Spatial Reasoning in Multimodal LLMs via CoT Distillation and Monte Carlo Tree Search for Dutch Facade-Element Detection: An Exploratory Study

Thesis submitted to Utrecht University for the degree of
MSc Artificial Intelligence, July 2025.

Riccardo Campanella

8175721

Examiners: Dr. Metehan Doyran, Dr. Itir Önal Ertuğrul
External Supervisor: Raphaël Gueulet (TNO)

Presentation Outline

1. Motivation & Problem Statement
2. Research Questions & Contributions
3. Background
4. Data Collection & Methodology
5. Exploratory Results Analysis
6. RQ Discussion & Limitations
7. Conclusions & Future Work

Motivation: Building Energy Renovation

The Challenge

- Dutch buildings need energy retrofits for sustainability goals
- Manual facade assessment is time-consuming and expensive
- Current Computer Vision models lack domain-specific knowledge

The Opportunity

- Multimodal LLMs offer contextual reasoning capabilities
- Can process natural language descriptions of architectural features
- Potential for zero-shot transfer to specialized domains

Problem Statement

MQA Target Features:

1. Weep holes
2. Crawling space
3. Chimneys
4. Pitched roof
5. Facade ventilation
6. Roof ventilation
7. Window ventilation
8. Dormers
9. Roof windows
10. Attics living spaces
11. Vegetation growth
12. Balconies
13. Photovoltaic panels
14. Window count

Reasoning Complexity:

Visual Recognition → Geometric Inference → Semantic Understanding → Context Analysis

Research Questions

Main Research Question

Are SoTA* Multimodal LLMs beneficial to identify applicable housing renovation concepts on Dutch building facades?

RQ1: Model Comparison

How does Chain-of-Thought reasoning (Qwen) compare to 3D scene graph methods (SpatialRGPT) in zero-shot prediction?

- Performance vs. SoTA models (GPT-4o)
- Impact of bounding box guidance

RQ2: Enhancement Methods

How can CoT reasoning MLLMs be enhanced for spatial recognition?

- Effect of 3D scene graph augmentation
- LoRA fine-tuning capabilities

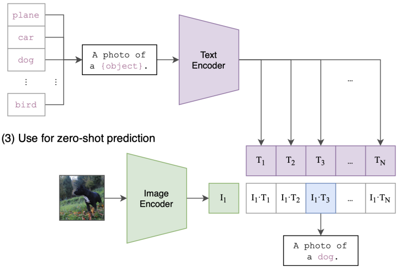
Key Contributions

1. **Comprehensive MLLM Evaluation Framework**
   - **First** systematic evaluation on real Dutch facade data
   - **Baseline performance** insights for architectural features
2. **DuTCh SpaCE: Spatial Reasoning Enhancement**
   - **Mitigation**: Novel dual-teacher distillation framework to mitigate Hallucinations
   - **Compensation**: Reasoning compensates for limited visual grounding
3. **Reasoning vs. Grounding Trade-off Analysis**
   - **Knowledge Transfer**: Domain expertise vs. model scale
   - **Fast Scaling**: Accessible path to spatial reasoning enhancement
4. **Practical Domain Adaptation Insights**
   - LoRA + Knowledge Distillation + Test-time Search
   - **Few-shot**: Specialization framework for low-data domains

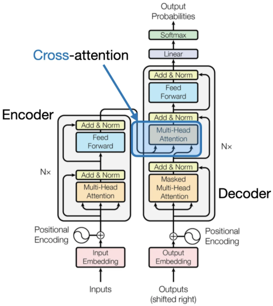
Background: Multimodal Large Language Models

- **June, 2017 — Transformer**: **LLM (Encore-Decoder)**
- **2021 — CLIP**: **Contrastive (Encoder-Encoder)**
- **2023 — LLaVA**: **Multi-stage (Decoder-only)**
- **May, 2024 — GPT-4o**: **Multi-stage Reasoning**
- **June, 2024 — SRGPT**: **Spatial Reasoning**
- **Sept, 2024 — Qwen2-VL**: **Multi-stage Reasoning**

