision-Language Models

He Zhu[1]*, Junyou Su[1]*, Minxin Chen[1]*, Wen Wang[1], Yijie Deng[1]
Guanhua Chen[2], Wenjia Zhang[1]
[1]Behavioral and Spatial AI Lab, Peking University & Tongji University
[2]Southern University of Science and Technology
zhuye140@gmail.com,wenjiazhang@pku.edu.cn

Abstract

In the field of urban planning, existing Vision-Language Models (VLMs) frequently fail to effectively analyze planning maps, which are critical for urban planners and educational contexts. Planning maps require specialized understanding of spatial configurations, regulatory requirements, and multi-scale analysis. To address this challenge, we introduce **PlanGPT-VL**, the first domain-specific VLM tailored for urban planning maps. PlanGPT-VL employs three innovations: (1) PlanAnno-V framework for high-quality VQA data synthesis, (2) Critical Point Thinking (CPT) to reduce hallucinations through structured verification, and (3) PlanBench-V benchmark for systematic evaluation. Evaluation on PlanBench-V shows that PlanGPT-VL outperforms general-purpose VLMs on planning map interpretation tasks, with our 7B model achieving performance comparable to larger 72B models.

1 Introduction

Vision-Language Models (VLMs) have achieved remarkable progress across general multimodal tasks, from image understanding (Hurst et al., 2024; DeepMind, 2023) to visual reasoning (Zhu et al., 2025; Guo et al., 2025) and multimodal dialogue (Liu et al., 2023; Wang et al., 2024a). Recent works have successfully adapted these models to specialized domains including medical imaging (Li et al., 2023a; Lai et al., 2025; Pan et al., 2025), geographical information systems (Zhang et al., 2024c,b), and mathematical reasoning (Chen et al., 2025a; Shen et al., 2025). However, urban planning remains a critical yet underexplored domain that demands specialized VLMs capable of *interpreting complex planning maps*—a task where even state-of-the-art commercial models exhibit substantial limitations.

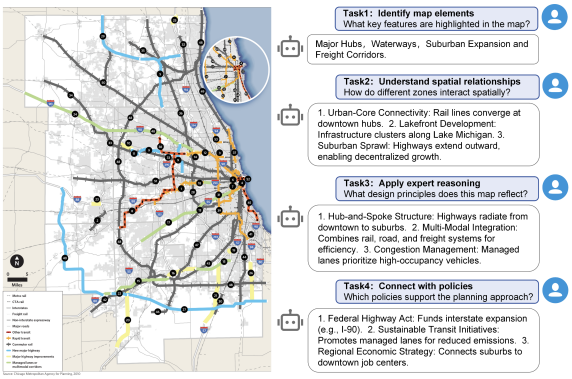


Figure 1: Urban planning multimodal tasks including map elements identification, spatial relationships understanding, expert reasoning, policy association and other key applications.

Planning maps are specialized visual representations that encode land use zones, transportation networks, and development policies through domain-specific symbols, color-coding systems, and professional annotations (Lynch and Hack, 1984; Steinitz, 1995; Healey, 1997). As shown in Figure 1, urban planning involves diverse multimodal tasks requiring both visual perception and domain expertise. Current general-purpose VLMs face three critical limitations: (1) high hallucination rates when processing information-dense planning maps, (2) responses misaligned with professional communication standards, and (3) lack of reliable evaluation frameworks for specialized map interpretation. These challenges stem from the scarcity of domain-specific training data and the prohibitive cost of expert annotation (Liu et al., 2024b; Li et al., 2023b).

We introduce **PlanGPT-VL**, the first specialized VLM for urban planning map interpretation. Our approach centers on three key innovations: (1) **PlanAnno-V**, a framework for synthesizing high-quality instruction-response pairs through expert-guided data preprocessing and professional communication alignment; (2) **Critical Point Think-**

* Equal contribution.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 2461–2483
November 4-9, 2025

ing (CPT), a novel methodology that reduces hallucinations via decomposition of complex visual information into verifiable critical points; and (3) **PlanBench-V**, the first comprehensive benchmark for evaluating VLM performance on planning map interpretation. PlanGPT-VL outperforms both open-source and commercial VLMs by 59.2% on average, with our 7B model achieving comparable performance to models exceeding 72B parameters. Our contributions include: (1) PlanGPT-VL-2B/7B, the first specialized VLM for urban planning with state-of-the-art performance and compact size; (2) PlanAnno-V framework for efficient high-quality training data generation; and (3) PlanBench-V for systematic evaluation of planning map interpretation capabilities.

2 Related Works

Domain-Specific Language and Vision-Language Models Large language models have evolved from general-purpose systems (OpenAI, 2023, 2022; Touvron et al., 2023; et al., 2023b; Anthropic, 2023; Mistral-AI, 2023; DeepMind, 2023) to specialized applications. Chinese models include DeepSeek, Baichuan (Baichuan, 2023), GLM (Du et al., 2022), and Qwen. Domain adaptation has produced specialized systems in medicine (HuaTuo (Wang et al., 2023), DoctorGLM (Xiong et al., 2023)), legal (ChatLaw (Cui et al., 2023)), finance (XuanYuan 2.0 (Zhang et al., 2023b)), and mathematics (MathGPT (Tycho Young, 2023)). Urban environment models include PlanGPT (Zhu et al., 2024b) for text-based planning, TrafficGPT (Zhang et al., 2023a) for transportation, NASA's Prithvi (et al., 2023a) for climate, and CityGPT (Feng et al., 2024) for spatial reasoning. However, none addresses visual interpretation of planning maps with specialized representational requirements, motivating PlanGPT-VL as the first vision-language model for urban planning map interpretation.

Multimodal Instruction Data Synthesis VLM effectiveness relies on high-quality instruction-response pairs, with synthetic data pipelines similar to our PlanAnno-V framework. General approaches like MAmmoTH-VL (Guo et al., 2024), MMInstruction (Liu et al., 2024a), and Infinity-Multimodal (Gu et al., 2024) employ template-based sampling, while OASIS (Zhang et al., 2025) uses visual prompting. Relevant to our Critical Point Thinking, MM-Verify (Sun et al., 2025) introduces verification mechanisms, and LLaVA-CoT (Xu et al., 2025) implements structured reasoning synthesis. Our PlanAnno-V framework extends these through domain-specific preprocessing, professional language alignment, and specialized verification for urban planning maps.

3 Method

To address the challenges of specialized planning map interpretation and training data scarcity, we introduce the PlanAnno-V framework with three key innovations: (1) systematic data synthesis for high-quality instruction-response pairs (Section 3.1); (2) Critical Point Thinking (CPT) for hallucination mitigation (Section 3.3); and (3) PlanBench-V for reliable evaluation (Section 3.4).

3.1 Overview of PlanAnno-V

The PlanAnno-V framework synthesizes high-quality visual instruction tuning data through three stages, as depicted in Figure 2: **Stage (1): Domain-Specific Data Preprocessing** involves collecting 5,000 planning maps from urban planning bureaus, applying diversity-based filtering to select 1,050 representative maps, and obtaining expert annotations on 800 high-quality examples from 50 selected maps. Details are in Appendix A. **Stage (2): Instruction-Response Synthesis** combines diversity-enhanced instruction generation (Section 3.2) that preserves professional expertise while expanding distributional coverage, and Critical Point Thinking (Section 3.3) for verifiable response synthesis that reduces hallucination. **Stage (3): Model-Specific Rewriting** employs target models to align responses with professional planning communication styles, incorporating planner examples as in-context demonstrations to ensure domain-appropriate linguistic patterns.

3.2 Distributional Instruction Synthesis

Manual annotation by domain experts yields high-quality, professionally relevant instructions but inherently suffers from limited diversity and complexity coverage (Wang et al., 2022). Our approach preserves this professional expertise while systematically expanding the distributional coverage through principled synthesis methods.

Instruction Spectrum Construction We begin with 1k professionally curated instructions from

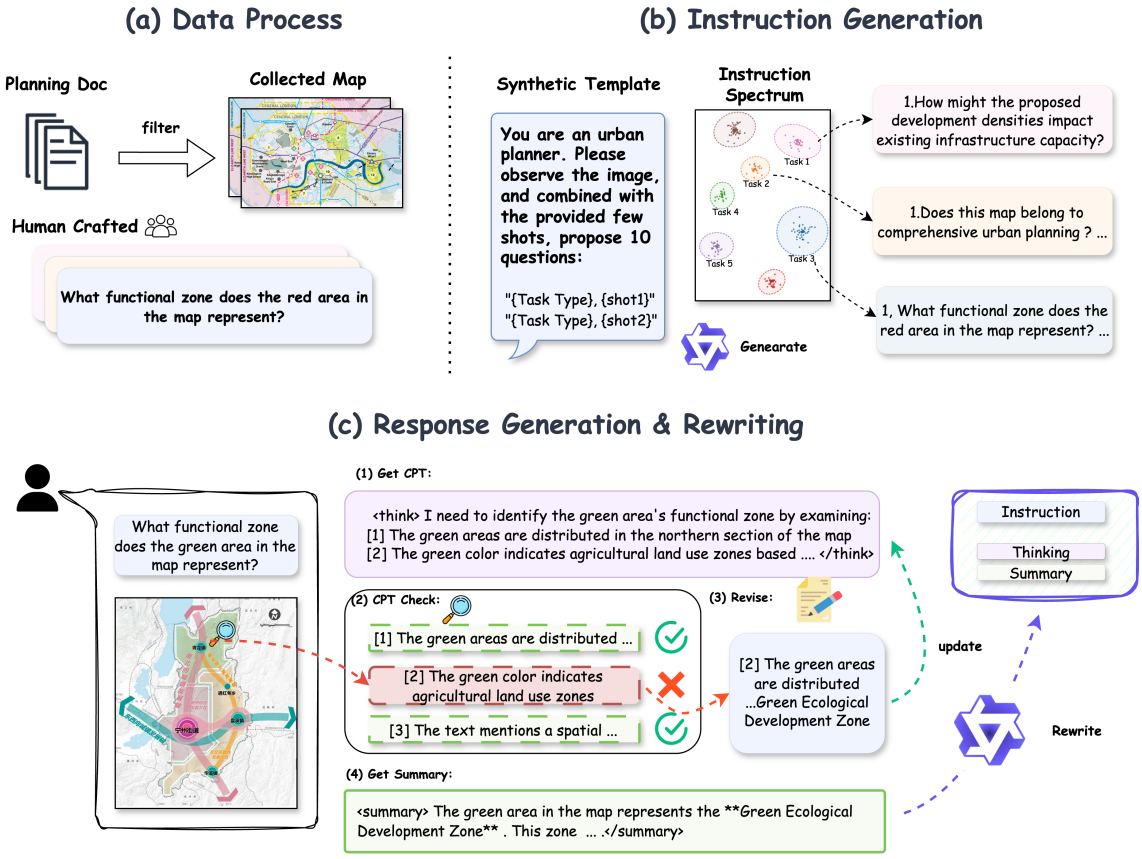


Figure 2: Overview of PlanAnno-V framework. Our approach synthesizes high-quality instruction-response pairs through a three-stage process: (1) domain-specific data preprocessing with expert annotation, (2) instruction-response synthesis using Critical Point Thinking for hallucination reduction, and (3) model-specific rewriting to align with professional planning communication patterns.

urban planning experts. Each instruction undergoes automated intent extraction via InstaTagger (Lu et al., 2023), identifying semantic components (e.g., "spatial_analysis", "location_identification"). Through clustering analysis, we categorize instructions into 8 distinct task types and establish a complexity hierarchy based on average intent count, creating a comprehensive instruction spectrum.

