OD-RASE: Ontology-Driven Risk Assessment and Safety Enhancement for Autonomous Driving

Supplementary Material

1. Structuring Infrastructure Improvement Process as Ontology

The 30 types of accident-causing road structures and 26 types of infrastructure improvement proposals defined through expert knowledge include elements that overlap or are time-dependent (e.g., traffic volume, moving vehicles), which fall outside the scope of this research. Moreover, because these elements were derived from a wide variety of real-world traffic-accident cases, they are very granular and therefore not well-suited as a data structure for training our models. For this reason, three experts reached consensus to merge similar elements and exclude those that are time-dependent.

1.1. Consolidation of Similar Elements and Exclusion of Time-Dependent Elements

The accident-causing road structures and countermeasure policies we defined are based on analyses of actual traffic accidents. Consequently, prior studies summarizing conventional infrastructure improvement processes [1, 7, 8, 10, 11, 16, 17] typically classify these elements at a fairly fine-grained level. As a result, multiple similar elements exist, which hinders effective model training. Additionally, since the objective of this study is to analyze and improve risks arising from road infrastructure, any time-dependent factors must be removed. For these reasons, from the initially defined 30 accident-causing road structures and 26 infrastructure improvement proposals, we carried out the merging of similar elements and the exclusion of time-dependent elements. This step was performed by the same experts who created the dataset.

Fig. 1 and Fig. 2 illustrate the processes by which the accident-causing road structures and the infrastructure improvement proposals were merged or excluded, respectively. The merging of similar elements was grouped by road environment. As a result of these procedures, the accident-causing road structures were reduced to 15 types, and the infrastructure improvement proposals were reduced to 12 types.

1.2. Exclusion of Elements Close to Corner Cases

In the Sec. 1.1, we consolidated elements by road environment and removed elements stemming from dynamic factors. However, in real-world road environments, countless corner cases exist. Because our original definitions are quite granular, certain elements ended up disproportionately addressing corner-case situations. In existing datasets as well,

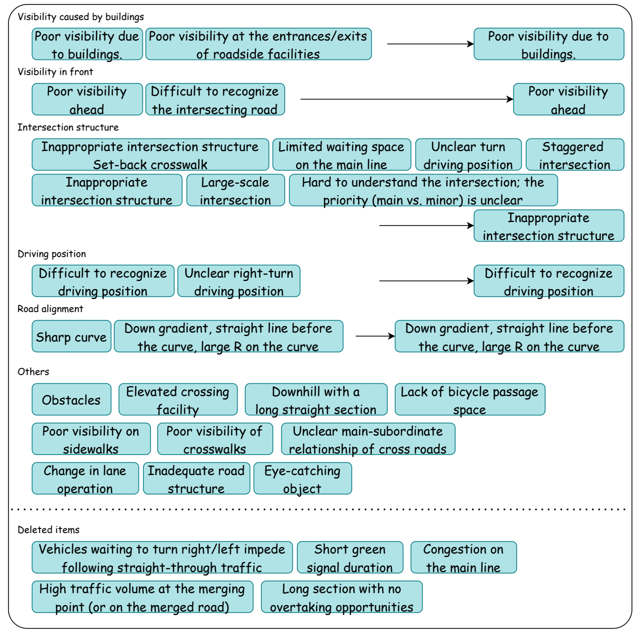

Figure 1. Merging and exclusion of elements in accident-causing road structures.

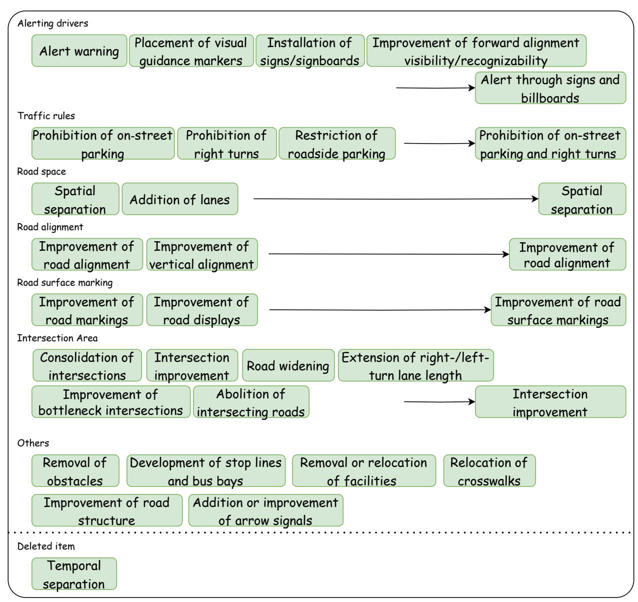

Figure 2. Merging and exclusion of elements in infrastructure improvement proposals.