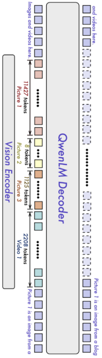
CLIP



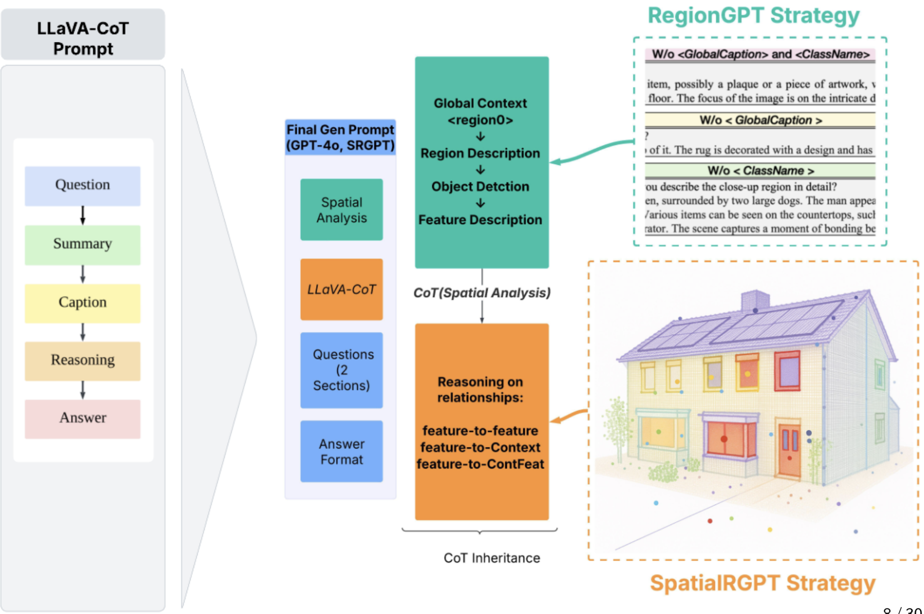
Transformers



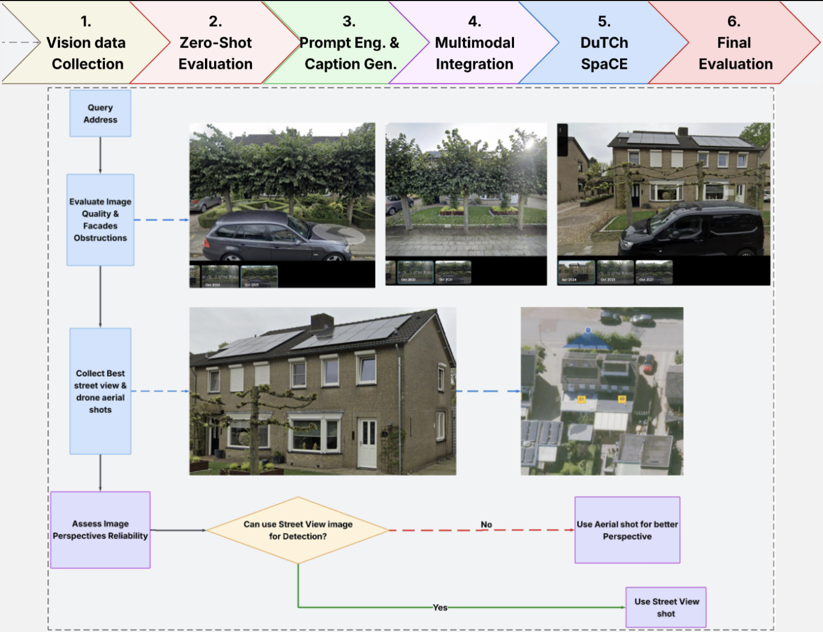
Qwen2-VL



Background: Grounding Chain-of-Thought Reasoning

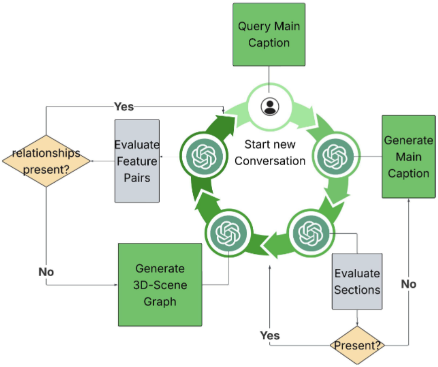


Data Collection: Few-shot Dutch Building Facade

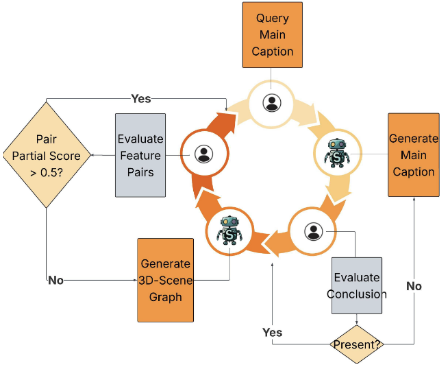


Multimodal integration: Captions and Scenes

Teacher 1: GPT-4o



