

How Well Do Vision–Language Models Understand Cities?

A Comparative Study on Spatial Reasoning from Street-View Images

Juneyoung Ro Namwoo Kim Yoonjin Yoon*
 Korea Advanced Institute of Science and Technology
 {juneyoung, namwoo, yoonjin}@spacetime.kaist.ac.kr

Abstract

Effectively understanding urban scenes requires fine-grained spatial reasoning about objects, layouts, and depth cues. However, how well current vision-language models (VLMs), pretrained on general scenes, transfer these abilities to urban domain remains underexplored. To address this gap, we conduct a comparative study of three off-the-shelf VLMs—BLIP-2, InstructBLIP, and LLaVA-1.5—evaluating both zero-shot performance and the effects of fine-tuning with a synthetic VQA dataset specific to urban scenes. We construct such dataset from segmentation, depth, and object detection predictions of street-view images, pairing each question with LLM-generated Chain-of-Thought (CoT) answers for step-by-step reasoning supervision. Results show that while VLMs perform reasonably well in zero-shot settings, fine-tuning with our synthetic CoT-supervised dataset substantially boosts performance, especially for challenging question types such as negation and counterfactuals. This study introduces urban spatial reasoning as a new challenge for VLMs and demonstrates synthetic dataset construction as a practical path for adapting general-purpose models to specialized domains.

1. Introduction

Understanding street-level urban scenes at fine-grained spatial scales is crucial for informing how cities are designed and experienced. Visual elements such as greenery, skyline openness, and building density significantly influence urban comfort, walkability, and perceived safety [3, 12, 13, 34]. While humans can intuitively grasp these features when viewing street-level imagery, current AI models—particularly those designed for general-purpose visual understanding—often face difficulties in reasoning about spatial relationships and compositional patterns in complex urban environments.

Out-of-Domain Challenge in Urban Scenes. Large-scale models like CLIP [38] excel on images resembling those in

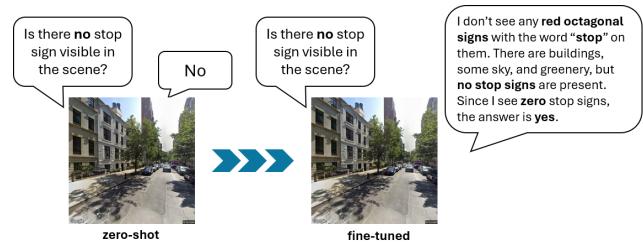


Figure 1. Example illustrating a zero-shot model’s failure in handling spatial reasoning under negation, contrasted with a fine-tuned model’s correct, step-by-step spatial reasoning process.

their pretraining corpora, yet frequently underperform on domains such as medical scans, satellite imagery, and “natural adversarial” photos that differ in texture, structure, or composition [18, 27]. A recent survey [27] attributes this limitation to the model’s tendency to rely on surface-level keyword associations rather than reasoning about spatial structures and relationships. Urban street scenes are particularly affected by these issues. While they share recurring elements, such as trees and buildings, fine-grained cues such as how much tall buildings block the sky, the canopy area of trees that creates shaded rest spots, or the overall greenery index that signals biophilic quality all convey crucial spatial information. Accurately interpreting these details is vital for evidence-based planning [19], navigation [17], and public-space design [8].

Street-view imagery serves as a primary tool for capturing and analyzing such fine-grained urban features. It has been extensively employed to study human-inhabited urban environments [26, 33]. Recent research has focused on streetscape perception evaluation [29, 37], and the integration of reasoning module may further enhance the depth of visual urban analysis. However, existing studies often emphasize perceptual assessment without systematically incorporating reasoning about spatial relationships. To address this gap, our pipeline generates both **perceptual** and **compositional** base QA pairs to assess how well state-of-the-art vision–language models interpret urban scenes. We then transform these base

*Corresponding author: yoonjin@spacetime.kaist.ac.kr

pairs into Chain-of-Thought (CoT) variants that verbalize the underlying reasoning process, supporting evaluation of final answers alongside reasoning fidelity. This design allows even simple yes/no questions to probe fine-grained spatial understanding and the ability to articulate metadata-grounded inferences.

In pursuit of this goal, we introduce a synthetic QA generation pipeline built to fine-tune vision–language models (VLMs) for urban street-scene understanding. Leveraging segmentation, depth estimation, and object detection predictions, we construct thousands of visually grounded questions and answers, providing targeted supervision for spatial and compositional understanding. We use this synthetic dataset to evaluate three off-the-shelf VLMs—BLIP-2 [25], InstructBLIP [11], and LLaVA-1.5 [30]—covering the spectrum from zero-shot transfer models to fully conversational multimodal assistants. These models were selected for their architectural diversity, distinct training strategies, and full open-source availability, which enables transparent inspection, reproducibility, and community-driven improvement. BLIP-2 serves as a strong general-purpose model with frozen vision encoders, InstructBLIP as an instruction-tuned variant optimized for multimodal following, and LLaVA-1.5 as a large-scale, conversational VLM with end-to-end fine-tuning. By comparing each model in its original form and after fine-tuning, we assess whether modest adaptation can yield reliable, human-aligned street-level reasoning. Below are our key research questions and contributions.

Research Questions

- **RQ1** – How well do current vision–language models understand fine-grained spatial relationships in urban street scenes out of the box?
- **RQ2** – How much can targeted fine-tuning on a synthetic but carefully structured domain-specific QA set close this gap?
- **RQ3** – How do model strengths and weaknesses vary across different question types, from perception-based to compositional reasoning tasks?

