CityLLaVA: Efficient Fine-Tuning for VLMs in City Scenario

Zhizhao Duan* Hao Cheng* Duo Xu Xi Wu Xiangxie Zhang Xi Ye Zhen Xie†
Alibaba Group
{zhizhao.dzz,luchen.ch,manii.xd,lingke.wx}@alibaba-inc.com
{zhangxiangxie.zxx,yx150449,xiezhen.xz}@alibaba-inc.com

Abstract

In the vast and dynamic landscape of urban settings, Traffic Safety Description and Analysis plays a pivotal role in applications ranging from insurance inspection to accident prevention. This paper introduces CityLLaVA, a novel fine-tuning framework for Visual Language Models (VLMs) designed for urban scenarios. CityLLaVA enhances model comprehension and prediction accuracy through (1) employing bounding boxes for optimal visual data preprocessing, including video best-view selection and visual prompt engineering during both training and testing phases; (2) constructing concise Question-Answer sequences and designing textual prompts to refine instruction comprehension; (3) implementing block expansion to fine-tune large VLMs efficiently; and (4) advancing prediction accuracy via a unique sequential questioning-based prediction augmentation. Demonstrating top-tier performance, our method achieved a benchmark score of 33.4308, securing the leading position on the leaderboard. The code will be released soon.

1. Introduction

With the rapid advancement of large language models (LLMs), an increasing number of fields are beginning to explore the capabilities of these models, investigating their potential impact on industry standards and societal practices. Particularly in research areas that straddle computer vision (CV) and natural language processing (NLP), such as traffic video analysis, these models have not only significantly raised the bar for automated analysis precision but have also unlocked unprecedented vistas of application. There are myriad foundational visual-language models (VLMs) such as GPT4-V [20], Qwen-VL-Chat [2], LLaVA [14] and others that stand as testaments to the synergetic potential of CV and NLP. While these large models exhibit formidable capabilities across a range of tasks, they often fall short of expectations when applied directly to highly specialized domains, such as traffic safety scenario captioning. It becomes evident that these models require essential fine-tuning to adequately capture the domain-specific nuances.

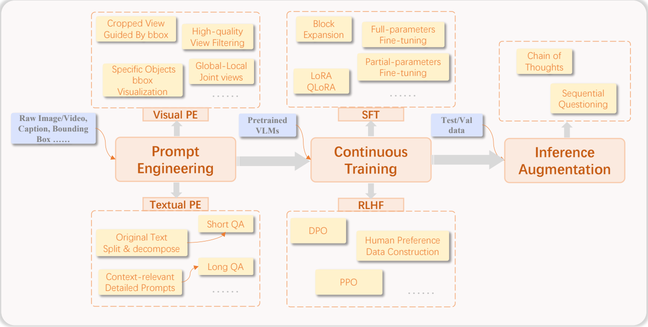

Cropped View Guided By bbox
High-quality View Filtering
Specific Objects bbox Visualization
Global-Local Joint views
......
Visual PE
Raw Image/Video, Caption, Bounding Box ......
Prompt Engineering
Textual PE
Short QA
Original Text Split & decompose
Long QA
Context-relevant Detailed Prompts
......
Block Expansion
Full-parameters Fine-tuning
Partial-parameters Fine-tuning
LoRA QLoRA
......
SFT
Pretrained VLMs
Continuous Training
RLHF
DPO
Human Preference Data Construction
PPO
......
Test/Val data
Chain of Thoughts
Sequential Questioning
Inference Augmentation

Figure 1. The efficient fine-tuning paradigm for VLMs. The efficient fine-tuning paradigm for VLMs. The paradigm first executes the prompt engineering, which includes visual and textual prompt engineering. Then continuous training including SFT and RLHF is implemented based on the pretrained VLMs. Finally, the Inference augmentation is used to improve performance.

Fine-tuning a large model to meet the granular needs of a particular application involves navigating a complex landscape of challenges, such as the construction of effective question-answer pairs, the design of appropriate prompts, and the selection of critical fine-tuning parameters. Furthermore, creating annotations that capture the multifaceted nature of real-world events can be labor-intensive.