Teacher 2: SpatialRGPT-bbox



Train Set (32)

Label	Distribution
No	44.7%
Yes	32.5%
Unknown	22.9%

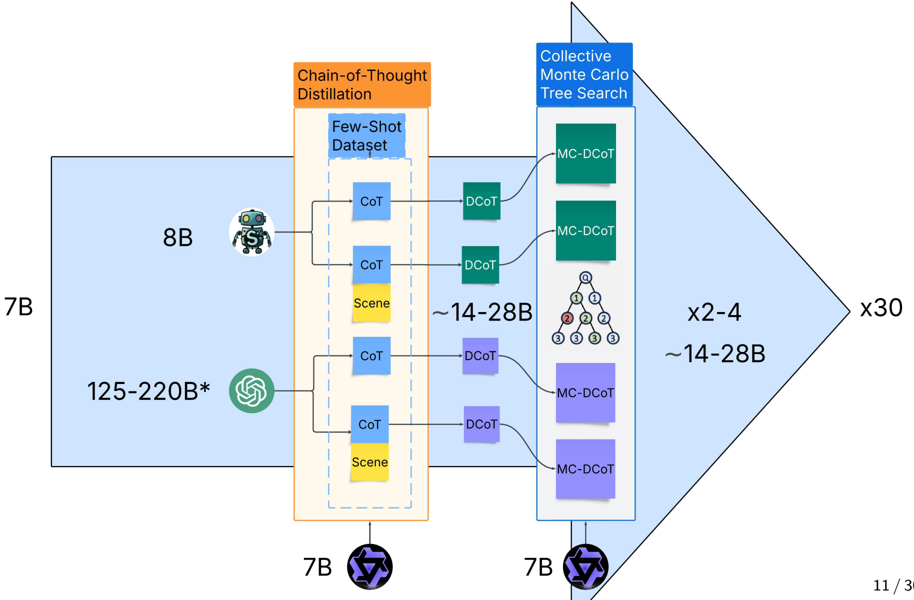
Val Set (4)

Label	Distribution
No	40.4%
Yes	46.2%
Unknown	13.5%

Test Set (9)

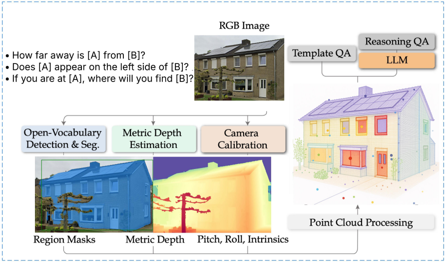
Label	Distribution
No	44.4%
Yes	39.3%
Unknown	16.2%

Dual-Teacher Chain-of-Thought Framework for Spatial Facade Reasoning (DuTCh Space)