Our Contributions

- We conduct a comparative study of BLIP-2, InstructBLIP, and LLaVA-1.5 on fine-grained spatial reasoning in urban street scenes, enabled by a synthetic VQA dataset we construct from street-view images to support both zero-shot and fine-tuned evaluation.
- We develop a modular QA generation pipeline that produces 280K questions across perceptual, compositional, and CoT formats, enabling diverse and progressively challenging supervision.
- We perform detailed quantitative and qualitative evaluations by question type, revealing model-specific strengths, weaknesses, and the effects and tradeoffs of fine-tuning at different scales. The code for

implementing our pipeline are publicly available at: https://github.com/eeyore22/urban_scope.

2. Related Work

Enhancing reasoning in vision–language models. A growing body of work has focused on improving the reasoning capabilities of vision–language models. Gokhale et al. (2020) introduced the VQA-LOL benchmark to specifically evaluate logical reasoning in VQA, demonstrating that models frequently fail on tasks involving negation and complex logic [15]. To address similar challenges, Niu et al. (2021) proposed a counterfactual VQA framework that mitigates language biases, showing that many VQA models rely on spurious correlations rather than true scene understanding [36]. Zhang et al. (2025) further emphasized that even modern multimodal models still misinterpret negated or hypothetical questions in urban driving scenarios, underscoring persistent reasoning gaps [51].

Synthetic and programmatically generated benchmarks have further advanced reasoning evaluation. Johnson et al. (2017) introduced the CLEVR benchmark, which used synthetic 3D scenes to systematically test compositional and logical reasoning in VQA [21]. Hudson and Manning (2019) developed GQA, a large-scale, programmatically generated VQA dataset that specifically targets multi-step reasoning on real images [20]. Chen et al. (2024) introduced SpatialVLM, which augmented VQA training with depth-aware, spatially grounded questions to improve quantitative spatial reasoning [7]. Wang et al. (2025) developed OmniDrive, which synthesized counterfactual driving scenarios (e.g., “If I decide to accelerate and make a left turn, what could be the consequences?”) to generate decision-oriented QA pairs tailored for driving applications [46].

Urban scene understanding and computer vision foundations. Meanwhile, the computer vision community has developed a wide range of tools and foundational datasets for urban scene understanding. Cityscapes [10] and Mapillary Vistas [35] remain among the most widely used resources in this domain. Cityscapes provides finely annotated urban street scenes and significantly advanced semantic segmentation in complex driving environments. Mapillary Vistas extends this foundation by offering diverse street-level imagery across a broader range of cities, countries, and viewpoints, enhancing generalization beyond uniform city structures.

Ranftl et al. (2020) proposed MiDaS, a cross-dataset monocular depth estimation framework that improved depth generalization across various visual domains, including urban scenes [39]. Caesar et al. (2020) introduced nuScenes, which provided multimodal sensory inputs—such as lidar, radar, and multi-camera setups—for urban driving tasks [5], enabling detailed spatial and temporal scene understanding. These perception pipelines now offer high-quality scene

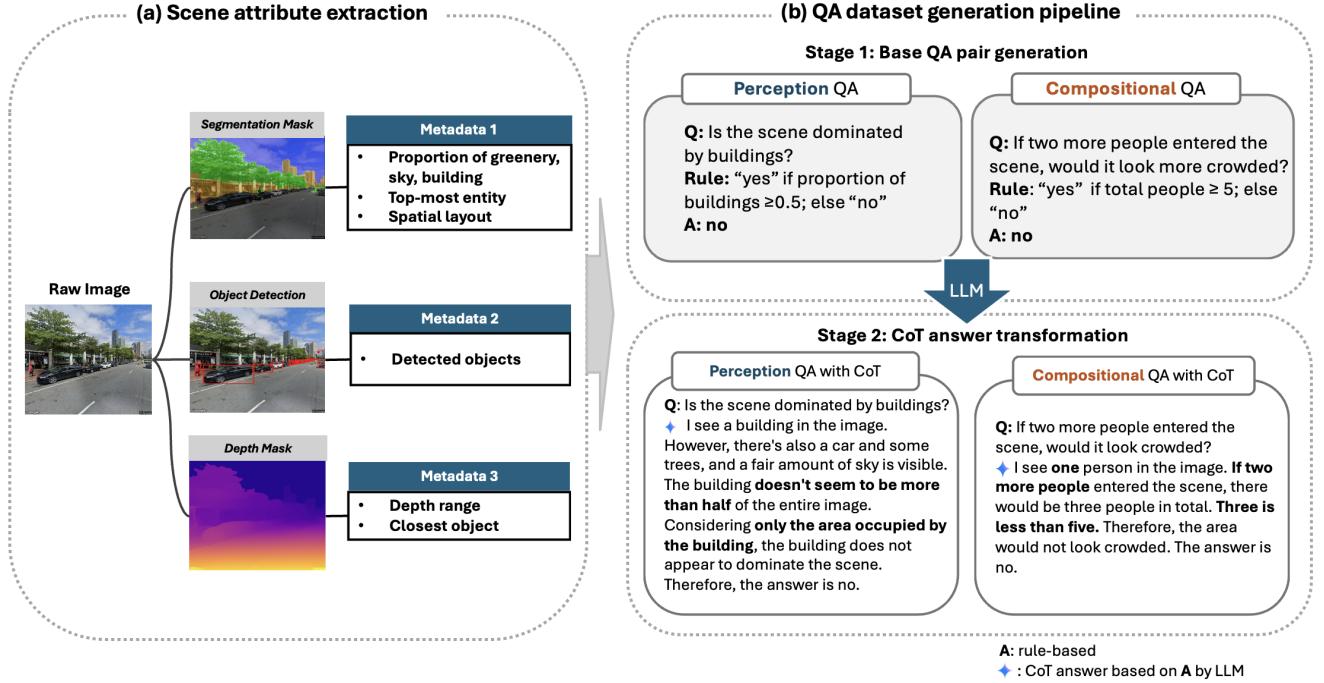


Figure 2. Overview of our pipeline. **(a)** We take raw street-view images as inputs and extract three key predictions: segmentation masks, object detection results, and monocular depth masks. Each output contributes to the assembled metadata describing the street scene, including view factor proportions (greenery, sky, building), object counts, spatial layout distribution, depth complexity, and closest object information. **(b)** The QA dataset generation module uses this metadata to generate diverse base QA pairs through perception-guided and compositional-guided components. These metadata-grounded answers provide the foundation for the Chain-of-Thought (CoT) QA transformation stage, which enriches each QA pair with step-by-step natural language reasoning using an LLM.

metadata that can support reasoning-centric evaluations in vision–language tasks.