Systematic Distributional Expansion Following recent advances in automated instruction generation (Liu et al., 2024b; Luo et al., 2024; Zhu et al., 2024a), we implement stratified replication to expand beyond seed limitations. For each planning image i, we sample diverse task types $\mathcal{T} = \{t_1, t_2, ..., t_{10}\}$ and exemplars $\mathcal{E} = \{e_1, e_2, ..., e_{10}\}$ as few-shot demonstrations. The instruction generation process is formalized as: $q_{new} \sim p(\cdot|i, \{(t_j, e_j)\}_{j=1}^{10}, \phi_{div})$ where q_{new} represents the synthesized instruction, and ϕ_{div} denotes diversification prompts encouraging task variety and complexity progression.

3.3 Critical Point Thinking for Hallucination Mitigation

Inspired by Yu et al. (2024), we introduce *Critical Point Thinking* (CPT), which decomposes reasoning into structured, verifiable components to systematically reduce factual errors through iterative verification and correction.

Algorithm 1 Critical Point Thinking (CPT)

Require: Planning map m, instruction i, verification threshold τ
Ensure: Verified response r_{final}

$\mathcal{P} \leftarrow \text{ExtractCriticalPoints}(m, i)$
for each $p_j \in \mathcal{P}$ **do**
  $q_j \leftarrow \text{FormulateVerificationQuery}(p_j, i)$
  $v_j \leftarrow \text{VerifyPoint}(q_j, m)$
  if $v_j < \tau$ **then**
    $p_j \leftarrow \text{CorrectPoint}(p_j, m, q_j)$
$\mathcal{P}_{merged} \leftarrow \text{MergeRedundantPoints}(\mathcal{P})$
$r_{final} \leftarrow \text{ReconstructResponse}(\mathcal{P}_{merged})$
return r_{final}

As shown in Algorithm 1, CPT employs a systematic "Generate-Verify-Revise" paradigm. The key insight is that models excel at focused verification tasks compared to open-ended generation (Hurst et al., 2024). We first extract critical points in structured format, then verify each atomic claim through targeted queries against the planning map and correct identified errors. Finally, we eliminate redundancy to mitigate overthinking (Chen et al., 2025b). Ablation experiments in Section 2 demonstrate effectiveness.

3.4 PlanBench-V: A Benchmark for Urban Planning VLMs

We introduce PlanBench-V, the first comprehensive benchmark for assessing VLM performance on planning map interpretation. PlanBench-V consists of 300 carefully curated examples spanning zoning analysis, infrastructure assessment, spatial reasoning, and regulatory compliance, with distribution shown in Figure 3(c). Each example is annotated by three professional urban planners. To address the open-ended nature of planning inquiries, we establish a multi-dimensional scoring framework where each question has n expert-defined evaluation criteria $\{c_1, c_2, ..., c_n\}$. We compute a normalized score $S = \frac{\sum_{i=1}^{n} \mathbb{I}(c_i \in R)}{n}$, where $\mathbb{I}(\cdot)$ indicates criteria satisfaction. This enables objective evaluation while maintaining alignment with professional standards. Details are in Appendix G.

3.5 Training Methodology

We implement PlanGPT-VL by fine-tuning Qwen2-VL-7B-Instruct (Wang et al., 2024a) while freezing vision encoder and projector layers to preserve general visual understanding. This strategy prevents severe degradation of general capabilities, as detailed in Section 5.2. We conduct Supervised Fine-Tuning using single-image QA pairs and multi-turn dialogues from PlanAnno-V, employing rejection sampling (Touvron et al., 2023) at inference for enhanced response quality. Detailed dataset analysis is provided in Appendix D and Appendix F.

4 Experiments

4.1 Experimental Setup

We fine-tune Qwen2-7B-VL-Instruct using VERL framework (Sheng et al., 2024) on 4 A100 GPUs with AdamW optimizer (2e-5 learning rate, 3 epochs). Our training corpus comprises 10k instruction-response pairs synthesized from 1k planning maps using PlanAnno-V. Evaluation encompasses both our PlanBench-V benchmark and standard VLM benchmarks via lmms-eval (Zhang et al., 2024a). Complete implementation details are provided in Appendix C.

4.2 Main Results

We evaluate PlanGPT-VL against state-of-the-art VLMs on domain-specific planning tasks. Table 1 presents comprehensive results on PlanBench-V. Key findings include: (1) Proprietary models like GPT-4o (1.342) underperform leading open-source alternatives, potentially reflecting training biases toward general rather than domain-specific content. (2) All models perform stronger in Description and Classification tasks but weaker in Evaluation and Decision Making dimensions requiring deeper domain expertise. (3) PlanGPT-VL-7B achieves top overall performance (1.566), with notable advantages in Professional Reasoning (1.729) and Implementation (1.520) dimensions where domain-specific training provides significant benefit, while maintaining competitive performance across all task categories.

4.3 Ablation Studies

We conduct extensive ablation studies to understand each component's contribution in the PlanAnno-V framework. Table 2 presents comprehensive comparisons of different model variants, evaluating both PlanBench-V performance (domain-specific) and general vision-language benchmarks. All PlanBench-V scores are normalized with GPT-4V as baseline (1.0).

Our ablation analysis reveals several key insights. **Critical Point Thinking (CPT)** provides modest improvements over standard Chain-of-Thought (+2.3% with 72B teacher), but enables effective verification mechanisms. **Verification component** significantly improves performance (+7.2% overall), particularly benefiting Implementation dimension (+19.2%), validating our hypothesis that iterative verification reduces hallucinations. **Teacher model quality** yields the most substantial gains—upgrading from 72B to 32B teacher models provides +25.0% overall improvement, with strong gains in Reasoning (+29.0%) and Association (+38.9%) tasks.

However, domain-specific fine-tuning creates a trade-off with general capabilities. MMMU scores drop by 11.2% (51.6 → 45.8) and GQA by 10.1%, while POPE remains stable (88.3 → 88.5), sug-

Model	PlanBench-V (Detailed Categories)								PlanBench-V (Main Categories)				Overall
	Element	Eval	Class	Assoc	Spatial	Prof	Desc	Dec	Perc	Reas	Assoc	Impl	
General Vision-Language Models													
Qwen2-VL-2B-Instruct	0.744	0.537	0.948	0.926	0.500	0.656	0.925	0.792	0.767	0.664	0.926	0.616	0.731 (-0.179)
Qwen2-VL-7B-Instruct	0.902	0.857	1.031	0.979	0.716	0.943	1.386	0.657	0.964	0.878	0.979	0.795	0.910 (base)
Qwen2-VL-72B-Instruct-AWQ	1.010	0.670	1.125	1.114	0.746	0.967	1.367	0.632	1.056	0.920	1.114	0.658	0.963 (+0.053)
Qwen2.5-VL-3B-Instruct	0.862	0.697	0.953	0.970	0.691	0.870	1.554	0.936	0.951	0.822	0.970	0.771	0.876 (-0.034)
Qwen2.5-VL-7B-Instruct	1.101	0.802	1.089	1.069	0.865	1.110	1.628	1.054	1.168	1.013	1.069	0.880	1.050 (+0.140)
Qwen2.5-VL-32B-Instruct	1.432	1.678	1.578	1.685	1.539	1.791	1.928	1.620	1.496	1.649	1.685	1.660	1.616 (teacher)
Qwen2.5-VL-72B-Instruct-AWQ	1.299	1.153	1.406	1.253	1.248	1.263	1.825	1.090	1.366	1.289	1.253	1.134	1.288 (+0.378)
InternVL3-8B	0.992	0.631	1.026	0.926	0.798	0.751	1.783	1.073	1.094	0.831	0.926	0.768	0.909 (-0.001)
InternVL3-9B	1.173	0.921	1.297	1.297	1.260	1.435	1.878	0.903	1.263	1.339	1.297	0.916	1.271 (+0.361)
InternVL3-14B	0.931	0.709	1.177	1.098	0.793	0.998	1.580	0.917	1.014	0.962	1.098	0.773	0.980 (+0.070)
GPT-4o-mini	0.664	0.636	1.021	0.963	0.789	1.030	0.890	1.175	0.693	0.938	0.963	0.803	0.866 (-0.044)
GPT-4o	1.051	1.223	1.260	1.527	1.305	1.564	1.708	1.429	1.136	1.399	1.527	1.287	1.342 (+0.432)
Our Models													
PlanGPT-VL-2B	1.174	1.305	1.219	1.453	1.328	1.485	1.744	**1.567**	1.247	1.366	1.453	1.386	1.352 (+0.442)
PlanGPT-VL-7B	**1.417**	**1.528**	**1.541**	**1.537**	**1.569**	**1.729**	**2.000**	1.501	**1.492**	**1.627**	**1.537**	**1.520**	**1.566** (+0.656)

Table 1: Performance comparison on PlanBench-V with detailed and main categories. Detailed categories: Element = Element Recognition, Eval = Evaluation, Class = Classification, Assoc = Association, Spatial = Spatial Relations, Prof = Professional Reasoning, Desc = Description, Dec = Decision Making. Main categories: Perc = Perception, Reas = Reasoning, Assoc = Association, Impl = Implementation.