Methodology: Models Under Study

1. GPT-4o
2. **SpatialRGPT-8B (Base/Bbox)**
3. Qwen2-7B-VL-Instruct



Categories	**Datasets**
Captioning	Image Paragraph Captioning [77], MSR-VTT [78], TextCaps [79]
Reasoning	CLEVR [80], NLVR [81], VisualMRC [82]
Translation	Multi30k [83]
VQA	ActivityNet-QA [84], DocVQA [85], GQA [86], iVQA [87], MSRVTT-QA [78], MSVD-QA [78], OCR-VQA [88], ST-VQA [89], ViQuAE [90], VQAv2 [91], Visual Dialog [92]

Categories	**Datasets**
Classification	V3Det [93], COCO [40], LVIS [94]
Caption	V3Det [93] VG [95], RefCOCO [96]
Relationship	VG [95]
REC	RefCOCO [96]

Table 11: The region-level SFT blend [19] we used.

User: Can you tell me if this house has photovoltaic panels?

SpatialRGPT: Yes, Ground Truth (GT)=Yes

Two photovoltaic arrays are visible on the pitched roof. The surface geometry and inclination match standard PV module dimensions (approx. 1m x 1.7m each), and they are aligned symmetrically on the left and right roof planes.

User: Does the house include a crawling space?

SpatialRGPT: Unknown, GT=Yes

There's no visible vent grid or elevation suggesting a ventilated plinth. Sub-floor access may be internal or obscured in this point cloud angle.

User: Are attics living spaces included?

SpatialRGPT: No, GT=No

No dormer extensions or roof windows suggestive of habitable attic conversion are visible. The upper roof volume appears sealed, likely used for insulation or storage only.

User: What's the window count on the front facade?

SpatialRGPT: 6 Windows, GT=7

3 on the ground floor, and 3 on the upper floor. No side windows detected from this viewpoint.

CoT Distillation by Imitation via CE Minimization

$$\mathcal{L}_{\text{DCoT}} = -\frac{1}{T}\sum_{t=1}^{T} \log p_{\theta_S}(y_t^{(T)} \mid y_{<t}^{(T)}, x_{mm})$$

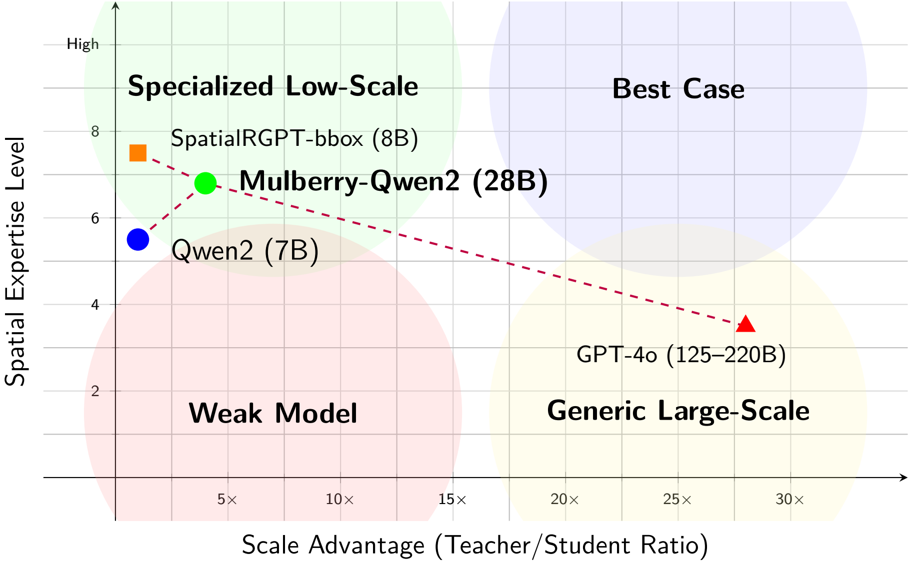
Cross-Entropy vs Entropy Minimization

- **Assumption**: Correct Reasoning yields Low Entropy
- **Alignment**: Convergence & Overfitting
- **Specificity**: Penalizes incorrect but confident predictions
- **Correctness**: Abstract reasoning transfer