3. Methods

3.1. Street-view Image Collection

We collected 50,000 street view images at a resolution of 512×512 from five global cities—Boston, New York City, Tokyo, Seoul, and Singapore—using the Google Street View API [1]. Sampling locations were systematically selected via OpenStreetMap (OSM) road network coordinates to ensure coverage of publicly accessible streets. At each location, we captured four headings (0° , 90° , 180° , and 270°) to provide a full panoramic view of the urban context. The dataset spans diverse urban typologies, including downtown cores, residential streets, waterfronts, green corridors, and mixed-use areas, enabling broad generalization across spatial environments.

3.2. Scene Attribute Extraction

For each street-view image I , we extract scene attributes using pretrained models: semantic segmentation, object detection, and monocular depth estimation. From these, we assemble structured metadata of each image such as greenery proportion, object counts, and depth range. This metadata serves

as pseudo-ground truth, offering scalable, interpretable supervision for spatial reasoning tasks.

Semantic Segmentation. We use SegFormer [49] pretrained on Cityscapes dataset [10] to obtain pixel-wise class labels and calculate the proportions of greenery, sky, and buildings—often quantified as Green View Index (GVI), Sky View Factor (SVF), and Building View Factor (BVF) [16, 26, 33]. These metrics help characterize visual structure in urban scenes, enabling tasks like detecting dominant vertical arrangements and left–right spatial bias.

Object Detection. DETR ResNet-50 [6] is used to detect key urban scene elements (e.g., pedestrians, cars, bicycles). We extract both counts and bounding box locations, enabling questions about object presence, quantity, and co-occurrence.

Monocular Depth Estimation. Using MiDaS [39], we generate per-pixel depth maps and derive scene-level statistics such as depth range, variance, and object-wise depth averages. These inform reasoning about spatial complexity and object proximity.

Metadata Assembly. The predictions are consolidated into a unified, structured metadata record for each image. This record directly drives base QA pair generation, ensuring that

every question is traceable to specific visual evidence in the scene.

Following prior benchmarks such as SpatialVLM [7] and OmniDrive [46], we adopt a synthetic supervision strategy based on pretrained model outputs rather than human annotations. This approach carries the risk of occasional errors such as reduced pixel-level precision, but offers a practical tradeoff: enabling large-scale, consistent QA generation with broad task coverage and reproducibility, thereby supporting more transparent and controlled evaluations of spatial reasoning in vision–language models.

Table 1. Example of extracted metadata for a single street-view image.

Field	Value
Greenery Proportion	0.35
Sky Proportion	0.15
Building Proportion	0.40
Objects	person: 2, car: 5, building: 2
Depth Range	41.5
Closest Object	person
Layout	buildings: left, cars: right
Top Entity	building

3.3. QA Generation Pipeline

We generate a QA dataset in two main phases: (1) creation of *base QA pairs* with short factual answers, and (2) transformation of the answers into *Chain-of-Thought (CoT) variants* that verbalize the underlying reasoning process. Base QA generation covers two broad perspectives: a **perceptual** perspective, which can be answered directly from a single scene attribute, and a **compositional** perspective, which requires combining multiple attributes through intermediate logic. Appendix 6.1 details the number of QA pairs per question type, with templates and type-specific rules in Appendices 6.2–6.3.

3.3.1. Base QA Generation

The first phase produces base QA pairs with short, factual answers. All questions are grounded in structured metadata of segmentation, object detection, and depth estimation predictions, and fall into two broad categories:

Perceptual QA. These questions can be answered directly from a single scene attribute, using deterministic rules and interpretable thresholds. Answers are typically numeric or binary, reflecting raw scene measurements.

- **Proportions:** Scalar or binary assessments of the pixel-wise proportions of greenery, sky, and buildings.
- **Depth:** Identification of the closest object using depth maps.

- **Layout:** Inference of vertical composition and left/right dominance.
- **Objects:** Counts, presence, and co-occurrence of urban scene elements.

Perceptual thresholds are drawn from prior studies linking feature proportions (e.g., GVI > 30%) to visual dominance and aesthetic appraisal [2, 4, 32, 42, 52]. These perception-level outputs serve as structured evidence for the reasoning and CoT stages, while also enabling standalone evaluation of intuitive visual patterns such as object prominence and spatial asymmetry.

Compositional QA. These questions require integrating multiple perceptual facts and applying intermediate logic to produce a higher-order answer. The output remains short, such as yes/no, integer, or a single word, but the derivation follows a fixed, question type–specific reasoning rule.

- **Negation:** Tests the ability to process exclusions or counter-statements based on perceptual evidence.
- **Counterfactuals:** Hypothetical scenarios constructed from plausible alternatives to the observed scene attribute.
- **Multi-hop:** Multi-step comparisons and chained logic that traverse multiple perceptual attributes.

Negation and counterfactuals pose well-known challenges to both humans and AI models [14, 22, 24]. Prior work has shown that LLMs often misinterpret these forms [28, 45], underscoring the need for rigorous evaluation of compositional reasoning grounded in explicit perceptual inputs.