In light of these challenges, this work aims to introduce an effective and comprehensive paradigm for fine-tuning large visual-language models, including prompt engineering, continuous training, and inference augmentation. Figure 1 shows the details of this effective paradigm, which is accumulated in our industry practice. To demonstrate the effectiveness of the proposed approach, we conduct detailed experiments on the WTS [9] dataset. Specifically, the proposed paradigm has achieved first place in the 2024 AI City

[*] Authors contributed equally to this work.
[†] Corresponding authors.

Challenge [28] - Traffic Safety Description and Analysis, contributing valuable insights and offering a pathway for future research to enhance or adapt LLMs for similar complex, domain-specific tasks.

In sum, the contributions of this paper are summarized as follows:

- Proposing an effective and comprehensive paradigm for fine-tuning large visual-language models for domain-specific tasks.
- Exploring visual and textual prompt engineering to construct informative and refined inputs for both training and inference.
- Investigating the adaptation of block expansion in VLMs, achieving superior performance compared to LoRA [8].
- Achieving state-of-the-art performance on the WTS dataset, with a detailed exploration of factors influencing fine-tuning efficacy.

2. Related Work

2.1. Vision-Language Models

Vision-language models (VLMs) utilize image and text data simultaneously and fuse knowledge in different domains for better performance. CLIP [22] is a pioneering work that aligns language and image by designing a pretraining task that matches images with text captions. It shows spectacular zero-shot transferability among several downstream tasks. In recent years, with the development of large language models [2, 21, 27], combining a visual encoder with an auto-regressive language decoder has become a prevalent approach in vision-language tasks. This type of method can benefit from both visual perception and linguistic expression and a more versatile model can be realized. An early study in this area is Flamingo [1], which leverages gated cross-attention to accept interleaved visual and language data as input and then generates text as output. BLIP-2 [10] introduces a lightweight but powerful module named Q-former to efficiently bridge the modality gap between image and text, while FlanT5 [6] is used as the language model. Built on the pretrained visual component of BLIP-2, MiniGPT4 [35] employs a single projection layer to align the visual features with text features and input to the Vicuna [5] language model. An improved version is MiniGPT-v2 [4], which applies a simpler strategy that directly projects the visual tokens from a ViT [7] encoder to the feature space of a large language decoder. LLaVA [14] adopts a similar model structure of utilizing a projection layer after the encoded visual features. With the proposed two-stage training strategy, LLaVA demonstrates impressive abilities in vision-language tasks and there are many following works based on it [11, 13, 15, 17, 26].

2.2. VLMs in Driving

Many researchers have attempted to apply vision-language models in driving since they have shown remarkable capabilities in visual signal perception and language understanding. A previous study [29] conducts an exhaustive evaluation on the state-of-the-art vision-language model GPT4-V [20] in the autonomous driving scenario and the experiment results illustrate superior performance. Dolphins [18], a novel vision-language model in which the pretrained OpenFlamingo [1] serves as the fundamental structure, demonstrates distinctive behaviors in the driving domain. DriveGPT4 [31] can process textual queries and multi-frame videos as the input and generate corresponding responses, while it is also capable of predicting low-level vehicle control actions and signals. Experiment results suggest that DriveGPT4 has comparable or even better ability in some cases compared with GPT4-V.

3. Methodology

3.1. Overview

CityLLaVA introduces an efficient fine-tuning pipeline aimed at enhancing spatial-temporal understanding and providing fine-grained perception within urban environments. As shown in Figure 2, the proposed paradigm consists of three major modules: visual prompt engineering, textual prompt engineering (i.e., text QA construction), and efficient fine-tuning for the large vision-language model. We will sequentially introduce the details of each module.