Multimodal vs Distilled CoT

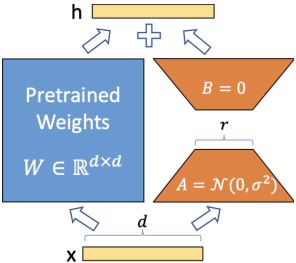
Configuration	**MCoT**	**DCoT (Ours)**
Train Data	Human-annotated rationales	Raw Teacher rationales
Train Size	29K (ScienceQA + A-OKVQA)	32 examples
Train Paradigm	SFT on CE Loss	KD via SFT on CE Loss
Train Components	Vision-language fusion layers	LLM attention modules only
Inference Process	Two-stage (rationale → answer)	Single-stage (end-to-end)
Parameter Efficiency	Full fine-tuning	LoRA
Reasoning Type	Multimodal	Complex chains

Scaling Laws: Teacher-Student Knowledge Distillation

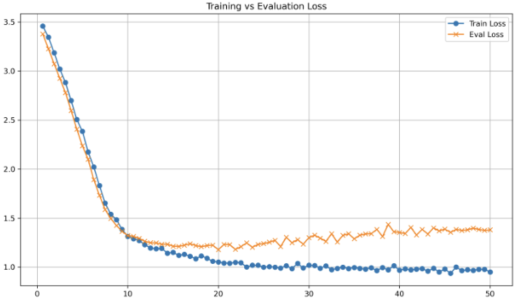


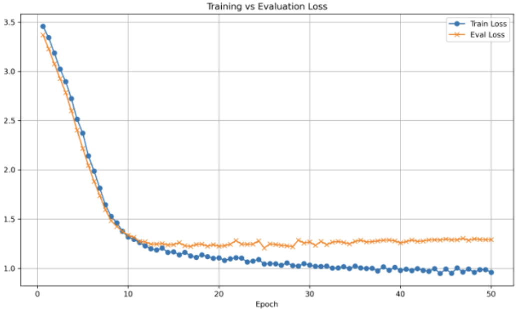
Our methodology can be further scaled!

Parameter-Efficient Fine-tuning with LoRA

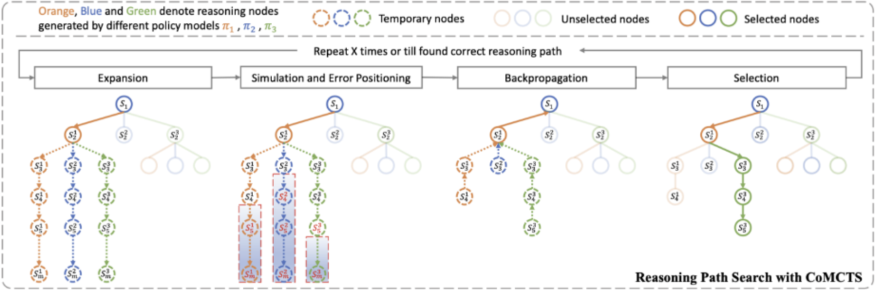


- `learning_rate`: 2e-5 (linear warmup)
- r (rank): 16
- `lora_alpha`: 32
- `early_stopping`: patience = 3
- `target_modules`: ["q_proj", "k_proj", v_proj", "o_proj"]