3.3.2. Chain-of-Thought QA Transformation

In the second phase, each base QA pair is transformed into a *CoT variant* by replacing its short factual answer with a step-by-step natural language rationale that reconstructs the original reasoning process. “Thinking-aloud” reasoning has been shown to enhance performance and interpretability in tasks such as arithmetic, commonsense inference, and multi-hop reasoning [23, 47, 48], and has recently been adapted to vision–language settings [9, 40, 50]. In our framework, CoT is not an additional question type but an *answer expansion layer* applied to both perceptual and compositional QA.

For each QA pair, Gemini 1.5-Flash [44] is prompted with the question, the scene’s metadata (Table 1), and the corresponding reasoning rule from a predefined question type-to-reasoning rule mapping table (shown in Appendix 6.4). The model is instructed to treat the metadata as visual evidence, follow the specified rule, and produce a step-by-step natural language rationale before stating the final answer, ensuring that the explanation remains fully grounded in the scene attributes extracted from Section 3.2. Figure 3 shows an example QA and detailed prompt templates are provided in Appendix 6.4.

3.3.3. Human Validation of Synthetic Supervision

Verifying the quality of generated supervision is essential in synthetic benchmarks, as it impacts the reliability and interpretability of downstream evaluations [31, 41, 43]. Because our QA pipeline uses pretrained models both to extract structured metadata and to produce CoT reasoning traces, we conducted a human evaluation of 500 randomly sampled QA pairs spanning all question types to check for (1) metadata accuracy and (2) CoT answer consistency and plausibility. Judgments were binary, focusing on perceptual plausibility rather than exact numeric precision (e.g., confirming that “3 cars” corresponds to roughly three visible cars, or that a greenery proportion of 0.47 appears visually reasonable).

Key results indicate that predictions from segmentation and depth estimation exhibited high accuracy (95% and 94%, respectively), whereas object detection showed comparatively more errors (88%), primarily due to over-counting objects not visible to human annotators. CoT reasoning was largely consistent with predefined rules (98% for consistency); however, in certain cases, the direct transformation of rule-based logic did not yield a fully plausible description of the scene’s complexity (90% for plausibility). Appendix 6.5 provides the full results with representative examples of both correct and incorrect cases.

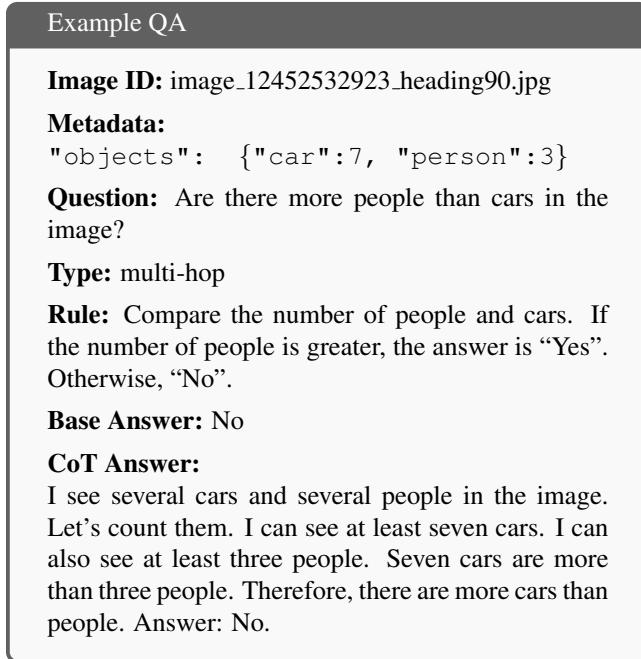


Figure 3. Example of a synthesized QA pair. The CoT answer faithfully verbalizes the rule-based derivation process using metadata.

4. Evaluation

4.1. Experimental Settings

We evaluate each question in both zero-shot and fine-tuned settings across BLIP-2 (Flan-T5-xl), InstructBLIP (Flan-T5-xl), and LLaVA-1.5-7B to isolate the effect of task-specific adaptation.

- **Zero-shot:** Evaluation on the synthetic VQA dataset without any fine-tuning, measuring each model’s inherent ability to perform spatial reasoning in urban scenes.
- **Fine-tuned:** Evaluation after fine-tuning on the synthetic VQA dataset, measuring how each model’s spatial reasoning performance changes with targeted supervision.

All models are fine-tuned using a batch size of 32, for 40 epochs, with a learning rate of 1e-4 and AdamW optimizer. The dataset is split into training, validation, and test sets with a 7:2:1 ratio, and we apply the same splits and hyperparameters across all models for fair comparison. Details regarding the metrics, answer parsing logic, and prompt constraints are in the Appendix 6.6.

4.2. Quantitative Evaluation

4.2.1. Heterogeneous Performance Trends

When we compare zero-shot performance to CoT fine-tuning across all task types, three distinct patterns emerge. **Significant improvement** group includes tasks such as *counterfactual reasoning*, *negation reasoning*, and *depth-categorical* questions. For example, as shown in Table 3, BLIP-2 shows a remarkable 509% gain in *depth-closest object* and a 591% improvement in *depth-categorical* questions. *Negation* and *counterfactual* reasoning tasks also see substantial improvements across models, with BLIP-2 achieving a 75% increase in *negation* and a 64% increase in *counterfactual* reasoning. These gains suggest that a few thousand in-domain examples can rapidly equip models to address reasoning gaps in urban scenes. **Marginal improvement** group—most notably *object presence* and *multihop reasoning*—shows more modest gains, with BLIP-2 improving by around 1% in *object presence* and 21% in *multihop* reasoning. Similarly, LLaVA-1.5 demonstrates a slight 11% improvement in *multihop* tasks and minimal gains in *object presence*. This suggests that fine-tuning aids logical chaining but may not fully overcome the inherent challenges in these complex, compositional tasks.