3.2. Dataset Construction

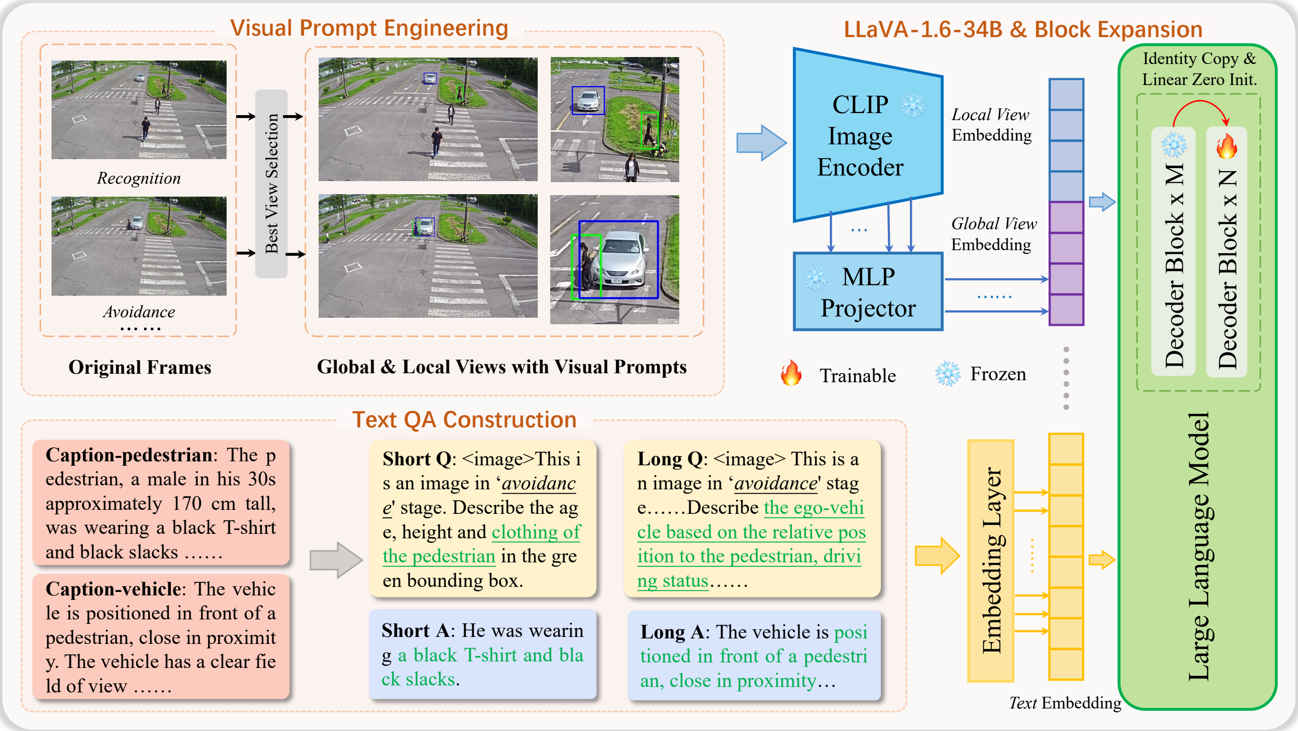
In this section, we elaborate on the details of data filtering and the construction of a vision-language instruction-tuning dataset for Supervised Fine-tuning (SFT). For the fine-tuning, an item $\mathbf{H}$ for training can be formulated as a tuple:

$$\mathbf{H} = (\mathbf{X_v}, \mathbf{X_q}, \mathbf{X_t}) \tag{1}$$

where $\mathbf{X_v}, \mathbf{X_q}, \mathbf{X_t}$ denote the visual inputs, textual instruction and textual response, respectively. The best construction of them is explored in the subsequent sections.

3.2.1 Bounding-box Guided View Selection

The WTS dataset consists of two parts: a. WTS data, which is a multi-view dataset with an uncertain number of vehicle views and overhead views; b. Filtered pedestrian-centric videos from BDD100K [32] data with vehicle view only. Given that the relevant vehicles and pedestrians might not be clearly visible or could be insignificant in size within some views of the WTS data, directly fine-tuning Visual Language Models (VLMs) on multi-perspective data poses a challenge. To address this, we introduce bounding-box



Caption-pedestrian: The pedestrian, a male in his 30s approximately 170 cm tall, was wearing a black T-shirt and black slacks ……

Caption-vehicle: The vehicle is positioned in front of a pedestrian, close in proximity. The vehicle has a clear field of view ……

Short Q: <image>This is an image in '*avoidance*' stage. Describe the age, height and clothing of the pedestrian in the green bounding box.