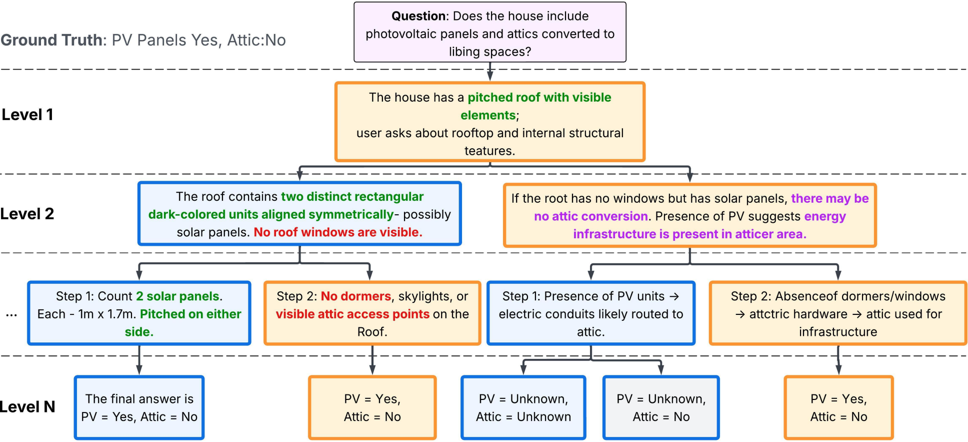


Test-Time Enhancement: Mulberry-Qwen CoMCTS



Configuration	Mulberry	Mulberry-Qwen (Ours)
Model Selection	GPT-4o, LLaMA, Qwen	Qwen-7B + Qwen-DCoT
Search Strategy	CoMCTS	CoMCTS
Training Strategy	SFT on Mulberry-260K	Zero-shot
Max Iterations	20	3
Reasoning Variation	High (inter-model)	Reduced (local/global)
Reasoning Robustness	Diverse patterns	Consistent domain-specific patterns
Bias Mitigation	Cross-model vote	Homogeneous model vote
Hallucination Reduction	High (multi-model)	Low (same-family)
Test Time Compute	Higher	Lower

Qualitative Illustration of Qwen-CoMCTS Spatial Reasoning



Evaluation Protocol & Metrics

Experimental Design

- **Zero-shot:** All models on 45 images
- **Fine-tuned:** 4 Qwen variants on test set
- **Multiple runs:** 10 evaluations per configuration
- **Robustness:** Average $\pm$ standard deviation

Table: Performance Metrics and Their Evaluation Focus

Metric(s)	**Evaluation Focus**
Accuracy, Balanced Accuracy	$\rightarrow$ Generalization
Precision, Recall, F1-score	$\rightarrow$ Hallucination Control
MAE / MSE	$\rightarrow$ Counting Accuracy (e.g., windows)
Feature-wise Analysis	$\rightarrow$ Spatial Reasoning via Feature Complexity

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Results Overview: Model Performance

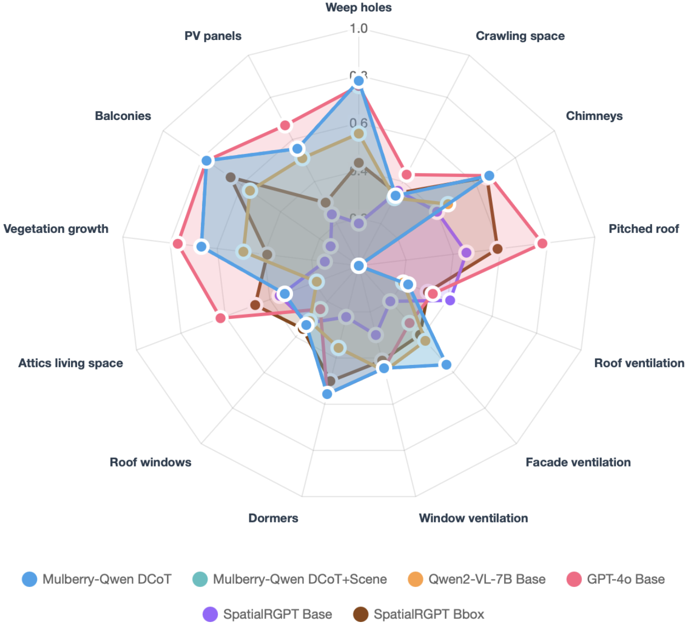
Table: Performance Comparison Across Models