Performance degradation group includes tasks such as *object co-occurrence*, *layout binary*, and *proportion binary*, where models frequently show drops after fine-tuning. For instance, LLaVA-1.5 and InstructBLIP experience a 9–28% decrease in *object co-occurrence* and *proportion binary*, suggesting that adaptation to in-domain dataset can sometimes erode zero-shot strengths in simpler perceptual tasks. One plausible explanation is *catastrophic forgetting*, where fine-tuning on a dataset dominated by compositional and

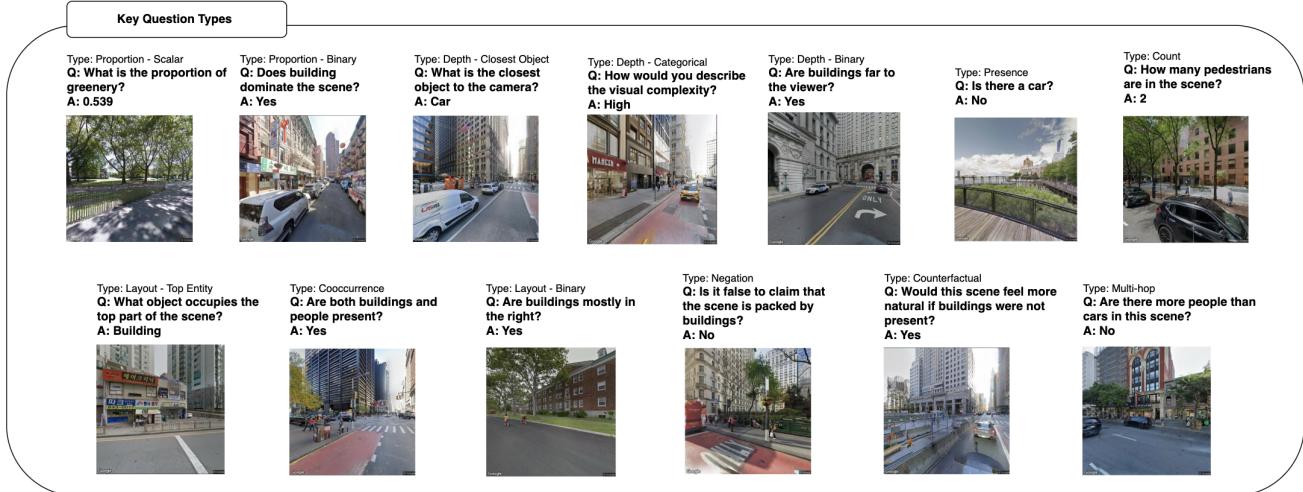


Figure 4. Example question types from the QA generation pipeline. Each example illustrates a representative question, the corresponding street-view image, and the provided answer. For brevity, the full CoT answer is shortened to the final answer.

Model	Setup	Proportions		Depth		Layout		Object		Compositional				
		Binary	Scalar	Categorical	Closest Obj.	Binary	Binary	Top entity	Count	Presence	Co-occ.	Negation	Counterf.	
		F1↑	MAE↓	F1↑	F1↑	F1↑	F1↑	F1↑	MAE↓	F1↑	F1↑	F1↑	F1↑	
LLaVA-1.5	Zero-shot	0.59	0.18	0.62	0.24	0.69	0.61	0.09	2.36	0.95	0.99	0.31	0.35	0.36
	Fine-tuned	0.44	0.22	0.50	0.12	0.80	0.58	0.15	3.03	0.97	0.90	0.72	0.33	0.40
InstructBLIP	Zero-shot	0.47	0.21	0.54	0.22	0.99	0.58	0.03	4.33	0.96	0.99	0.53	0.37	0.80
	Fine-tuned	0.11	0.27	0.62	0.10	0.88	0.45	0.02	4.05	1.00	0.95	0.40	0.48	0.72
BLIP-2	Zero-shot	0.33	0.21	0.11	0.11	0.99	0.58	0.06	4.10	0.97	0.91	0.52	0.55	0.67
	Fine-tuned	0.89	0.11	0.76	0.67	0.89	0.57	0.87	1.70	0.98	1.00	0.91	0.90	0.81

Table 2. Performance of vision–language models on perceptual QA and compositional QA tasks. Results are reported for zero-shot and fine-tuned settings. Bold indicates the better performance within each model and metric.

reasoning-heavy samples shifts the model’s feature representations away from low-level perceptual cues. A second factor could be *data distribution mismatch*, as our fine-tuning set contains relatively fewer straightforward perception cases. Our fine-tuning also emphasizes rule-based reasoning traces, which could draw attention toward abstract scene logic at the expense of rapid, single-step perceptual judgments. These findings suggest pairing domain-specific adaptation with strategies such as rehearsal data, multi-task balancing, or selective freezing to preserve perceptual competence while boosting reasoning performance.

4.2.2. Model Efficiency and Practical Trade-offs

As shown in Figure 5, BLIP-2 offers the most parameter-efficient gains, achieving substantial perception and reasoning improvements despite being the smallest model in this comparison. In contrast, LLaVA-1.5 specializes in reasoning tasks, showing strong reasoning gains with minimal parameter overhead but a noticeable decline in perception tasks, such

as a 25% drop in *proportion binary* and a 28% increase in *proportion scalar* MAE. Notably, LLaVA-1.5’s robust zero-shot performance makes it an appealing option for deployment in settings with minimal or no task-specific supervision. InstructBLIP, despite its larger model size, demonstrates limited parameter efficiency and suffers from perceptual performance degradation across several tasks, including a 77% decrease in proportion binary and a 29% increase in scalar MAE, suggesting that more targeted fine-tuning strategies may be required to fully leverage its potential. Taken together, these results emphasize the need for careful model selection based on the target domain and computational constraints. BLIP-2 emerges as an effective, lightweight model for domain-specific fine-tuning pipelines like ours, providing strong perception and reasoning gains with low computational cost, while LLaVA-1.5 remains a valuable option in reasoning-focused applications, especially when fine-tuning budgets are limited and zero-shot robustness is critical.