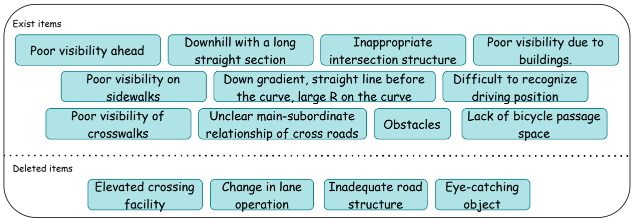

Figure 3. Exclusion of accident-causing road structures related to corner cases.

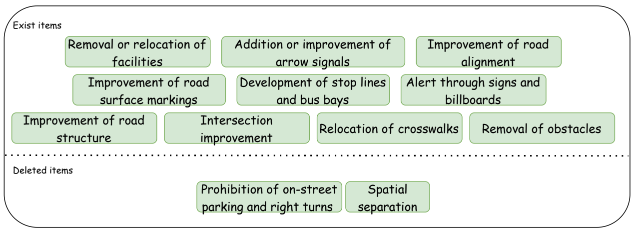

Figure 4. Exclusion of corner-case elements in infrastructure improvement proposals.

these corner-case elements appear so infrequently that they could risk further imbalancing the data. Consequently, using expert knowledge, we carried out an additional exclusion of such corner-case elements from the accident-causing road structures and countermeasure policies that remained after the earlier merging and exclusion step. An overview of this process is shown in Fig. 3 and Fig. 4. As a result, the final accident-causing road structures are reduced to 11 types and the infrastructure improvement proposals to 10 types.

2. G2CoT:Graph-Based Grounded CoT Prompt

The proposed G2CoT uses a carefully designed CoT (chain-of-thought) prompt [19] to mimic the expert reasoning process when drafting infrastructure improvement proposals. Fig. 7 shows the details of our proposed G2CoT. As shown in Fig. 7, it generates outputs in four stages: (1) traffic risks, (2) accident-causing road structures, (3) accident occurrence processes, and (4) infrastructure improvement proposals. Specifically, Step 1 produces a textual explanation of static traffic risks from any given driving-scene image. In Step 2, referencing both the image and the results from Step 1, the model infers the accident-causing factors and selects all the elements that match from the accident-causing road structures we defined in Sec. 1. In Step 3, referencing Steps 1 and 2, it predicts how an accident might unfold. Since infrastructure improvements for the same accident-causing factor can differ depending on the accident occurrence process, Step 3 aids in predicting infrastructure improvements by considering the accident process. Finally, in Step 4, referencing Steps 2 and 3, the model infers the infrastructure

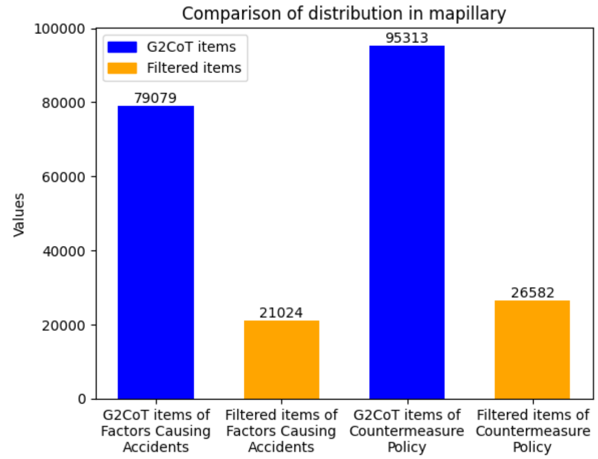

Figure 5. Changes in data distribution before and after filtering for Mapillary Vistas. Blue: before filtering, orange: after filtering.

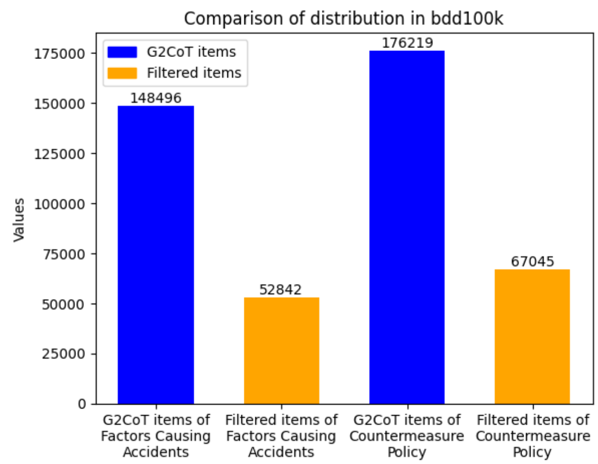

Figure 6. Changes in data distribution before and after filtering for BDD100K. Blue: before filtering, orange: after filtering.

improvement proposals (i.e., countermeasure policies) and selects all applicable elements from those defined in 1. By performing this context-aware inference at each step, we can automatically build a dataset.