Model	Accuracy	B. Acc.	Prec.	Recall	F1	MAE
GPT-4o*	**0.567**	**0.458**	**0.362**	**0.429**	**0.373**	**2.60 ± 0.14**
Mulberry-Qwen	0.479	0.370	0.178	0.333	0.226	3.78
Qwen2-VL-7B*	0.371	0.341	0.181	0.274	0.200	4.44
SpatialRGPT-bbox*	0.446	0.371	0.197	0.372	0.240	3.31 ± 0.29
SpatialRGPT*	0.287	0.330	0.139	0.311	0.171	6.67

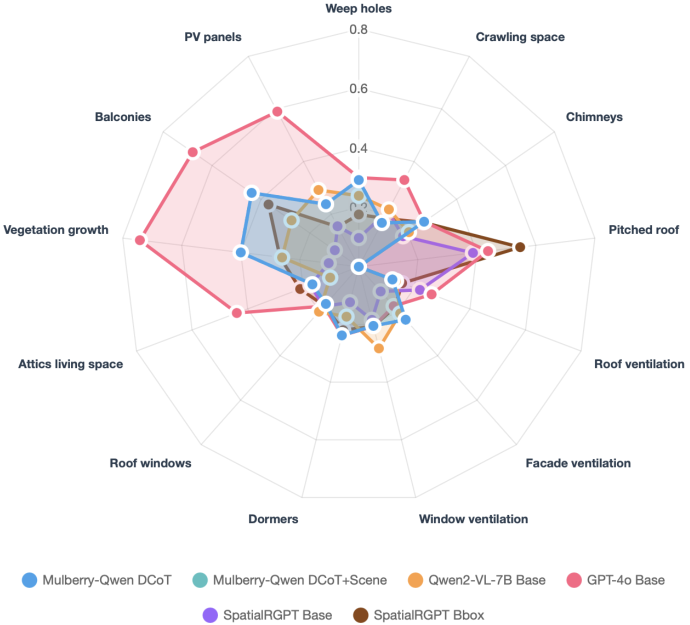
Key Findings

- **GPT-4o dominates** across all metrics
- **Mulberry-Qwen achieves 11% accuracy improvement** over baseline
- **Bounding box guidance helps SpatialRGPT** significantly
- **Gap with GPT-4o reduced**: 20%-8% in Acc., 11-8% in B. Acc.
- **Hallucinations** reduced significanlty with DuTCh SpaCE
- **Robustness:** All models with (*) showed std. dev. so Mulberry-Qwen is the most robust overall and feature-wise

Feature Performance (Accuracy)



Feature Hallucinations (F1-score)



Summary: DuTCh Space Achievements

MLLM Benefits Before	with DuTCh Space	Limitations After
• Zero-shot transfer capabilities • Contextual reasoning about features • Interpretable explanations via CoT • Scalable to large building stocks	• Knowledge distillation works • Test-time compute helps • Small models can compete • Net effect: Reduced under-prediction (FN), slight over-confidence (FP)	• Spatial complexity challenges • Numerical reasoning gaps • Domain-specific biases • Computational requirements

Key Takeaway

- **Successful Mitigation:** DCoT + MC Tree Search reduces hallucinations by promoting evidence-based reasoning
- **Reasoning compensates** for limited visual grounding
- **Domain expertise** can rival raw scale

RQ1 Findings: Model Architecture Comparison

RQ1a: SoTA Comparison

GPT-4o superior across all metrics (20-30% gap with smaller models)

- Benefits from scale (∼125-220B vs. 7-8B parameters)
- Comprehensive multimodal training
- Better generalization to Dutch-specific features

RQ1b: Bounding Box Guidance

SpatialRGPT with bounding boxes: 15% accuracy improvement

- Helps with spatial localization
- Reduces false positives
- Doesn't improve balanced accuracy significantly
- Additional annotation step

Architecture Insights: Different MLLMs have complementary strengths for different feature types