Model Variant	PlanBench-V (Domain-Specific)					General Benchmarks		
	Perception	Reasoning	Association	Implementation	Overall	MMMU	GQA	POPE
Baseline Model								
Qwen2-VL-7B-Instruct	0.964	0.878	0.979	0.795	0.910	**51.6**	**62.3**	88.3
Qwen2.5-72B-VL-Instruct Teacher Models								
+ CoT	1.172 (+0.208)	1.159 (+0.281)	1.108 (+0.129)	0.904 (+0.109)	1.129 (+0.219)	48.3 (-3.3)	61.5 (-0.8)	88.7 (+0.4)
+ CPT	1.231 (+0.267)	1.196 (+0.318)	1.013 (+0.034)	0.990 (+0.195)	1.155 (+0.245)	49.4 (-2.2)	**62.3** (+0.0)	**89.0** (+0.7)
+ CPT + Verification	1.326 (+0.362)	1.225 (+0.347)	1.172 (+0.193)	1.180 (+0.385)	1.238 (+0.328)	49.1 (-2.5)	61.5 (-0.8)	88.7 (+0.4)
Qwen2.5-32B-VL-Instruct Teacher Models								
+ CPT	1.464 (+0.500)	1.580 (+0.702)	**1.628** (+0.649)	1.464 (+0.669)	1.547 (+0.637)	46.3 (-5.3)	60.5 (-1.8)	88.5 (+0.2)
+ CPT + Verification	**1.492** (+0.528)	**1.627** (+0.749)	1.537 (+0.558)	**1.520** (+0.725)	**1.566** (+0.656)	45.8 (-5.8)	56.0 (-6.3)	88.5 (+0.2)

Table 2: Comprehensive ablation study results on PlanBench-V and general vision-language benchmarks.

gesting preserved object detection. Our method achieves +72.1% improvement on domain tasks, but at the cost of general capabilities. We aim to address this limitation through mixed training strategies incorporating general benchmark data.

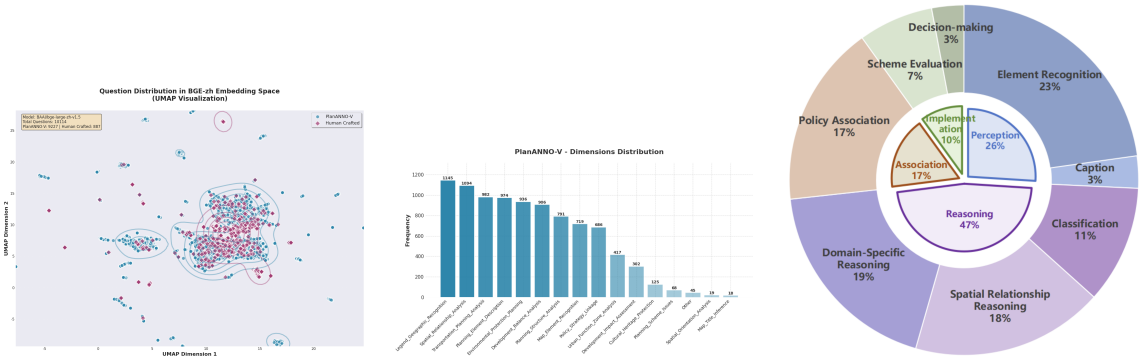
5 Analysis

5.1 Instruction Synthesis Effectiveness Analysis

To validate our automated instruction generation pipeline, we analyze PlanAnno-V synthesized instructions across three dimensions: **(1) Distributional Alignment** Using bge-zh-base embeddings (Xiao et al., 2024), we compute distributional metrics between seed and synthesized data, revealing strong semantic alignment (Cosine Similarity = 0.9350) with controlled variation (MMD = 0.0515). Figure 3(a) shows our method maintains core distribution while introducing beneficial diversity. **(2) Categorical Expansion** Intent tagging analysis shows synthesis expands coverage from 8 to 15 planning categories while maintaining proportional representation of core tasks, preventing categorical drift. **(3) Quality Preservation** Professional evaluation of 100 sampled pairs shows synthesized instructions maintain comparable quality to expert seed data (planning expertise: 0.87 vs. 0.89; correctness: 0.85 vs. 0.88; fluency: 0.91 vs. 0.90) with no significant differences. Detailed analysis is provided in Appendix E.

5.2 Preserving General Capabilities While Enhancing Domain Expertise

We investigate how to prevent general visual capability degradation while enhancing planning specialization. We explore three key strategies: (1) Data Mixing: incorporating 5k examples from ShareGPT4-V (Chen et al., 2024); (2) Architecture Modifications: unfreezing the vision encoder; and (3) Caption Training: including or removing caption data. Table 3 presents our findings across both planning expertise (PlanBench-V) and gen-



(a) (b) (c)

Figure 3: Analysis of PlanAnno-V instruction synthesis: (a) UMAP projection of instruction embeddings showing synthesized instructions (blue) maintain similar distribution to expert-annotated seed data (red) while introducing beneficial diversity; (b) Categorical distribution of synthesized instructions; (c) Statistical distribution of PlanBench-V Dataset.

	Model	# Training Data	Planning Skill	General Skill	Avg
1	Qwen2-VL-7B-Instruct	-	0.910	**51.6**	**26.26**
2	PlanGPT-VL	11k	1.566	45.8	23.68
3	2 (w/ Mix Data)	**16k**	1.52	47.3	24.41
4	2 (unfreeze Vision tower)	11k	1.54	44.3	22.92
5	2 (w/o caption)	10k	**1.59**	45.4	23.50

Table 3: Comparison of training configurations showing the trade-off between planning expertise and general visual understanding.

eral visual understanding (MMMU). Our analysis reveals that the optimal configuration combines mixed-domain data, a frozen vision encoder, and caption integration. Our findings demonstrate that a balanced approach with mixed training data and caption integration effectively preserves general visual capabilities while enhancing domain-specific expertise.

5.3 Base Model Architecture and Data Scaling Analysis

Following the configuration in Section 4, we analyze PlanGPT-VL's performance across architectures and training data scales (Table 4). Qwen2-VL-7B-Instruct achieves highest performance (1.566), while LLaVA-1.5-7B shows largest improvement (+127.5%). Smaller models demonstrate substantial gains, suggesting effective domain specialization. For data scaling, performance improves dramatically from 100 images (-4.1%) to 500 images (+73.6%), then stabilizes with 1,000 images (+72.1%). This demonstrates that our PlanAnno-V framework consistently improves performance across models of different sizes once a minimum data threshold is reached, highlighting the effectiveness of our approach.

Base Model Analysis				
Model	**Params**	**Original**	**Ours**	**Improv.**
Qwen2-2B	2B	0.731	1.352	+85.0%
Qwen2-7B	7B	0.910	**1.566**	+72.1%
LLaVa-7B	7B	0.171	0.389	**+127.5%**
LLaVa-13B	13B	0.223	0.474	+113.6%
Training Data Analysis				
Data Size	**Params**	**Original**	**Ours**	**Improv.**
100 imgs	7B	0.910	0.873	-4.1%
500 imgs	7B	0.910	**1.580**	**+73.6%**
1,000 imgs	7B	0.910	1.566	+72.1%

Table 4: Analysis of model architectures and training data configurations.

6 Conclusions

We introduce PlanGPT-VL, the first domain-specific Vision-Language Model for urban planning map interpretation. Through our PlanAnno-V framework, Critical Point Thinking methodology, and PlanBench-V benchmark, we address challenges of data scarcity, hallucination reduction, and evaluation in this specialized domain. Experiments show that PlanGPT-VL outperforms general-purpose VLMs on planning tasks, with our 7B model achieving performance comparable to larger 72B models. This work advances AI applications in urban planning and provides a blueprint for developing specialized VLMs in other domains. PlanGPT-VL offers planners, policymakers, and educators a reliable tool for map analysis and decision support.

Limitations

Despite PlanGPT-VL's improvements, several limitations remain. While our Critical Point Thinking approach substantially reduces hallucinations, complete elimination of factual errors remains challenging, particularly for complex planning maps with ambiguous visual elements or when interpreting multiple scales simultaneously. Additionally, our approach requires a trade-off between domain specialization and general capabilities, as evidenced by performance degradation on general benchmarks. Our model's effectiveness is also constrained by training data diversity, with current implementation primarily focused on Chinese urban planning contexts. Future work should address these limitations through enhanced verification mechanisms, balanced training strategies, and expanded cross-cultural planning data.

Acknowledgements

The paper is supported by the Key Project of the Shanghai Municipal Education Commission's AI-Enabled Research Paradigm Reform and Discipline Leap Program (Development of a Domain-Specific Large Language Model in the Field of Urban and Rural Planning for Enhancing Spatial Cognition and Decision-Making Capabilities) and by the Fundamental Research Funds for the Central Universities (22120250239). Guanhua was supported by National Natural Science Foundation of China (No. 62306132). We would like to thank the support from the Spatial Planning Bureau of the Ministry of Natural Resources of China, the China Land Surveying and Planning Institute, the Planning and Natural Resources Bureau of Shenzhen Municipality, the Planning and Research Center of Guangzhou Municipality, the Shenzhen Marine Development Promotion Research Center, the China Academy of Urban Planning and Design, and Guangzhou Planning Corporation. We would also like to express our sincere gratitude to all members of the BSAI Lab for their invaluable support.

References

Anthropic. 2023. Model card and evaluations for claude models.

Baichuan. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2024. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pages 370–387. Springer.

Shiqi Chen, Jinghan Zhang, Tongyao Zhu, Wei Liu, Siyang Gao, Miao Xiong, Manling Li, and Junxian He. 2025a. Bring reason to vision: Understanding perception and reasoning through model merging. *arXiv preprint arXiv:2505.05464*.

Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. 2025b. Do not think that much for 2+3=? on the overthinking of o1-like llms. *Preprint*, arXiv:2412.21187.

Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. 2023. Chatlaw. https://github.com/PKU-YuanGroup/ChatLaw.

Google DeepMind. 2023. Gemini. https://gemini.google.com.

Qianlong Du, Chengqing Zong, and Jiajun Zhang. 2023. Mods: Model-oriented data selection for instruction tuning. *arXiv preprint arXiv:2311.15653*.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Jakubik et al. 2023a. Prithvi-100M.

Rohan Anil et al. 2023b. Palm 2 technical report.

Jie Feng, Yuwei Du, Tianhui Liu, Siqi Guo, Yuming Lin, and Yong Li. 2024. Citygpt: Empowering urban spatial cognition of large language models. *Preprint*, arXiv:2406.13948.

Shuhao Gu, Jialing Zhang, Siyuan Zhou, Kevin Yu, Zhaohu Xing, Liangdong Wang, Zhou Cao, Jintao Jia, Zhuoyi Zhang, Yixuan Wang, et al. 2024. Infinity-mm: Scaling multimodal performance with large-scale and high-quality instruction data. *arXiv preprint arXiv:2410.18558*.

Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, et al. 2025. Seed1. 5-vl technical report. *arXiv preprint arXiv:2505.07062*.