Model	Setup	Proportions (%Δ)		Depth (%Δ)			Layout (%Δ)		Object (%Δ)			Compositional (%Δ)		
		Binary	Scalar	Categorical	Closest Obj.	Binary	Binary	Top Entity	Count	Presence	Co-occ.	Negation	Counterf.	Multihop
LLaVA-1.5	%Δ	-25.4	-22.2	-19.4	-50.0	+15.9	-4.9	+66.7	-28.4	+2.1	-9.1	+132.3	-5.7	+11.1
InstructBLIP	%Δ	-76.6	-28.6	+14.8	-54.5	-11.1	-22.4	-33.3	+6.5	+4.2	-4.0	-24.5	+29.7	-10.0
BLIP-2	%Δ	+169.7	+47.6	+590.9	+509.1	-10.1	-1.7	+1350.0	+58.5	+1.0	+9.9	+75.0	+63.6	+20.9

Table 3. Percentage change from zero-shot to fine-tuned for each model and task type. Positive values indicate improvement (higher F1 or lower MAE), while negative values indicate performance degradation.

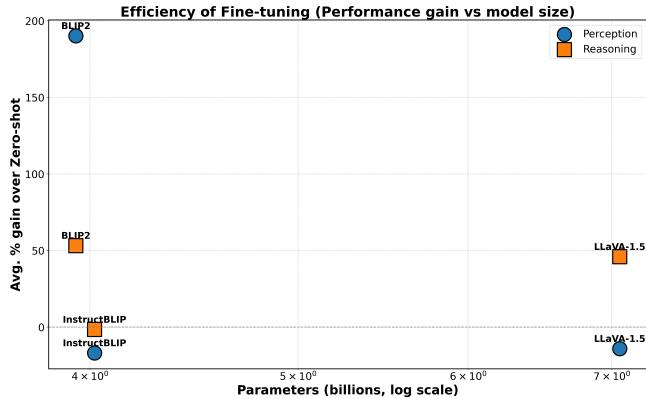


Figure 5. Parameter efficiency of vision–language models based on fine-tuning gains. The x-axis shows the model size in billions of parameters (log scale): LLaVA-1.5 (7.06B), InstructBLIP (4.02B), and BLIP-2 (3.94B). The y-axis shows the average percentage performance gain after fine-tuning compared to zero-shot performance.

4.3. Qualitative Evaluation

Beyond numerical performance, our qualitative evaluation highlights that different models exhibit distinct strengths across question types. Notably, BLIP-2 consistently demonstrates versatility and robust reasoning capabilities across a wide range of tasks. For example, as illustrated in Figure 6, in a representative negation-type question, fine-tuning enhances reasoning quality for both InstructBLIP and BLIP-2 compared to their zero-shot counterparts. However, while InstructBLIP tends to remain at the level of question repetition—stating that “the statement says that neither buildings nor sky dominates this scene”—BLIP-2 engages in more explicit logical reasoning, accurately resolving the double-negative structure of the question by deducing: “while neither the sky nor the building takes up the majority of the image, they are both present and visible. Therefore, it’s false to say that neither dominates the scene.”

While LLaVA-1.5 exhibits comparatively lower average performance, as indicated by its F1 scores in Table 2, it nonetheless produces occasional high-quality answers. We report one example of a depth task in Figure 6. The model consistently provides detailed and contextually appropriate assessments of scene depth, even when other models strug-

gle. Additionally, InstructBLIP shows notable proficiency in counterfactual reasoning. Fine-tuning substantially improves its ability to perform multi-step hypothetical reasoning beyond surface-level descriptions, as shown in the counterfactual example in the same figure.

Collectively, these qualitative observations complement the quantitative results, indicating that while fine-tuning with a domain-specific synthetic QA dataset generally improves model performance, each model retains distinct strengths more apparent through qualitative analysis.

5. Conclusions and Future Work

In this study, we evaluated the ability of general-purpose vision–language models to understand fine-grained spatial relationships in street-view images. By introducing a structured pipeline for generating diverse, spatially grounded QA tasks, our work establishes a new problem domain in VL research and creates opportunities for advancing domain-specific perception and reasoning.

Comprehensive experiment results show that fine-tuning with our synthetic QA dataset leads to substantial performance gains. Lightweight models like BLIP-2 particularly benefit from this structured supervision, achieving gains in perception and reasoning capabilities with minimal task-specific data. These findings highlight the potential of synthetic QA with CoT supervision as a versatile approach for enhancing spatial understanding in urban scenes and other domain-specific contexts. For larger models like LLaVA-1.5, our study suggests that synthetic QA alone may be insufficient to shift its pretrained distribution. Addressing this will likely require more complex QA structures and advanced instruction tuning methods, presenting a key direction for future research.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No.2022M3J6A1063021, No.RS-2025-00517342).