2.1. Changes in Data Distribution

We now provide a quantitative comparison of the data distribution before and after applying our expert-knowledge-based filtering on the dataset automatically constructed via G2CoT. We compare the distributions for accident-causing road structures and countermeasure policies. Fig. 5 and Fig. 6 show the changes in data distribution for Mapillary Vistas [14] and BDD100K [20], respectively. From both figures, we see that the output from GPT-4o [2] contains a substantial amount of incorrect data, resulting in more than 50% of generated annotations being discarded by our data

Vision Encoder	Text Encoder	Mapillary				BDD100K			
		Recall	Precision	F1	Acc	Recall	Precision	F1	Acc
ResNet-50[9]	RoBERTa-Base[13]	74.85	82.06	79.19	64.25	86.54	88.55	87.53	73.50
	Flan-T5-xl[5]	72.07	12.50	21.30	0.00	73.94	13.54	22.89	0.00
	Long-CLIP[21]	77.17	76.96	77.06	58.92	83.35	88.74	85.96	71.04
ViT-B[6]	RoBERTa-Base[13]	71.26	78.76	68.05	63.50	88.80	86.44	87.60	72.69
	Flan-T5-xl[5]	29.01	3.76	6.65	0.00	24.25	3.33	5.85	0.00
	Long-CLIP[21]	76.64	80.53	78.54	61.68	85.27	89.35	87.26	73.33
CLIP[15]	RoBERTa-Base[13]	64.08	77.80	70.28	57.00	89.17	85.92	87.51	72.36
	Flan-T5-xl[5]	24.21	4.44	7.50	0.00	22.58	4.28	7.20	0.00
	Long-CLIP[21]	77.10	82.66	79.79	65.06	84.91	90.02	87.39	73.60
Long-CLIP[21]	RoBERTa-Base[13]	64.08	77.80	70.28	57.00	89.10	85.92	87.48	72.30
	Flan-T5-xl[5]	27.80	6.50	10.54	0.00	24.07	5.74	9.27	0.00
	Long-CLIP[21]	76.44	83.89	79.99	65.09	86.23	88.99	87.59	73.96

Table 1. Quantitative evaluation results for predicting accident-causing road structures. The best, second and third best performances are shown in First, Second, Third, respectively.

filtering.

Modal	Precision	Recall	F1	Acc
CLIP	54.98	72.54	62.55	32.60
Long-CLIP	66.34	57.94	61.86	28.22
Ours	76.44	83.89	79.99	65.09

Table 2. Ablation study on grounding block is most effective for predicting infrastructure improvement proposals.

3. Versatility of the proposed method

This research has demonstrated that our OD-RASE model can accurately predict infrastructure improvement proposals for road structures. We have also shown that it can generalize to unknown road structures. In this section, we show that OD-RASE is also capable of predicting accident-causing road structures, not just the infrastructure improvement proposals. We employ the Mapillary Vistas [14] and BDD100K [20] datasets.

3.1. Predicting Accident-Causing Road Structures

We evaluate the ability of OD-RASE to predict road structures that lead to traffic accidents, using supervised learning. Tab. 1 presents quantitative evaluations on the Mapillary and BDD100K datasets. From Tab. 1, it is evident that models using RoBERTa-Base [13] or Long-CLIP [21] as text encoders successfully predict road structures that cause accidents, across multiple vision encoders. By contrast, models using Flan-T5-xl [5] exhibit high recall but low precision, resulting in too many false positives. Overall, the best performance is obtained when both the vision and text encoders are Long-CLIP.

Table 2 shows an ablation study evaluating the impact of the grounding block in OD-RASE. For this experiment, we used Long-CLIP as the vision and text encoder and trained on Mapillary. From Tab. 2, OD-RASE, which includes the grounding block that integrates image and text, outperforms vanilla CLIP or Long-CLIP by a large margin. This confirms that our proposed grounding block is effective.

Data filtering	Precision	Recall	F1	Acc
	38.25	84.33	52.63	1.01
✓	76.44	83.89	79.99	64.09

Table 3. Ablation study on effectiveness of ontology-driven data filtering. Results indicate that filtering improved overall model performance.