RQ2 Findings: Enhancement Methods

RQ2a: Scene Graph Augmentation

No performance difference between DCoT with/without scene graphs

- Qwen architecture may not effectively utilize explicit spatial representations
- DCoT reasoning already captures sufficient spatial relationships
- Need deeper integration beyond simple text augmentation

RQ2b: LoRA Fine-tuning

Significant improvements across all metrics

- 10% accuracy improvement over baseline
- Reduces gap with GPT-4o from 20% to 8%
- Quality reasoning in few examples can match mitigate lack of extensive
- Maintains general capabilities while adding domain expertise

Limitations & Challenges

Dataset Limitations

- **Scale:** Only 45 images (vs. typical 100K+ datasets)
- **Annotation:** Subjective "unknown" categories
- **Imbalance:** Most features absent in dataset

Technical Limitations

- **Visual Grounding:** No vision encoder fine-tuning
- **Spatial Integration:** Scene graphs not effectively utilized
- **Problem Complexity:** 14 questions in single prompt
- **Computational:** Limited test-time search iterations

Methodological Considerations

- Cross-entropy loss vs. preference learning
- LoRA vs. full fine-tuning trade-offs
- Teacher quality vs. scale in distillation

Future Work & Research Directions

Immediate Extensions

- **Scale dataset:** 1000+ images from web scraping + automated filtering
- **Problem decomposition:** Split into building-section-specific questions
- **Vision Grounding:** Fine-tune multimodal components
- **Advanced search:** Increase CoMCTS iterations and model diversity

Advanced Methodologies

- **RLHF:** Human preference optimization for spatial reasoning
- **Multi-adapter:** Feature-specific LoRA modules
- **Grounded CoT:** Visual evidence linking in reasoning chains
- **DoRA:** Weight decomposition for stable fine-tuning

Broader Impact & Applications

Energy & Sustainability

- Accelerate building renovation assessment
- Reduce manual inspection costs (time & labor)
- Enable large-scale retrofit planning
- Support EU Green Deal objectives

Technical Contributions

- **Domain Adaptation:** Framework for specialized applications
- **Multimodal Reasoning:** Insights into reasoning vs. grounding
- **Knowledge Distillation:** Teacher-student dynamics in few-shot settings
- **Test-Time Compute:** Practical application in complex reasoning

Conclusions

DuTCh SpaCE key findings

1. **MLLMs are beneficial** for facade analysis with proper enhancement
2. **Reasoning can compensate** for limited visual grounding capabilities
3. **Domain expertise rivals scale** in knowledge transfer scenarios

Theoretical Contributions

- First systematic MLLM evaluation on few-shot Dutch architecture
- Novel dual-teacher Distillation Framework for Reasoning vs Grounding
- Scaling laws in few-shot multimodal domain adaptation

Practical Impact

- Accessible path to spatial reasoning enhancement
- Framework for domain-specific MLLM adaptation
- Cost-effective alternative to explicit spatial grounding
- Foundation for automated building assessment systems

Acknowledgements

Supervision & Guidance

- **Dr. Metehan Doyran** - First Examiner, Utrecht University
- **Dr. Itir Önal Ertuğrul** - Second Examiner, Utrecht University
- **Raphaël Gueulet** - External Supervisor, TNO Machine Learning Engineer

TNO Research Support

- **Paolo de Heer** - Data Scientist, **Wietske van Kanten-Roos & J.M. Tang** - Research managers

Personal Support

- **Family** - Unwavering support throughout studies
- **Friends** Aryan Ashar, Shang-Jen Wang, Tom Slik - IT Support
- **Sarthak Anand** - Chat template implementation advice

Thank you for making this research possible!

Thank You!

Questions & Discussion

Riccardo Campanella
r.campanella@students.uu.nl
Utrecht University, MSc Artificial Intelligence

Spatial Reasoning in Multimodal LLMs
via CoT Distillation and Monte Carlo Tree Search
for Dutch Facade-Element Detection: An Exploratory Study