Image	Question	Model	Setting	Answer
	Q: If two more people entered the scene, would it look crowded? Respond "yes" or "no".	InstructBLIP	Zero-shot	no
			Fine-tuned	I see four people in the image. If two more people were to enter the scene, there would be six people in total. Six is greater than or equal to five. Therefore, the area would look crowded. answer: yes
	BLIP-2	Zero-shot	Fine-tuned	if two more people entered the scene would it look crowded
			Fine-tuned	i see one person in the image. if two more people entered the scene, there would be a total of three people. three is less than five. therefore, the scene would not look crowded. answer: no
	LLaVA-1.5	Zero-shot	Zero-shot	No
			Fine-tuned	yes
Ground-truth Yes				
	Q: Is it false to say that neither buildings nor sky dominate this scene? Respond in "yes" or "no".	InstructBLIP	Zero-shot	No
			Fine-tuned	the image shows a scene with a building and sky dominating the scene. the statement says that neither buildings nor sky dominates this scene. answer: no
	BLIP-2	Zero-shot	Zero-shot	No
			Fine-tuned	looking at the image, i see a significant portion of the sky at the top. there's also a building visible on the right side. while neither the sky nor the building takes up the majority of the image, they are both present and visible. therefore, it's false to say that neither dominates the scene. answer: no
	LLaVA-1.5	Zero-shot	Zero-shot	No
			Fine-tuned	no no no no (...omitted repetitive text...) answer: no
Ground-truth No				
	Q: Does the scene appear visually complex? Answer in yes or no.	InstructBLIP	Zero-shot	Low is the overall depth complexity of the scene, as there are only a few objects in the scene. There is a low amount of detail in the scene, which can be attributed to the low number of objects in the scene. (...)
			Fine-tuned	the looking at the overall depth complexity of this scene. label the overall depth complexity of this scene label the overall depth complexity of this scene label the overall depth complexity of this scene (...) answer: (n/a)
	BLIP-2	Zero-shot	Zero-shot	low
			Fine-tuned	looking at the scene, i observe a significant variation in the distances of objects from the camera. the depth range appears quite extensive. the depth complexity is labeled 'high'. answer: high
	LLaVA-1.5	Zero-shot	Zero-shot	Moderate
			Fine-tuned	since the depth of the scene is dominated by a wide, clear, and unobstructed road with several cars and a few potted plants, there's not much complexity. the depth range is quite large, exceeding 20 units. therefore, the depth complexity is high. answer: high
Ground-truth High				

Figure 6. Qualitative answers from three VLMs—InstructBLIP, BLIP-2, and LLaVA-1.5—under zero-shot and fine-tuned settings for representative question types.

References

- [1] Dragomir Anguelov, Carole Dulong, Daniel Filip, Christian Frueh, Stéphane Lafon, Richard Lyon, Abhijit Ogale, Luc Vincent, and Josh Weaver. Google street view: Capturing the world at street level. *Computer*, 43(6):32–38, 2010. 3
- [2] Yoji Aoki. Evaluation methods for landscapes with greenery. *Landscape Research*, 16:3–6, 1991. 4
- [3] Mondira Bardhan, Fu Li, Mathew H.E.M. Browning, Jiaying Dong, Kuiran Zhang, Shuai Yuan, Hüseyin Ertan İnan, Olivia McAnirlin, Dani T. Dagan, Allison Maynard, Katie Thurson, Fan Zhang, Ruoyu Wang, and Marco Helbich. From space to street: A systematic review of the associations between visible greenery and bluespace in street view imagery and mental health. *Environmental Research*, 2024. 1
- [4] Anna-Maria Bolte, Benjamin Niedermann, Thomas Kistemann, Jan-Henrik Haunert, Youness Dehbi, and Theo Kötter. The green window view index: automated multi-source visibility analysis for a multi-scale assessment of green window views. *Landscape Ecology*, 39(3):71, 2024. 4
- [5] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giacarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2
- [6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. 2020. 3
- [7] Boyuan Chen, Zhuo Xu, Sean Kirmani, Danny Driess, Pete Florence, Brian Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *CVPR*, 2024. 2, 4
- [8] Mingze Chen, Yuxuan Cai, Shuying Guo, Ruilin Sun, Yang Song, and Xiwei Shen. Evaluating implied urban nature vitality in san francisco: An interdisciplinary approach combining census data, street view images, and social media analysis. *Urban Forestry Urban Greening*, 95:128289, 2024. 1
- [9] Zhenfang Chen, Qinhong Zhou, Yikang Shen, Yining Hong, Zhiqing Sun, Dan Gutfreund, and Chuang Gan. Visual chain-of-thought prompting for knowledge-based visual reasoning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(2):1254–1262, 2024. 4
- [10] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 2, 3
- [11] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructclip: Towards general-purpose vision-language models with instruction tuning. In *NIPS*, 2023. 2
- [12] Fabio Duarte, Deborah Lefosse, Rohit Sanatani, Yuhao Kang, Arjan Timmeren, and Carlo Ratti. Feeling Nature: Measuring perceptions of biophilia across global biomes using visual AI, 2024. ISSN: 2693-5015. 1
- [13] Abhimanyu Dubey, Nikhil Naik, Devi Parikh, Ramesh Raskar, and César A. Hidalgo. Deep Learning the City: Quantifying Urban Perception at a Global Scale. In *Computer Vision – ECCV 2016*, pages 196–212, Cham, 2016. Springer International Publishing. 1
- [14] Carolin Dudschig, Barbara Kaup, Mingya Liu, and Juliane Schwab. The processing of negation and polarity: An overview. *Journal of Psycholinguistic Research*, 50(6):1199–1213, 2021. 4
- [15] Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. Vqa-lol: Visual question answering under the lens of logic. In *European conference on computer vision*, pages 379–396. Springer, 2020. 2
- [16] Fang-Ying Gong, Zhao-Cheng Zeng, Fan Zhang, Xiaojiang Li, Edward Ng, and Leslie K. Norford. Mapping sky, tree, and building view factors of street canyons in a high-density urban environment. *Building and Environment*, 134(1):155–167, 2018. 3
- [17] Xuan He and Sylvia Y. He. Using mobile phone big data and street view images to explore the mismatch between walkability and walking behavior. *Transportation Research Part A: Policy and Practice*, 180:103946, 2024. 1
- [18] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural Adversarial Examples, 2021. arXiv:1907.07174 [cs]. 1
- [19] Yanjun Hu, Fengtao Qian, Hai Yan, Ariane Middel, Renwu Wu, Minghui Zhu, Qian Han, Kechun Zhao, Han Wang, Feng Shao, and Zhiyi Bao. Which street is hotter? street morphology may hold clues -thermal environment mapping based on street view imagery. *Building and Environment*, 262:111838, 2024. 1
- [20] Drew A. Hudson and Christopher D. Manning. 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 (CVPR)*, 2019. 2
- [21] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2
- [22] Nora Kassner and Hinrich Schütze. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. *arXiv preprint arXiv:1911.03343*, 2019. 4
- [23] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *NIPS*, 2022. 4
- [24] Eugenia Kulakova and Mante S Nieuwland. Understanding counterfactuality: A review of experimental evidence for the dual meaning of counterfactuals. *Language and Linguistics compass*, 10(2):49–65, 2016. 4
- [25] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023. 2
- [26] Xiaojiang Li, Chuanrong Zhang, Weidong Li, Robert Ricard, Qingyan Meng, and Weixing Zhang. Assessing street-level urban greenery using google street view and a modified green view index. *Urban Forestry Urban Greening*, 14:675–685, 2015. 1, 3

