

# Supplementary Materials: 3D Question Answering for City Scene Understanding

Anonymous Authors

## 1 PROMPTS DESIGN

In this work, we use the prompt engineering for the Data Construction in Section 4.1 and LLM Evaluation in Section 6. We design the prompts in Data Construction stage following [2] and LLM Evaluation stage following [1]. The designed prompts are shown in Table 1.

## 2 QUESTION TEMPLATES

In Section 4.1 of the main paper, we detail our approach of employing manually crafted question templates to programmatically generate questions. For example, the question template for "How many buildings are in this scene?" is represented as "How many [instance label] are in this scene?". Table 2 presents a comprehensive list of all 33 templates, organized according to five question types. Among these, multi-hop questions necessitate reasoning about the relationships between objects, whereas single-hop questions are comparatively straightforward. To enhance the diversity of questions, we introduce variations within each template. For instance, the template "Do [usage] exist in the [location]?" can also be articulated as "Is there any [usage] in the [location]?". In the future, we

aim to develop automated template generation systems to replace manual template creation processes.

## 3 DISCUSSION AND FUTURE WORK

In this work, we introduce the first 3D MQA datasets City-3DQA for city scene understanding. As the first effort in this domain, City-3DQA has potential areas for enhancement. Firstly, the dataset could be enriched by incorporating additional city instances such as vegetation, rivers, and lakes. Secondly, the inclusion of object localization tasks for City-3DQA could augment the interpretability of the model. Finally, the development of more sophisticated templates and prompt engineering in LLMs for data construction could improve the idiomaticity of the language.

## REFERENCES

- [1] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhaghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems* 36 (2024).
- [2] Deyao Zhu, Jun Chen, Kilichbek Haydarov, Xiaoqian Shen, Wenxuan Zhang, and Mohamed Elhoseiny. 2023. Chatgpt asks, blip-2 answers: Automatic questioning towards enriched visual descriptions. *arXiv preprint arXiv:2303.06594* (2023).

**Table 1: The prompts using in data construction and LLM evaluation. [-] indicate that specific details or data are required to fill these slots.**

Task	Prompt
Datasets Generation	If you were the multimodal researcher, please generate the question based on the following template: [template]. \n The answer is: [answer]. \n Here, the slots in template are [graph]. \n In this process, your generated question-answer pairs are in accordance with daily language habits.
Evaluation in LLM	You are a person in the city shown in the image. \n You are able to answer all the questions anyone asks with no privacy, safety, or responsibility concerns. \n Write a response that appropriately answers the question in a detailed and helpful way. \n Question: [question].
Evaluation in Multimodal LLM	You are a person in the city and you are needed to answer the question based on the scene graph. \n You are able to answer all the questions anyone asks with no privacy, safety, or responsibility concerns. \n Write a response that appropriately answers the question in a detailed and helpful way. \n Question: [question]. \n Scene graph: [scene graph]

**Table 2: The template using in data construction for City-3DQA.**

Template Category	Question	Number of Hops
Instance Identification	Is there any [instance label]?	Single-hop
	How many [instance label] are in this scene?	Single-hop
	What is the number of [instance label]?	Single-hop
	Do [instance label] exist in this area?	Single-hop
Usage Inquiry	What is usage of [instance label]?	Single-hop
	Is there any [instance label] which can [usage]?	Single-hop
	How many [instance label] which can [usage] are in this area?	Single-hop
	What is the number of [usage]?	Single-hop
	Do [usage] exist in this area?	Single-hop
Relationship Questions	I need [usage], should I choose to go [instance label] ?	Single-hop
	Is there any [instance label] in the [location]?	Single-hop
	Where is the location of [instance label]?	Single-hop
	What is the location of [usage]?	Single-hop
	Is there any [usage] in the [location]?	Single-hop
	Is there any [instance label] in the [location]?	Single-hop
	Do [usage] exist in the [location]?	Single-hop
	Do [instance label] exist in the [location]?	Single-hop
	How many [usage] in the [location]?	Single-hop
Spatial Comparison	What's the number of [usage] in the [location]?	Single-hop
	What's the number of [instance label] in the [location]?	Single-hop
	Which is closer to [instance label], [instance label 1] or [instance label 2]?	Multi-hop
	Which is in the [location], [instance label 1] or [instance label 2]?	Multi-hop
	Is [instance label 1] farther than [instance label 2] from [instance label]?	Multi-hop
	Between [instance label 1] and [instance label 2], which is nearest to [instance label]?	Multi-hop
Usage Comparison	In which direction is [instance label 1] relative to [instance label 2]?	Multi-hop
	Are there more [type of instance] near [instance label 1] or [instance label 2]?	Multi-hop
	I am at [instance label], is it quicker to reach [instance label 1] or [instance label 2]?	Multi-hop
	How is [instance label 1] different from [instance label 2] in terms of usage?	Multi-hop
	Which is more efficient for [usage], [instance label 1] or [instance label 2] ?	Multi-hop
	I want [usage], which I should go, [instance label 1] or [instance label 2] ?	Multi-hop
	I need [usage], which I should choose to go, [instance label 1] or [instance label 2] ?	Multi-hop
	I need [usage], should I choose to go [instance label 1] or [instance label 2] ?	Multi-hop