3.2. Effectiveness of Dataset Filtering

In this experiment, we used Long-CLIP as the vision and text encoder, using Mapillary as our dataset.

Tab. 3 shows the performance with and without data filtering. When the training data were not filtered, the accuracy on the filtered evaluation set was notably low. Specifically, in the absence of filtering, the model demonstrated a high F1-Score of 52.63 pt but a low Accuracy of 1.01 pt, making it difficult to correctly identify the road structures leading to accidents. In contrast, once data filtering was used, the model achieved an F1-Score of 79.99 pt and an Accuracy of 64.09 pt, indicating a more robust learning outcome. These results strongly support the necessity of our ontology-based data filtering grounded in expert knowledge.

3.3. Zero-shot Prediction

We conduct zero-shot prediction experiments, analogous to the infrastructure improvement proposals task, for accident-causing road structures. Tab. 4 shows results for models trained on BDD100K and evaluated on Mapillary Vistas, and Tab. 5 for those trained on Mapillary Vistas and evaluated

Method		Recall		Precision		F1-Score		Accuracy	
Vision Encoder	Text Encoder	val	test	val	test	val	test	val	test
Ours Baseline									
ResNet-50[9]	Long-CLIP[21]	71.68	74.76	77.28	79.76	74.38	77.18	58.48	61.38
ViT-B[6]	Long-CLIP[21]	76.84	77.77	79.68	80.58	78.24	79.15	62.43	63.83
CLIP[15]	Long-CLIP[21]	74.75	75.63	80.81	82.22	77.66	78.79	61.68	63.25
Long-CLIP[21]	Long-CLIP[21]	75.49	77.10	80.22	81.54	77.78	79.26	62.50	64.11
Generalist Models									
GPT-4o[2]		47.19	51.87	17.86	19.66	25.27	27.81	16.86	18.59
LLaVA-1.5[12]		75.52	75.74	14.71	15.51	22.03	22.93	14.05	14.86
Qwen2-VL[18]		83.42	84.42	21.43	22.51	32.98	34.28	21.07	22.18
Phi-3[3]		52.15	51.35	21.69	21.60	28.94	28.73	18.38	18.24
InternVL2[4]		68.27	72.36	21.07	22.63	30.73	32.74	20.45	21.85

Table 4. Zero-shot prediction of accident-causing road structures. Models are trained on BDD100K and evaluated on Mapillary.

Method		Recall		Precision		F1-Score		Accuracy	
Vision Encoder	Text Encoder	val	test	val	test	val	test	val	test
Ours Baseline									
ResNet-50[9]	Long-CLIP[21]	81.10	81.50	88.37	88.16	84.58	84.70	68.35	68.24
ViT-B[6]	Long-CLIP[21]	75.12	76.64	82.14	83.74	78.47	80.03	63.02	65.00
CLIP[15]	Long-CLIP[21]	81.70	82.04	90.86	90.95	86.04	86.26	71.15	71.66
Long-CLIP[21]	Long-CLIP[21]	83.14	83.72	90.07	89.96	86.46	86.73	71.45	72.02
Generalist Models									
GPT-4o[2]		51.31	48.32	19.92	18.66	27.94	26.21	18.88	17.71
LLaVA-1.5[12]		73.88	75.60	17.34	18.16	25.00	26.21	16.30	17.08
Qwen2-VL[18]		86.31	86.92	24.81	25.08	37.09	37.48	24.36	24.67
Phi-3[3]		41.70	41.27	22.13	22.01	27.48	27.48	17.47	17.48
InternVL2[4]		76.32	76.91	26.68	26.78	37.48	37.65	25.75	25.91

Table 5. Zero-shot prediction of accident-causing road structures. Model was trained on Mapillary and evaluated on BDD100K.

on BDD100K. In both cases, the vision and text encoder pair of Long-CLIP achieves the highest performance, for instance an F1-Score of 79.26 pt and Accuracy of 64.11 pt in Tab. 4. In contrast, using a generalist model like Qwen2-VL [18] yielded an F1-Score of 34.28 pt and an Accuracy of 22.18 pt, indicating that predicting the factors leading to accidents in unknown domains is difficult.

Tab. 5 shows the results for models trained on Mapillary and evaluated on the validation and test sets of BDD100K. Among our baseline variants, the combination of Long-CLIP for both the vision encoder and text encoder yielded an F1-Score of 86.73 pt and an Accuracy of 72.02 pt on the Mapillary test set. Our broad experiments show that generalist models alone struggle to identify road structures that accident-causing road structures or propose meaningful infrastructure improvements.