Short A: He was wearing a black T-shirt and black slacks.

Long Q: <image> This is an image in '*avoidance*' stage……Describe the ego-vehicle based on the relative position to the pedestrian, driving status……

Long A: The vehicle is positioned in front of a pedestrian, close in proximity…

Figure 2. The overview of CityLLaVA. Our method is anchored on the pretrained LLaVA-1.6-34B [15] equipped with block expansion [30], combining the textual prompt engineering and visual prompt engineering guided by bounding boxes.

guided view selection. Initially, we filter out overhead views from the WTS data that do not match the officially recommended perspectives. The data is then segmented into tuples $\mathbf{S} = \{\mathbf{V}, \mathbf{T}, \mathbf{B}\}$, representing the video clips, descriptions, and bounding boxes, following the segment set forth by the authorities. For each tuple $\mathbf{S}_i$ in the training dataset, we compute the average vehicle area $\mathbf{A_v}_i$ and the average pedestrian area $\mathbf{A_p}_i$ based on the bounding boxes $\mathbf{B}_i$. We consider the area to be zero if a bounding box is not present. We then apply the following criteria to obtain the filtered tuple data $\hat{\mathbf{S}}_i$, ensuring that only relevant views are selected for model training:

$$\hat{\mathbf{S}}_i = \begin{cases} \text{False}, & \text{if } A_{p_i} > thr_p \text{ and } A_{v_i} > thr_v, \\ \text{False}, & \text{if } A_{p_i} > thr_p \text{ and } A_{v_i} = 0, \\ \text{False}, & \text{if } A_{v_i} > thr_v \text{ and } A_{p_i} = 0, \\ \text{True}, & \text{otherwise}. \end{cases} \quad (2)$$

where thr_p and thr_v denotes the area threshold of pedestrian and vehicle.