Jarvis Guo, Tuney Zheng, Yuelin Bai, Bo Li, Yubo Wang, King Zhu, Yizhi Li, Graham Neubig, Wenhu Chen, and Xiang Yue. 2024. Mammoth-vl: Eliciting multimodal reasoning with instruction tuning at scale. *arXiv preprint arXiv:2412.05237*.

Patsy Healey. 1997. *Collaborative Planning: Shaping Places in Fragmented Societies*. Macmillan, London.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Siddharth Karamcheti, Suraj Nair, Ashwin Balakrishna, Percy Liang, Thomas Kollar, and Dorsa Sadigh. 2024. Prismatic vlms: Investigating the design space of visually-conditioned language models. *Preprint*, arXiv:2402.07865.

Yuxiang Lai, Jike Zhong, Ming Li, Shitian Zhao, and Xiaofeng Yang. 2025. Med-r1: Reinforcement learning for generalizable medical reasoning in vision-language models. *arXiv preprint arXiv:2503.13939*.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023a. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023b. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023c. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Preprint*, arXiv:2304.08485.

Jihao Liu, Xin Huang, Jinliang Zheng, Boxiao Liu, Jia Wang, Osamu Yoshie, Yu Liu, and Hongsheng Li. 2024a. Mm-instruct: Generated visual instructions for large multimodal model alignment. *arXiv preprint arXiv:2406.19736*.

Yangzhou Liu, Yue Cao, Zhangwei Gao, Weiyun Wang, Zhe Chen, Wenhai Wang, Hao Tian, Lewei Lu, Xizhou Zhu, Tong Lu, et al. 2024b. Mminstruct: A high-quality multi-modal instruction tuning dataset with extensive diversity. *Science China Information Sciences*, 67(12):1–16.

Keming Lu, Hongyi Yuan, Zheng Yuan, Runji Lin, Junyang Lin, Chuanqi Tan, Chang Zhou, and Jingren Zhou. 2023. #instag: Instruction tagging for analyzing supervised fine-tuning of large language models. *Preprint*, arXiv:2308.07074.

Run Luo, Haonan Zhang, Longze Chen, Ting-En Lin, Xiong Liu, Yuchuan Wu, Min Yang, Minzheng Wang, Pengpeng Zeng, Lianli Gao, Heng Tao Shen, Yunshui Li, Xiaobo Xia, Fei Huang, Jingkuan Song, and Yongbin Li. 2024. Mmevol: Empowering multimodal large language models with evol-instruct. *Preprint*, arXiv:2409.05840.

Kevin Lynch and Gary Hack. 1984. *Site Planning*. MIT Press, Cambridge, MA.

Mistral-AI. 2023. mistral. `https://mistral.ai/`.

OpenAI. 2021. Clip-vit-base-patch32. `https://huggingface.co/openai/clip-vit-base-patch32`.

OpenAI. 2022. Chatgpt. `https://chat.openai.com`.

OpenAI. 2023. Gpt-4 technical report.

Jiazhen Pan, Che Liu, Junde Wu, Fenglin Liu, Jiayuan Zhu, Hongwei Bran Li, Chen Chen, Cheng Ouyang, and Daniel Rueckert. 2025. Medvlm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning. *arXiv preprint arXiv:2502.19634*.

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*.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2024. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*.

Carl Steinitz. 1995. A framework for planning practice and education. *Landscape and Urban Planning*, 32(3):173–195.

Linzhuang Sun, Hao Liang, Jingxuan Wei, Bihui Yu, Tianpeng Li, Fan Yang, Zenan Zhou, and Wentao Zhang. 2025. Mm-verify: Enhancing multimodal reasoning with chain-of-thought verification. *arXiv preprint arXiv:2502.13383*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Krish Mangroila Tycho Young, Andy Zhang. 2023. Mathgpt - an exploration into the field of mathematics with large language models.

Haochun Wang, Chi Liu, Nuwa Xi, Zewen Qiang, Sendong Zhao, Bing Qin, and Ting Liu. 2023. Huatuo: Tuning llama model with chinese medical knowledge. *arXiv preprint arXiv:2304.06975*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *Preprint*, arXiv:2409.12191.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pages 641–649.

Honglin Xiong, Sheng Wang, Yitao Zhu, Zihao Zhao, Yuxiao Liu, Qian Wang, and Dinggang Shen. 2023. Doctorglm: Fine-tuning your chinese doctor is not a herculean task. *arXiv preprint arXiv:2304.01097*.

Guowei Xu, Peng Jin, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. 2025. Llava-cot: Let vision language models reason step-by-step. *Preprint*, arXiv:2411.10440.

Tianyu Yu, Haoye Zhang, Yuan Yao, Yunkai Dang, Da Chen, Xiaoman Lu, Ganqu Cui, Taiwen He, Zhiyuan Liu, Tat-Seng Chua, et al. 2024. Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. *arXiv preprint arXiv:2405.17220*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. 2024a. Lmms-eval: Reality check on the evaluation of large multimodal models. *Preprint*, arXiv:2407.12772.

Letian Zhang, Quan Cui, Bingchen Zhao, and Cheng Yang. 2025. Oasis: One image is all you need for multimodal instruction data synthesis. *arXiv preprint arXiv:2503.08741*.

Siyao Zhang, Daocheng Fu, Zhao Zhang, Bin Yu, and Pinlong Cai. 2023a. Trafficgpt: Viewing, processing and interacting with traffic foundation models.

Xuanyu Zhang, Qing Yang, and Dongliang Xu. 2023b. Xuanyuan 2.0: A large chinese financial chat model with hundreds of billions parameters.

Yifan Zhang, Zhengting He, Jingxuan Li, Jianfeng Lin, Qingfeng Guan, and Wenhao Yu. 2024b. Mapgpt: an autonomous framework for mapping by integrating large language model and cartographic tools. *Cartography and Geographic Information Science*, 51(6):717–743.

Yifan Zhang, Zhiyun Wang, Zhengting He, Jingxuan Li, Gengchen Mai, Jianfeng Lin, Cheng Wei, and Wenhao Yu. 2024c. Bb-geogpt: A framework for learning a large language model for geographic information science. *Information Processing & Management*, 61(5):103808.

He Zhu, Junyou Su, Tianle Lun, Yicheng Tao, Wenjia Zhang, Zipei Fan, and Guanhua Chen. 2024a. Fanno: Augmenting high-quality instruction data with open-sourced llms only. *Preprint*, arXiv:2408.01323.

He Zhu, Wenjia Zhang, Nuoxian Huang, Boyang Li, Luyao Niu, Zipei Fan, Tianle Lun, Yicheng Tao, Junyou Su, Zhaoya Gong, et al. 2024b. Plangpt: Enhancing urban planning with tailored language model and efficient retrieval. *arXiv preprint arXiv:2402.19273*.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, et al. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

A Domain-Specific Data Preprocessing Details

Our data preprocessing pipeline involves several sophisticated steps to ensure high-quality planning maps for annotation and model training:

Map Collection and Extraction We collected approximately 5,000 master plans and detailed planning maps from urban planning bureaus across China. These documents were primarily in PDF format, requiring specialized extraction techniques. We employed PDF parsers with custom configurations to extract high-resolution visual content while preserving spatial relationships and annotations critical to planning interpretation.

Quality Filtering Initial filtering employed a multi-stage approach:

- Resolution-based filtering: We established minimum resolution thresholds (1000×1000 pixels) to ensure sufficient detail for fine-grained planning elements.
- Information density assessment: We used computer vision techniques to quantify the information content of each map, filtering out overly sparse or dense representations.
- LLM-as-judge evaluation: We designed specialized prompts (detailed in Appendix B) that enabled large language models to assess information density, clarity, and planning relevance of extracted maps.

This rigorous preprocessing approach ensured our seed dataset represented authentic professional planning expertise while maintaining high visual and informational standards.

B Filter Image Prompts

Figure shows the full prompt used to guide LLMs in filtering planning maps. It specifies criteria such as completeness, visual clarity, and the presence of essential map elements, enabling automated and consistent quality assessment.

C Experimental Setup

Implementation Details We conducted experiments using 4 NVIDIA A100 GPUs (80GB each). Our implementation is based on the VERL framework (Sheng et al., 2024) for efficient VLM fine-tuning. We employed Qwen2-7B-VL-Instruct

Image Filter Prompt

PLANNING_MAP_PROMPT = """You are an urban planning expert. Please determine if the image below is a complete and independent urban or territorial spatial planning map.

Please first provide a brief description of the image, then determine if it is a planning map.

Judgment criteria:
1. It must be a complete planning map, not a part or screenshot of a planning map
2. The planning map should be the main content of the image, occupying the main area of the image
3. It should not contain many page elements unrelated to the planning map (such as large text descriptions, tables, etc.)
4. It should have typical features of planning maps:
- Clear visual structure of spatial layout or land zoning
- Essential planning map elements such as legend, scale, direction
indicator, map title, planning unit, etc.

Please note:
- If the image is a scan of an entire document page, and the planning map is only part of the page, it should be determined as not meeting the requirements
- If the image contains multiple planning maps, it should also be determined as not meeting the requirements
- If the image content is blurry and difficult to recognize the planning content, it should be determined as not meeting the requirements

After analyzing the image, please output in the following format:

Analysis: [Provide your analysis here]
Determination: If it is a complete and independent planning map, please output: $\boxed{1}$
If it is not a complete and independent planning map, please output: $\boxed{0}$
"""

(Wang et al., 2024b) as our base model and conducted supervised fine-tuning without a pre-training phase (Karamcheti et al., 2024). For SFT, we used the AdamW optimizer with a learning rate of 2e-5, cosine learning rate scheduler with 5% warmup steps, and trained for 3 epochs with a global batch size of 128 and maximum sequence length of 8192 tokens.

Datasets Our training corpus consists of approximately 10k instruction-following examples generated from 1k selected urban planning maps using our PlanAnno-V framework. The dataset spans multiple query categories including zoning analysis, infrastructure assessment, spatial reasoning, policy compliance, and regulatory alignment. The maps primarily originate from diverse Chinese cities, ensuring geographical coverage. Detailed data analysis and validation procedures are documented in Appendix D.

Evaluation Benchmarks We evaluate our model using both domain-specific and general benchmarks: (1) **PlanBench-V**, our newly created benchmark described in Section 3.4; and (2) **General VLM Benchmarks** including MMMU (Yue et al., 2024), GQA (Hudson and Manning, 2019), and POPE (Li et al., 2023c) to assess preservation of general visual understanding capabilities. We use the lmms-eval framework (Zhang et al., 2024a) for standardized evaluation.