- [27] Xinyao Li, Jingjing Li, Fengling Li, Lei Zhu, Yang Yang, and Heng Tao Shen. Generalizing Vision-Language Models to Novel Domains: A Comprehensive Survey, 2025. arXiv:2506.18504 [cs]. 1
- [28] Yian Li, Wentao Tian, Yang Jiao, Jingjing Chen, and Yu-Gang Jiang. Eyes can deceive: Benchmarking counterfactual reasoning abilities of multi-modal large language models. *arXiv e-prints*, pages arXiv–2404, 2024. 4
- [29] Xiucheng Liang, Jiat Hwee Chang, Song Gao, Tianhong Zhao, and Filip Biljecki. Evaluating human perception of building exteriors using street view imagery. *Building and Environment*, 263:111875, 2024. 1
- [30] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, 2024. 2
- [31] Gaurav Maheshwari, Dmitry Ivanov, and Kevin El Haddad. Efficacy of synthetic data as a benchmark. *arXiv preprint arXiv:2409.11968*, 2024. 5
- [32] Chunping Miao, Shuai Yu, Yuanman Hu, Huiwen Zhang, Xingyuan He, and Wei Chen. Review of methods used to estimate the sky view factor in urban street canyons. *Building and Environment*, 168:106497, 2020. 4
- [33] Shohei Nagata, Tomoki Nakaya, Tomoya Hanibuchi, Shiro Amagasa, Hiroyuki Kikuchi, and Shigeru Inoue. Objective scoring of streetscape walkability related to leisure walking: Statistical modeling approach with semantic segmentation of google street view images. *HealthPlace*, 66:102428, 2020. 1, 3
- [34] Nikhil Naik, Jade Philipoom, Ramesh Raskar, and Cesar Hidalgo. Streetscore - predicting the perceived safety of one million streetscapes. In *CVPR*, 2014. 1
- [35] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kortschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *ICCV*, 2017. 2
- [36] Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12700–12710, 2021. 2
- [37] Yoshiki Ogawa, Takuya Oki, Chenbo Zhao, Yoshihide Sekimoto, and Chihiro Shimizu. Evaluating the subjective perceptions of streetscapes using street-view images. *Landscape and Urban Planning*, 247:105073, 2024. 1
- [38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *NeurIPS*, 2021. 1
- [39] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, 2020. 2, 3
- [40] Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. In *Advances in Neural Information Processing Systems*, pages 8612–8642. Curran Associates, Inc., 2024. 4
- [41] Krishnakant Singh, Thanush Navaratnam, Jannik Holmer, Simone Schaub-Meyer, and Stefan Roth. Is synthetic data all we need? benchmarking the robustness of models trained with synthetic images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2505–2515, 2024. 5
- [42] Marie K. Svensson. Sky view factor analysis – implications for urban air temperature differences. *Meteorological Applications*, 11(3):201–211, 2004. 4
- [43] Hui Tang and Kui Jia. A new benchmark: On the utility of synthetic data with blender for bare supervised learning and downstream domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15954–15964, 2023. 5
- [44] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 4
- [45] Tereza Vrabcová, Marek Kadlecík, Petr Sojka, Michal Štefánik, and Michal Spiegel. Negation: A pink elephant in the large language models' room? *arXiv preprint arXiv:2503.22395*, 2025. 4
- [46] Shihao Wang, Zhiding Yu, Xiaohui Jiang, Shiyi Lan, Min Shi, Nadine Chang, Jan Kautz, Ying Li, and Jose M. Alvarez. Omnidrive: A holistic vision-language dataset for autonomous driving with counterfactual reasoning. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 22442–22452, 2025. 2, 4
- [47] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *ICLR*, 2023. 4
- [48] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *NIPS*, 2022. 4
- [49] Enze Xie, Wenhui Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *NIPS*, 2021. 3
- [50] Ruohong Zhang, Bowen Zhang, Yanghao Li, Haotian Zhang, Zhiqing Sun, Zhe Gan, Yinfei Yang, Ruoming Pang, and Yiming Yang. Improve vision language model chain-of-thought reasoning, 2024. 4
- [51] Yuhui Zhang, Yuchang Su, Yiming Liu, and Serena Yeung-Levy. Negvqa: Can vision language models understand negation? *arXiv preprint arXiv:2505.22946*, 2025. 2
- [52] Huaizhen Zhu, Fan Yang, Zhiyi Bao, and Xinge Nan. A study on the impact of visible green index and vegetation structures on brain wave change in residential landscape. *Urban Forestry Urban Greening*, 64:127299, 2021. 4