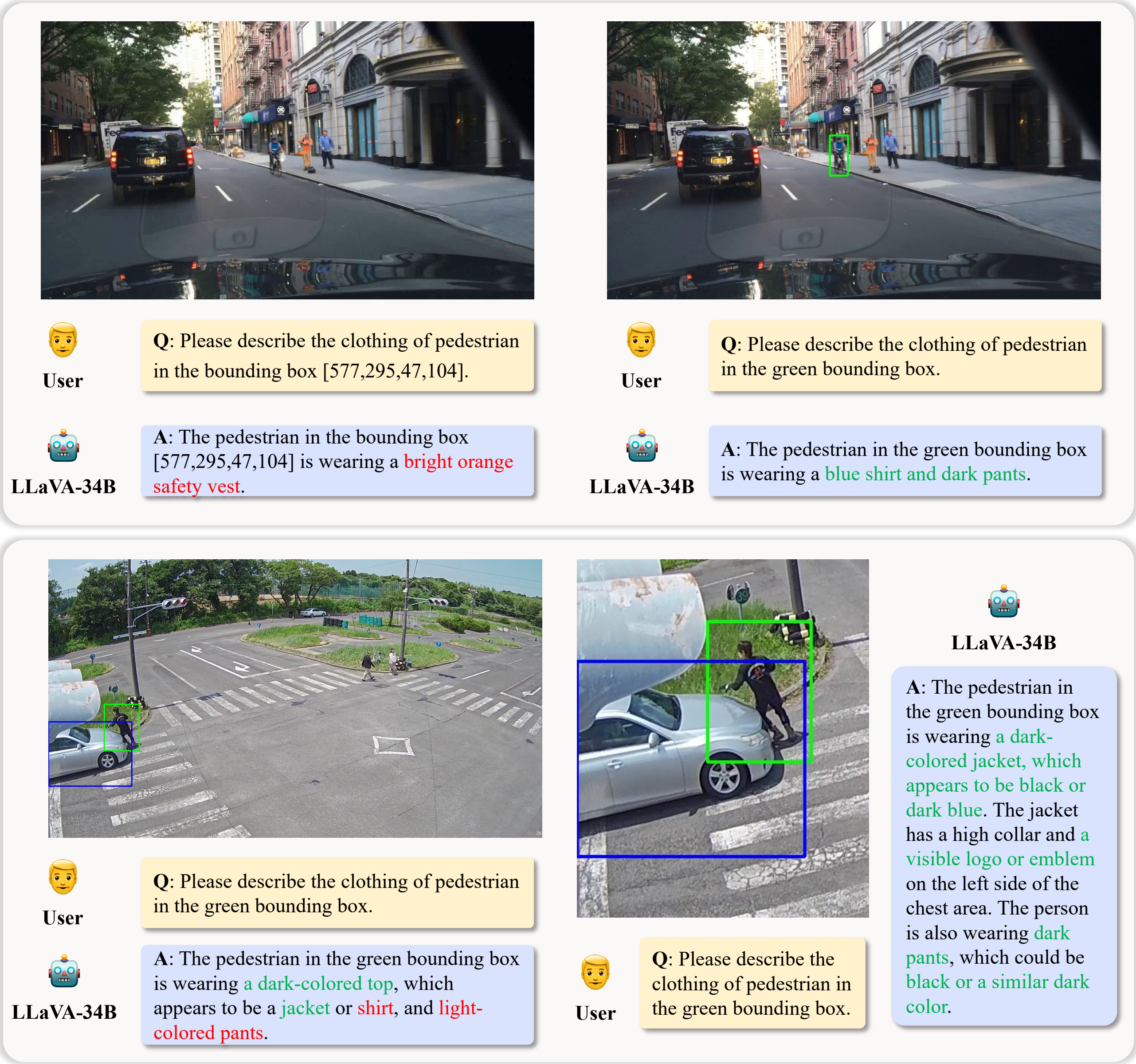
For the `test` dataset, we initially group the tuples $\mathbf{S}$ by scenario. Within a scenario, we select the video that not only features bounding boxes across five stages but also showcases the largest average pedestrian area $\mathbf{A_p}$ as the optimal view for testing purposes. This process ensures the selection of the most suitable perspective for inference.

3.2.2 Visual Prompt Engineering

In this section, the visual prompt engineering (VPE) is elaborated from two aspects.

Visual Prompt. A visual prompt introduces an innovative approach to prompt engineering. VLMs are inherently multimodal and provide the chance to manipulate both visual and textual modality. A simple red circle in pixel space can direct CLIP's attention to the interested region improving performance in zero-shot referring expressions comprehension [25]. ViP-LLava [3] also leverages a red bounding box or point arrow to enhance the region-level perception of VLM. In this paper, we introduce bounding box rectangles as visual prompts to localize the interested pedestrian and vehicle, for the sake of fine-grained vision-language alignment and information extraction of local region.

We implement visual prompts by drawing the scaled-up bounding boxes on the corresponding video frame. Given a frame $\mathbf{I}$, and its pedestrian bounding box $\mathbf{B}_p$, vehicle bounding box $\mathbf{B}_v$, we visualize the bounding box with green rectangle for the pedestrian and blue rectangle for the vehicle. It is critical to scale up the bounding box to ensure it encompasses the entire region of interest around the pedestrian or vehicle, as the original bounding box might cover only a partial area. Given a bounding box $\mathbf{B}$, the



Q: Please describe the clothing of pedestrian in the bounding box [577,295,47,104].

A: The pedestrian in the bounding box [577,295,47,104] is wearing a bright orange safety vest.

Q: Please describe the clothing of pedestrian in the green bounding box.

A: The pedestrian in the green bounding box is wearing a blue shirt and dark pants.

Q: Please describe the clothing of pedestrian in the green bounding box.