D Data Analysis

We conduct a comprehensive analysis of the dialogue dataset from two perspectives:

First, we analyze the token distribution across each round of dialogue, including the number of tokens in the instruction, response, and their total. This provides insights into the input-output complexity of the dataset. Additionally, we examine the distribution of critical points per round to understand the density of semantic shifts or decision points in the dialogues. The results are shown in Figure 4.

Second, to assess the semantic diversity and complexity of the instructions, we employ the Instagger model to map each instruction into a predefined tag space. This allows us to analyze the diversity of task types and compute the number of tags associated with each instruction to estimate its semantic complexity. The corresponding analysis is illustrated in Figure 5.

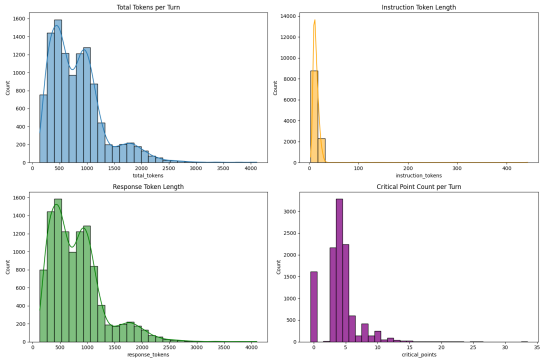


Figure 4: Token Distribution and Cirtical Point Distribution Analysis

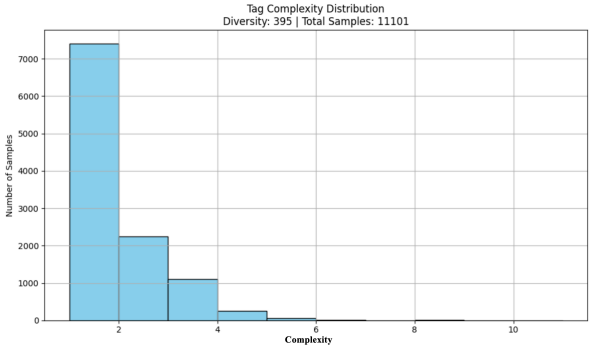


Figure 5: Instagger Analysis

E Instruction Synthesis Effectiveness Analysis

To validate the quality and diversity of our automated instruction generation pipeline, we conducted a comprehensive comparative analysis between human-annotated seed data and PlanAnno-V synthesized instructions across three critical dimensions: distributional alignment, categorical diversity, and complexity preservation.

Distributional Alignment We analyze whether synthesized instructions maintain the underlying distributional characteristics of expert-curated examples. Using the bge-zh-base model (Xiao et al., 2024) to map instructions to embeddings, we compute distributional metrics between seed and synthesized data to ensure quality preservation.

We employ three complementary metrics: (1) *Cosine Similarity* measures semantic coherence between centroids of seed $\mathcal{S}$ and synthesized $\mathcal{T}$ embeddings: $\text{CosSim}(\mathcal{S}, \mathcal{T}) = \frac{\bar{\mathbf{s}} \cdot \bar{\mathbf{t}}}{|\bar{\mathbf{s}}||\bar{\mathbf{t}}|}$ where $\bar{\mathbf{s}}$ and $\bar{\mathbf{t}}$ are mean embeddings. Higher values (closer to 1.0) indicate better semantic alignment. (2) *Maximum Mean Discrepancy (MMD)* quantifies distributional divergence in reproducing kernel Hilbert space: $\text{MMD}^2(\mathcal{S}, \mathcal{T}) =$

$\mathbb{E}[k(\mathbf{s}, \mathbf{s}')] + \mathbb{E}[k(\mathbf{t}, \mathbf{t}')] - 2\mathbb{E}[k(\mathbf{s}, \mathbf{t})]$ using RBF kernel $k(\mathbf{x}, \mathbf{y}) = \exp(-\gamma|\mathbf{x} - \mathbf{y}|^2)$ to detect subtle distributional shifts, where lower values indicate better distributional similarity. (3) *Wasserstein Distance* measures optimal transport cost for spatial arrangement analysis: $W_2(\mathcal{S}, \mathcal{T}) = \inf_{\pi \in \Pi(\mathcal{S}, \mathcal{T})} \left(\int |\mathbf{s} - \mathbf{t}|^2 d\pi(\mathbf{s}, \mathbf{t}) \right)^{1/2}$ where lower values indicate closer spatial distributions in the embedding space.

Our analysis reveals strong semantic alignment (Cosine Similarity = 0.9350) with controlled variation (MMD = 0.0515), despite some spatial distribution differences (Wasserstein Distance = 2.0255). As visualized in Figure 3(a) using UMAP dimensionality reduction and Gaussian kernel density estimation, our method maintains the core distribution while introducing beneficial diversity, demonstrating strong distributional fidelity with controlled variations.

Categorical Expansion We employ intent tagging to analyze task type distributions as introduced in section 3.2. While seed data covers 8 primary planning categories, our synthesis expands coverage to 15 categories. Notably, this expansion maintains proportional representation of core planning tasks, preventing categorical drift.

Quality Preservation Professional urban planners evaluated 100 randomly sampled instruction-response pairs from both seed and synthesized data across three dimensions (0-1 scale): planning expertise, factual correctness, and fluency. Synthesized instructions maintained comparable quality to expert-created seed data (planning expertise: 0.87 vs. 0.89; correctness: 0.85 vs. 0.88; fluency: 0.91 vs. 0.90), with no statistically significant differences ($p>0.05$). This confirms our approach preserves expert annotation quality while achieving substantial scale improvements.

F Data Leakage Analysis

To ensure evaluation validity, we conducted data leakage analysis between our training corpus and evaluation datasets, addressing concerns about contamination in large-scale model development (Du et al., 2023). We focused on detecting potential image duplicates by extracting CLIP-ViT-L/32 (OpenAI, 2021) embeddings from all images and computing cosine similarities between pairs. Images exceeding a 0.9 similarity threshold underwent manual inspection. As shown in Table 5, we identified minimal high-similarity pairs, and human verification confirmed these were distinct content rather than duplicates. This analysis confirms our performance improvements represent genuine domain-specific capabilities rather than memorization artifacts.

Dataset	Total Images	High Similarity	Verified Leakage
Seed Data	50	0 (0.00%)	0 (0.00%)
PlanAnno-V	1k	9 (0.09%)	0 (0.00%)

Table 5: Data leakage analysis showing minimal image duplication.

G PlanBench-V

G.1 Benchmark Overview

PlanBench-V is the first comprehensive benchmark for evaluating VLM performance on urban planning map interpretation tasks. The benchmark comprises 101 spatial planning maps sourced from official master plans and the Chinese Certified Urban-Rural Planner Qualification Examination (CURPQE), with 1,134 expert-annotated question-answer pairs. Professional urban planners with domain expertise manually curated all annotations to ensure accuracy and relevance to real-world planning practice. We sample 300 representative examples from this dataset as the baseline evaluation set for this paper.

G.2 Evaluation Framework

Our evaluation framework assesses VLM capabilities across four main categories, each containing specific detailed subcategories as shown in Table 1:

Perception (Perc) evaluates fundamental visual understanding capabilities:

- *Element Recognition*: Identifying layout configurations, textual annotations, geographic features, and drawing elements
- *Description*: Generating comprehensive captions that capture planning-relevant visual details

Reasoning (Reas) assesses structured inference and domain-specific analysis:

- *Classification*: Recognizing different types of planning maps (master plans, detailed plans, specialized plans)
- *Spatial Relations*: Understanding topological, ordinal, and metric spatial relationships between geographic elements

- *Professional Reasoning*: Applying planning principles and domain knowledge to interpret map content

Association (Assoc) evaluates policy and regulatory knowledge integration:

- *Policy Association*: Connecting visual elements with relevant planning policies, regulations, and guidelines

Implementation (Impl) measures practical application and critical evaluation:

- *Evaluation*: Assessing strengths and weaknesses of planning proposals
- *Decision Making*: Making value-driven choices and recommendations in planning scenarios

G.3 Example Questions

Element Recognition: "Where is the north arrow located on this map? What transportation features are shown?"
Spatial Relations: "What is the topological relationship between the Riverside Agricultural Zone and the Lixiahe Agricultural Zone?"
Professional Reasoning: "Based on the spatial layout shown, analyze the urban development strategy and its implications for regional connectivity."
Policy Association: "Which relevant policies or guidelines should be considered when evaluating the implementation of this planning proposal?"
Evaluation: "What are the strengths and weaknesses of this planning scheme in terms of land use efficiency and ecological protection?"

G.4 Scoring Methodology

Each question is evaluated using expert-defined criteria with normalized scores ranging from 0 to 2. We employ an automated evaluation protocol using GPT-4o-mini as judge, which achieves high correlation (average pairwise error < 0.1) with human expert ratings. All scores in Table 1 are normalized with GPT-4o as the reference baseline (1.0). The specific prompt is shown in Figure 6.

Prompt of Evaluation

""" Please evaluate the following answer based on the question, the list of scoring points, and the image content.

Question: {question}

Scoring Points List:
{critical_points_text}

Answer to be Evaluated: {summary}

Scoring Criteria:
Please check whether the model's answer covers each scoring point:
- For each scoring point, if the model's answer includes relevant content, give 1 point.
- If the scoring point is not covered or is incorrectly described, give 0 points.
- The scoring points are mutually exclusive, and each point can earn a maximum of 1 point.

Please score according to the following format:
1. Scoring Point 1: [0/1] – Briefly explain whether this point is covered and provide reasoning
2. Scoring Point 2: [0/1] – Briefly explain whether this point is covered and provide reasoning
...

Final Score: X/Y (X is the total score, Y is the total number of scoring points)
"""

Figure 6: The specific prompt for Scoring Methodology of PlanBench-V.