References

[1] Road infrastructure guidelines: New eu-wide guidelines to assess safety of road infrastructure. European Commission: Mobility & Transport - Road Safety, 2023. 1

[2] GPT-4o system card, Accessed:2024-08-06. 2, 4

[3] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, et al. Phi-3 technical report: A highly capable language model locally on your phone. arXiv preprint arXiv:2404.14219, 2024. 4

[4] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024. 4

[5] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models. Journal of Machine Learning Research, 25(70):1–53, 2024. 3

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner,

G2CoT Prompt

You are an AI responsible for proposing infrastructure improvement plans to reduce traffic accidents. Given images from a car's front camera, you analyze the environment, assess potential traffic risks, and suggest infrastructure enhancements.

Step 1: Explanation of Traffic Risk

Carefully observe the images and focus on the surrounding road structure environment and describe the traffic risk in a minimum of 300 words and a maximum of 400 words.

Step 2: Factors Causing Accidents

Based on the traffic risks identified in Step 1 and surrounding environment, explain the factors that could cause accidents. Also, select all applicable items from the [Factors Causing Accidents] class.

Step 3: Accident Occurrence Process

Using the content from Steps 1 and 2, explain the process by which an accident might occur.

Step 4: Countermeasure Policy

Based on the contents of Steps 2 and 3, propose infrastructure improvement plans to prevent accidents. Also, select all applicable items from the [Countermeasure Policy] class.

Factor Causing Accidents Class

- Poor visibility ahead
- Downhill with a long straight section
- Inappropriate intersection structure
- Poor visibility due to buildings.
- Poor visibility on sidewalks
- Difficult to recognize driving position
- Poor visibility of crosswalks
- Lack of bicycle passage space
- Unclear main-subordinate relationship of cross roads
- Obstacles
- Down gradient, straight line before the curve, large R on the curve

Countermeasure Policy Class

- Removal or relocation of facilities
- Improvement of road alignment
- Improvement of road surface markings
- Alert through signs and billboards
- Development of stop lines and bus bays
- Intersection improvement
- Improvement of road structure
- Removal of obstacles
- Relocation of crosswalks
- Addition or improvement of arrow signals

Figure 7. Details of the G2CoT prompt used when constructing the OD-RASE Dataset.

Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In International Conference on Learning Representations (ICLR), 2021. 3, 4

[7] Wenlu Du, Ankan Dash, Jing Li, Hua Wei, and Guiling Wang. Safety in traffic management systems: A comprehensive survey. Designs, 7(4), 2023. 1

[8] Dorin-Ion Dumitrascu. Influence of road infrastructure design over the traffic accidents: A simulated case study. Infrastructures, 9(9), 2024. 1

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and

Pattern Recognition (CVPR), 2016. 3, 4

[10] Maher Holozadah and Sahar Tawfiq. Road safety audit for the intersection of cedar road and brainard road and i-271 interchange with brainard road and cedar road. Northeast Ohio Areawide Coordinating Agency, 2013. 1

[11] Sam Thompson David Phipps Carlos Moya Shawn A. Troy Mary Bea Kolbe Christopher Oliver Phillip Vereen Dom Ciaramitaro Joe Seymour Tim Williams Kimberly Hinton Jason Schronce Brian Wert Hanna Cockburn Kristina Solberg Ed Johnson Julie Bogle, Sarah Lee and Jimmy Travis. Action plan for implementing pedestrian crossing countermeasures at uncontrolled locations. North Carolina Department of Transportation, 2018. 1

[12] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. Advances in Neural Information Processing Systems (NeurIPS), 2023. 4

[13] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. In Proceedings of the 20th Chinese National Conference on Computational Linguistics, 2020. 3

[14] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2017. 2, 3

[15] Alec Radford, Jong Wook Kim, Chris Hallacy, A. 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 Proceedings of the 38th International Conference on Machine Learning (ICML), 2025. 3, 4

[16] R.A. Retting, B.N. Persaud, P.E. Garder, and D. Lord. Crash and injury reduction following installation of roundabouts in the United States. American Journal of Public Health, 91(4): 628–631, 2001. 1

[17] Richard A. Retting, Allan F. Williams, David F. Preusser, and Helen B. Weinstein. Classifying urban crashes for countermeasure development. Accident Analysis Prevention, 27(3): 283–294, 1995. 1

[18] 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. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191, 2024. 4

[19] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Huai hsin Chi, F. Xia, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems (NeurIPS), 2022. 2

[20] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020. 2, 3

[21] Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-clip: Unlocking the long-text capability of clip. In European Conference on Computer Vision (ECCV), 2024. 3, 4