H Compare Attention Score Map

I Comparison of PlanGPT-VL and Qwen

Evaluation Example 1 (Part 1)

Question:
How does Jiangsu Province promote urban-rural integration in territorial spatial planning?
Answer:
In its territorial spatial planning, Jiangsu Province promotes urban-rural integration by scientifically delineating the urban development boundary. The plan explicitly states that the expansion multiplier of the urban development boundary is controlled within 1.3 times the scale of urban construction land in 2020, ensuring orderly urban development. At the same time, by strengthening infrastructure construction, the province promotes integrated urban-rural development and narrows the urban-rural gap. Additionally, Jiangsu optimizes the functional layout of towns to improve their overall carrying capacity, providing a higher quality living environment for urban and rural residents.
Critical Points:

1 Scientifically delineate urban development boundaries, controlling the expansion multiplier within 1.3 times.

2 Ensure orderly urban development.

3 Strengthen infrastructure construction to promote urban-rural integration.

4 Narrow the urban-rural gap.

5 Optimize town functional layout and improve overall carrying capacity.

6 Provide higher-quality living environments for urban and rural residents.

Model Response:
Jiangsu Province promotes urban-rural integration in territorial spatial planning mainly through the following measures:
1. **Comprehensive Planning**: Develop integrated urban-rural spatial development plans to ensure coordinated development in functions, infrastructure, and public services between cities and rural areas.
2. **Infrastructure Construction**: Increase investment in rural infrastructure, improving transportation, water supply, electricity, and internet access to enhance rural living standards.
3. **Land Use Policy**: Optimize land use structure through scientific zoning and intensive utilization to rationally allocate agricultural and construction land.
4. **Industrial Transformation and Upgrading**: Encourage industrial coordination between urban and rural areas, support the development of characteristic rural industries, and boost rural economic vitality.
5. **Ecological Protection**: Strengthen environmental protection and promote green development to ensure harmony between development and nature.
6. **Policy Support**: Implement policies favorable to urban-rural integration, including financial support and tax incentives to stimulate investment and innovation.
7. **Social Services**: Improve rural education, healthcare, and cultural services to raise living standards and attract talent back to rural areas.
In summary, Jiangsu's approach to urban-rural integration aims to achieve comprehensive and harmonious development through coordinated planning and multifaceted policy support.

Evaluation Example 1 (Part 2)

Score Text:
1. Point 1: [1] – The model mentions "comprehensive planning" and "integrated urban-rural spatial development plans," which align with the requirement to scientifically delineate urban development boundaries.
2. Point 2: [1] – The response discusses coordinated development in functions, infrastructure, and public services, which corresponds to ensuring orderly urban development.
3. Point 3: [1] – The response clearly mentions increasing investment in rural infrastructure, which matches the point of strengthening infrastructure construction.
4. Point 4: [1] – Improving education, healthcare, and cultural services in rural areas helps narrow the urban-rural gap.
5. Point 5: [1] – Scientific zoning and rational land allocation reflect optimizing town functions and improving carrying capacity.
6. Point 6: [0] – Although the response mentions raising living standards, it does not explicitly emphasize providing a higher-quality living environment, and thus does not fully meet this point.

Final Score: 5/6

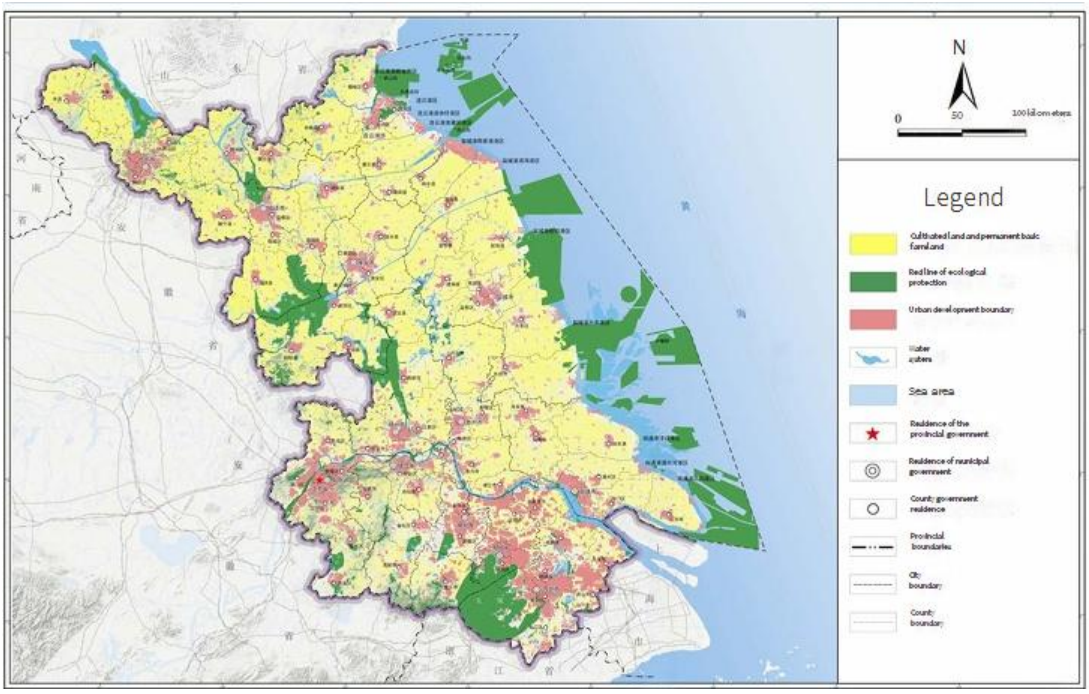


Figure 7: Image of Evaluation Example 1

Evaluation Example 2

Question:
Which category does this map belong to: master plan, regulatory detailed plan, construction detailed plan, or special plan?
Answer:
Construction detailed plan.
Critical Points:

1 Correctly identify the map type as a "construction detailed plan."

2 Recognize the features and purpose of a construction detailed plan.

Model Response:
This map belongs to a regulatory detailed plan. Regulatory detailed plans are typically used to show specific details such as land use, road layout, and urban infrastructure.
Score Text:
1. Point 1: [0] – The model incorrectly identifies the map as a "regulatory detailed plan" instead of the correct type "construction detailed plan," thus failing to meet this point.
2. Point 2: [0] – The model fails to recognize the features and intended use of a construction detailed plan, so this point is not fulfilled.

Final Score: 0/2

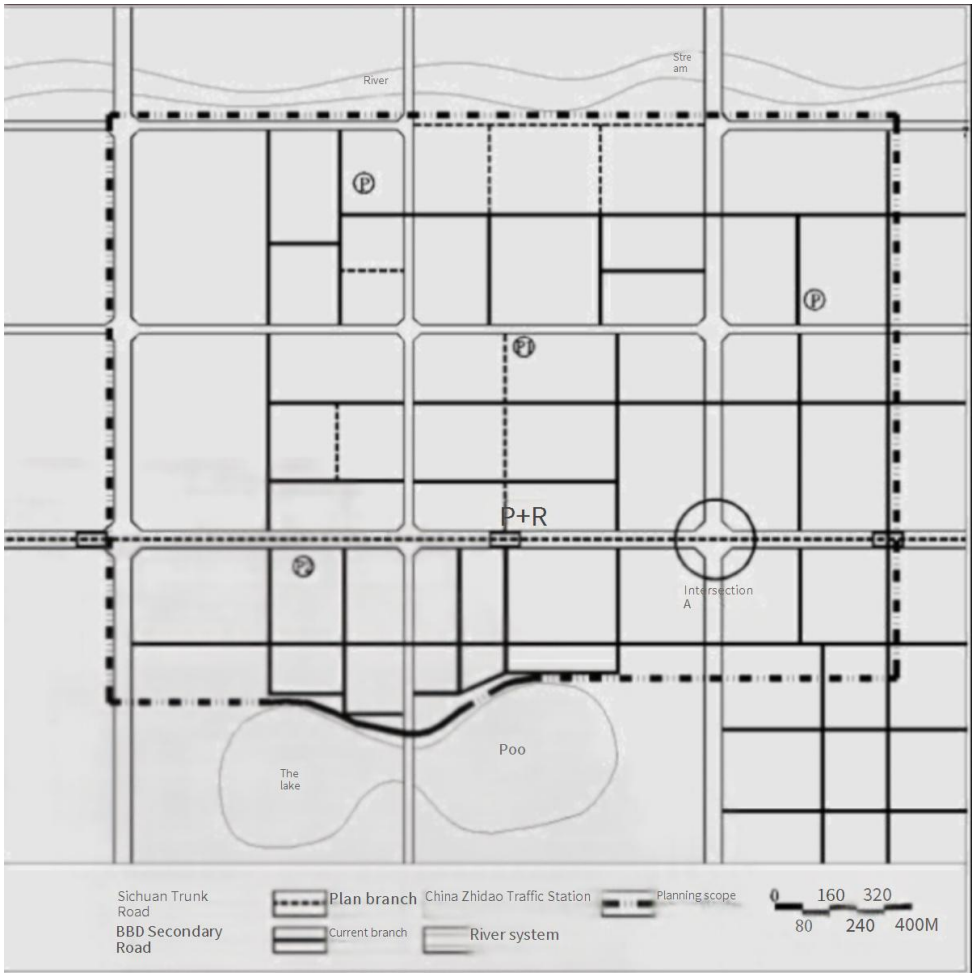
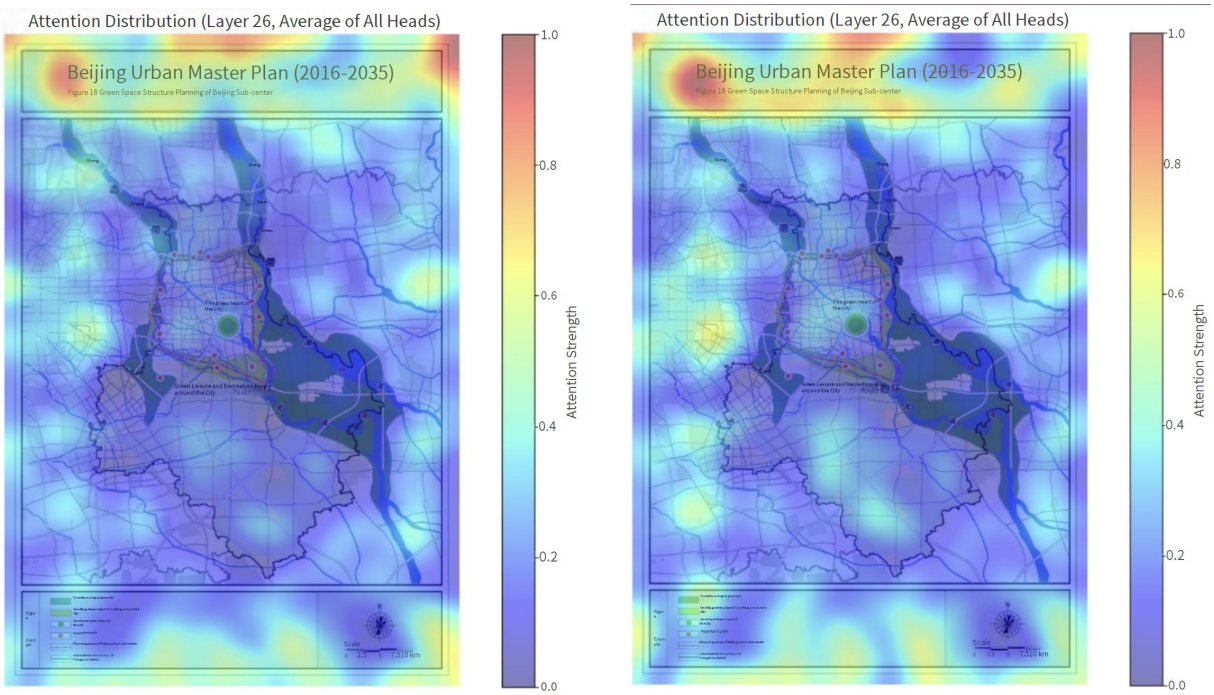
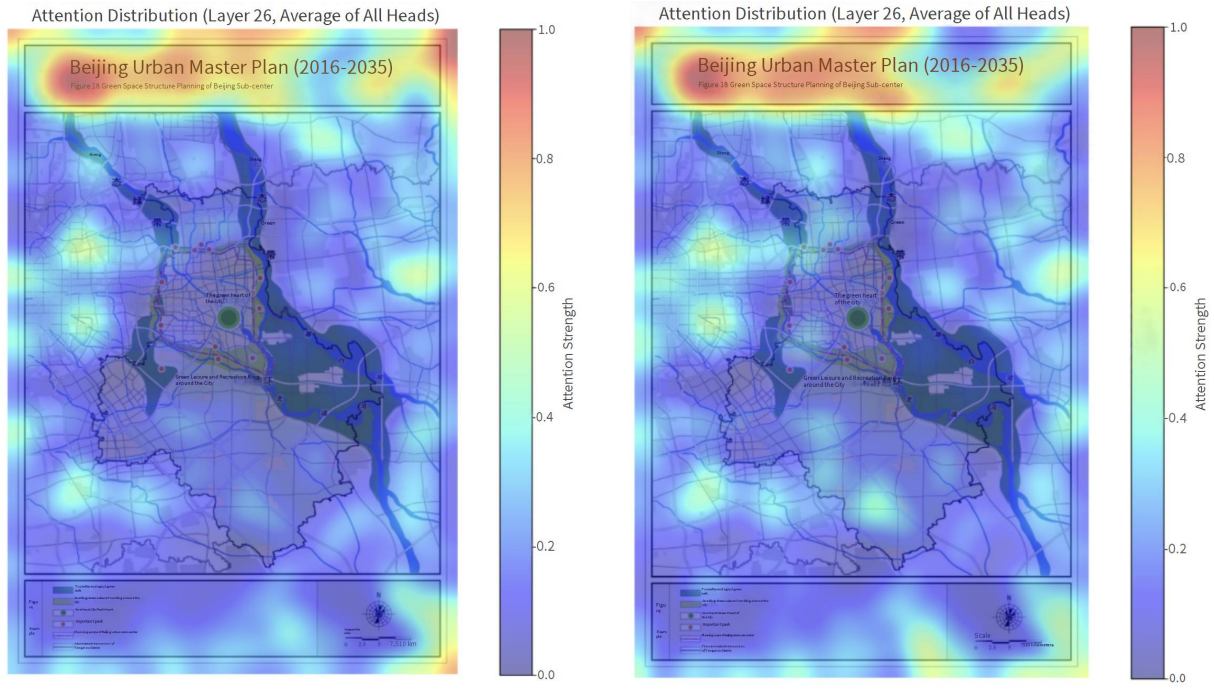


Figure 8: Image of Evaluation Example 2



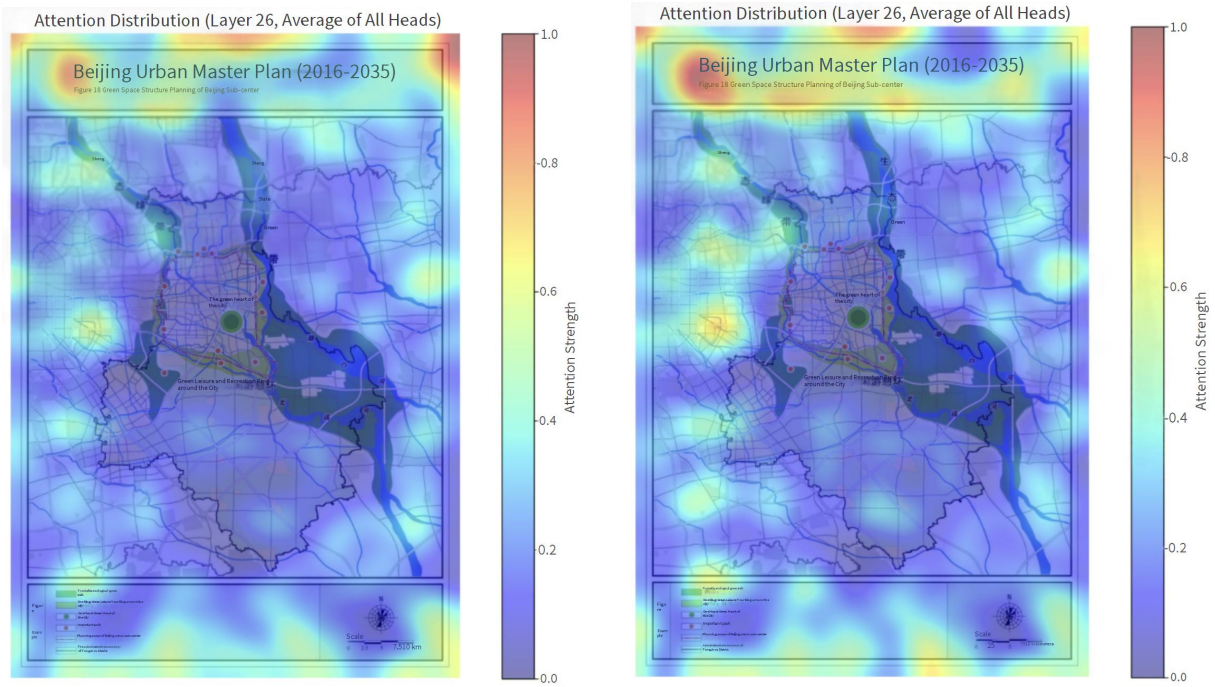
(a) Base Model (b) Mix Model

Figure 9: Attention scores for question: Where is the green heart of the city?



(a) Base Model (b) Mix Model

Figure 10: Attention scores for question: Please describe this image



(a) Base Model (b) Mix Model

Figure 11: Attention scores for question: Where is the ecological green belt

Example 1

Question: Please describe the national key wetlands in the picture.(image see Fig 12)

PlanGPT-VL: According to the legend, the National Key Wetland is represented by a light gray shadow. It is located on the southern side of the planned expressway, near the river (within the river embankment) area. The wetland is situated close to the county town and the food industry park, but maintains a certain distance from them, thereby avoiding direct disturbance. The primary functions of the National Key Wetland include protecting biodiversity, regulating the climate, conserving water resources, and purifying water quality. At the same time, it serves as an important ecological buffer zone for the city, preventing excessive interference with natural ecosystems caused by urban expansion. In urban planning, the establishment of the National Key Wetland reflects the coordination between ecological protection and urban development, contributing to the harmonious coexistence between the city and nature.
Qwen2-7B-Instruct: According to the information in the picture, the national key wetlands are located in the lower right corner of the map and are represented by light gray areas.
Qwen2-72B-Instruct-AWQ: In this map, the national key wetland is marked as a small area near the river. This area is labeled "National Key Wetland" and outlined with a dashed frame.

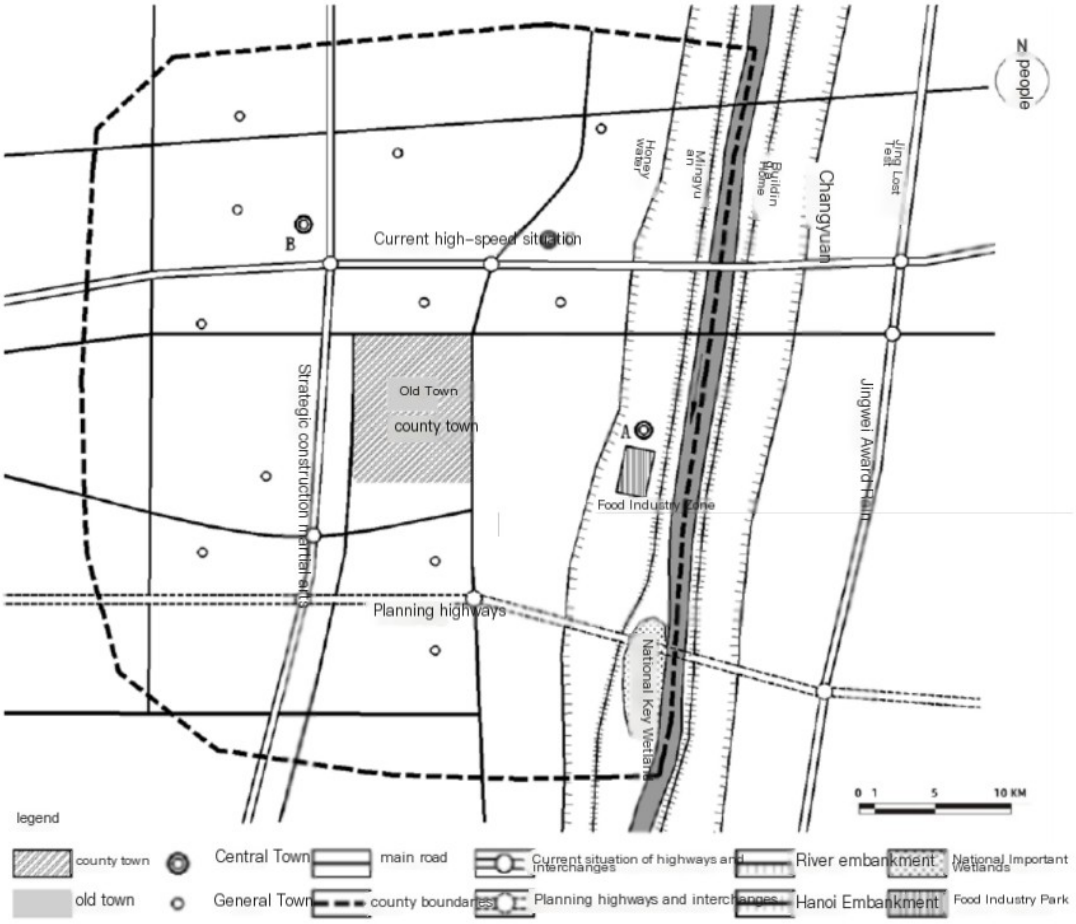


Figure 12: Image of Example 1

Example 2

Question: 世界遗产在地图上是如何标识的？北京市有哪些世界遗产被列入了这张地图？(image see Fig 13)

PlanGPT-VL: 世界遗产在地图上以红色圆点表示，用于标识其位置。北京市的世界遗产包括：故宫、天坛、颐和园、周口店北京人遗址、长城、大运河、西山永定河文化带（如妙峰山、潭柘寺等）、大运河文化带（如通州大运河、元大都遗址等）。这些世界遗产被纳入北京市市域历史文化名城保护结构规划中，体现了北京市对世界遗产的高度重视和保护。

Qwen2-7B-Instruct: 世界遗产在地图上用紫色圆圈标识。北京市被列入这张地图的世界遗产有：故宫、天坛、颐和园、周口店北京人遗址、长城、大运河。

Qwen2-72B-Instruct-AWQ: 世界遗产在地图上是用红色圆圈标识的。北京市被列入这张地图的世界遗产包括：长城、故宫、天坛、颐和园、十三陵、周口店北京人遗址、明十三陵、清东陵、清西陵、圆明园、北海公园、景山公园、中山公园、天安门广场、北京动物园、北京植物园、北京天文馆、北京自然博物馆、北京科技馆、北京艺术博物馆、北京历史博物馆、北京军事博物馆、北京革命博物馆、北京民俗博物馆、北京工艺美术博物馆、北京古建筑博物馆、北京古生物博物馆、北京古陶瓷博物馆、北京古钱币博物馆、北京古玉器博物馆、北京古家具博物馆、北京古钟表博物馆、北京古乐器博物馆、北京古书画博物馆、北京古籍博物馆、北京古籍修复博物馆、北京古籍装裱博物馆、北京古籍装订博物馆、北京古籍装订工具博物馆、北京古籍装订材料博物馆、北京古籍装订技术博物馆、北京古籍装订历史博物馆、北京古籍装订艺术博物馆、北京古籍装订工艺博物馆...

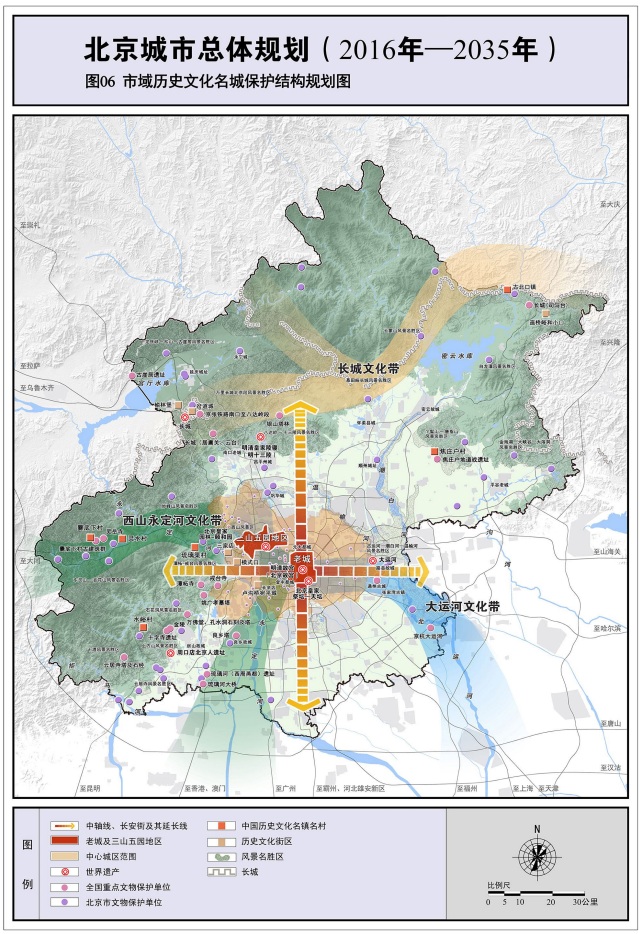


Figure 13: Image of Example 2

Example 3

Question How is the transportation connection between the "Beijing New Airport" and the surrounding areas in the image? (image see Fig 14)

PlanGPT-VL: Translation (without changing the format):

—

According to the information in the image, the transportation connection between Beijing New Airport and the surrounding areas is as follows:

1. **Location of Beijing New Airport**: Beijing New Airport is located in the lower right corner of the image, near Daxing District. It serves as an important transportation hub in the image.

2. **Surrounding Transportation Lines**: * **Operational Lines**: The image shows that there are operational lines near Beijing New Airport, including Line 8, Line S1, etc. These lines directly connect Beijing New Airport with the central urban area and surrounding regions. * **Planned Lines**: The image also indicates planned lines (including those under construction), such as the southern and northern extensions of Line S1. These lines will further improve the transportation network of Beijing New Airport.

3. **Transfer Stations and Regular Stations**: * **Transfer Stations**: There are several transfer stations around Beijing New Airport, such as Daxing Airport Station and Daxing Airport North Station. These stations enable efficient transfers between rail transit and surface transportation. * **Regular Stations**: There are also regular stations around Beijing New Airport, such as Daxing Airport Station and Daxing Airport North Station, providing passengers with convenient transfer and travel services.

4. **Transportation Accessibility**: * **Distance**: According to the scale in the image, Beijing New Airport is relatively close to the central urban area and other regions, indicating good transportation accessibility. * **Transfer Convenience**: The distribution of transfer and regular stations around Beijing New Airport is reasonable, enabling seamless integration of rail and surface transportation and enhancing passengers' transfer experience.

5. **Conclusion**: * The transportation connection between Beijing New Airport and the surrounding areas is relatively strong. The combination of operational and planned lines enables efficient linkage with the central urban area and surrounding regions. * The layout of transfer and regular stations is well-designed to meet passengers' transfer needs and improve overall transportation accessibility.

In summary, the transportation connection between Beijing New Airport and the surrounding areas is well-developed. With the coverage of the rail transit network and the layout of transfer stations, it enables convenient transportation accessibility, supporting the efficient operation of Beijing New Airport and coordinated regional development.

Example 3

Question How is the transportation connection between the "Beijing New Airport" and the surrounding areas in the image? (image see Fig 14)

Qwen2-7B-Instruct: The transportation connection between "Beijing New Airport" and the surrounding areas in the image is mainly achieved through rail transit. As shown in the image, there are multiple rail transit lines around Beijing New Airport, including Line 8, Line 19, and the New Airport Line. These lines have stations near the new airport, providing convenient access for passengers. In addition, the New Airport Line connects to the central area of Beijing, further strengthening the transportation link between the new airport and the urban center.

Qwen2-72B-Instruct-AWQ: The image shows that the transportation connection between "Beijing New Airport" and the surrounding areas is very close. Multiple rail transit lines link the new airport with the city center of Beijing and surrounding regions. Specifically, there are several blue lines near the new airport, which represent planned lines, including those under construction. These lines are connected to other routes through transfer stations, forming a convenient transportation network. Through these lines, passengers can quickly reach the city center of Beijing and various surrounding areas from the new airport.

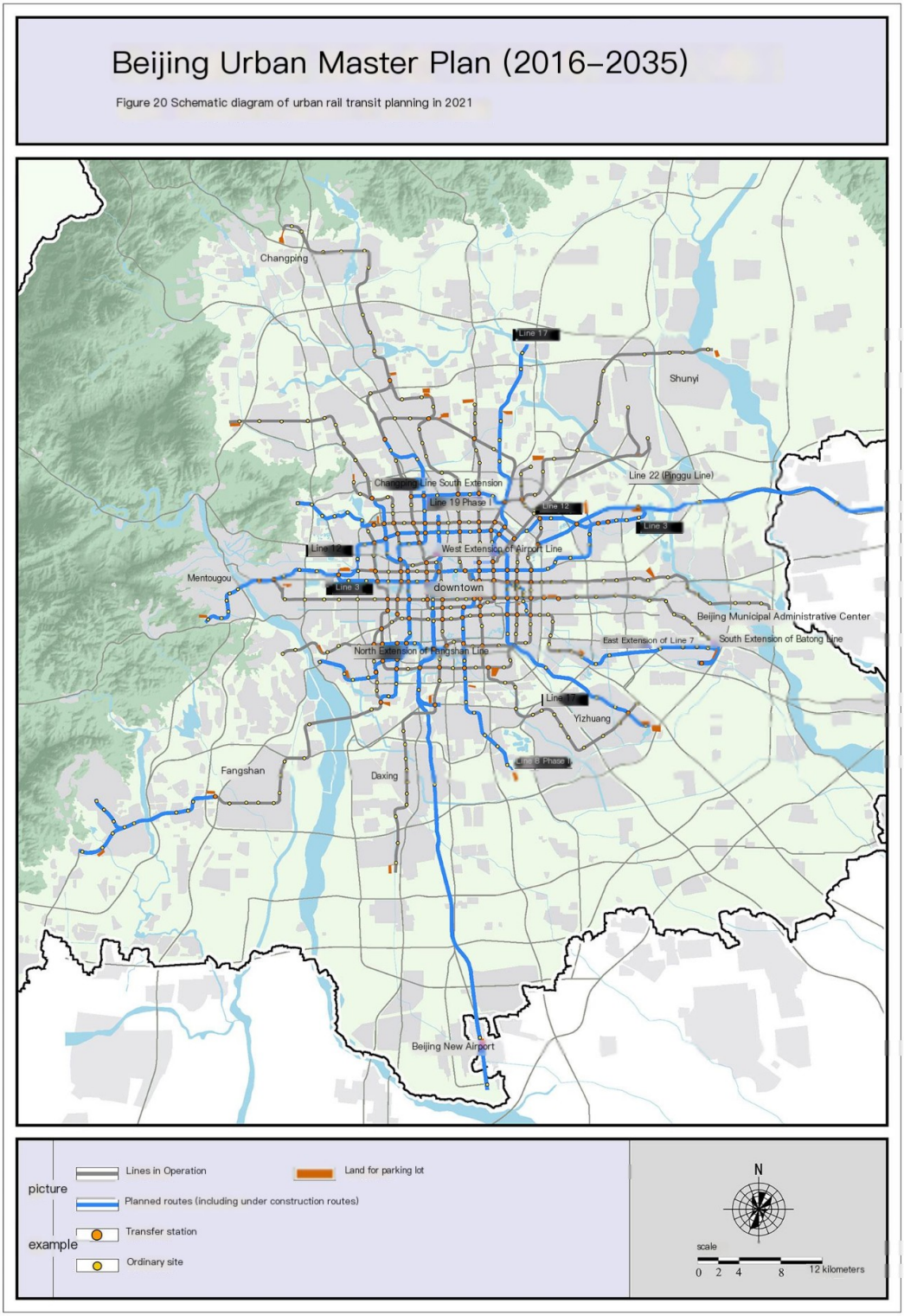


Figure 14: Image of Example 3