A: The pedestrian in the green bounding box is wearing a dark-colored top, which appears to be a jacket or shirt, and light-colored pants.

Q: Please describe the clothing of pedestrian in the green bounding box.

A: The pedestrian in the green bounding box is wearing a dark-colored jacket, which appears to be black or dark blue. The jacket has a high collar and a visible logo or emblem on the left side of the chest area. The person is also wearing dark pants, which could be black or a similar dark color.

Figure 3. Examples of usages of visual prompt (Top) and cropped view guided by bounding boxes (Bottom).

scaled-up $\hat{\mathbf{B}}$ can be formulated as the follow:

$$\hat{\mathbf{B}} = \text{Scale}(\mathbf{B}, c) = (x+\frac{w}{2}-\frac{cw}{2}, y+\frac{h}{2}-\frac{ch}{2}, cw, ch) \quad (3)$$

where c denotes the scaling coefficient, x and y represent the coordinates of the top-left corner, and w and h represent the width and height of the box, respectively. The default c is set to 1.2. Based on the scaled-up bounding boxes $\hat{\mathbf{B}}_p = (\hat{x}_p, \hat{y}_p, \hat{w}_p, \hat{h}_p)$ and $\hat{\mathbf{B}}_v = (\hat{x}_v, \hat{y}_v, \hat{w}_v, \hat{h}_v)$, we can generate the augmented frame $\hat{\mathbf{I}}$ with visual prompts, which is then fed into VLM.

Global-Local Joint Views. To enhance the understanding of both global contexts and specific interests in pedestrians and vehicles, we concatenate the augmented frame $\hat{\mathbf{I}}$ and its cropped view guided by bounding boxes $\hat{\mathbf{I}}_m$ as joint visual inputs. Prior research [12, 15] suggests that using a concatenation of multi-cropped and full-view images as visual inputs can enhance the performance on fine-grained multimodal understanding and reduce the hallucination in outputs. However, this manner brings a greater computational burden. Furthermore, the visual redundancy introduced in this procedure may yield fewer improvements, even decrements for certain tasks. As a result, We replace multi-cropped views in [15] with local cropped views guided by bounding boxes. This replacement directs the model's attention to the key region while reducing unnecessary visual information. Given bounding boxes $\hat{\mathbf{B}}_p$ and $\hat{\mathbf{B}}_v$, the cropped view boundary can be formulated as:

$$\begin{aligned}\mathbf{V}_m &= \text{Crop}(\hat{\mathbf{B}}_p, \hat{\mathbf{B}}_v) \\ &= (x_{\min}, y_{\min}, x_{\max} - x_{\min}, y_{\max} - y_{\min}) \\ x_{\min} &= \min(\hat{x}_p, \hat{x}_v) \\ y_{\min} &= \min(\hat{y}_p, \hat{y}_v) \\ x_{\max} &= \max(\hat{x}_p + \hat{w}_p, \hat{x}_v + \hat{w}_v) \\ y_{\max} &= \max(\hat{y}_p + \hat{h}_p, \hat{y}_v + \hat{h}_v)\end{aligned} \tag{4}$$

where $\mathbf{V}_c$ denotes the boundary of the cropped view, which can be interpreted as the smallest external rectangle of two bounding boxes. We can use Scale$(\cdot, \cdot)$ defined in Eq. 3 to scale up the boundary for more context attributes:

$$\hat{\mathbf{V}}_m = \text{Scale}(\mathbf{V}_m, c^*) \tag{5}$$

Here c^* is set to 1.5. The final cropped view $\hat{\mathbf{I}}_m$ can be formulated as:

$$\hat{\mathbf{I}}_m = \hat{\mathbf{I}}[:, \hat{x}_m : \hat{x}_m + \hat{w}_m, \hat{y}_m : \hat{y}_m + \hat{h}_m] \tag{6}$$

where $\hat{x}_m, \hat{y}_m$ represent the coordinates of top-left corner of $\hat{\mathbf{V}}_m$, and $\hat{w}_m, \hat{h}_m$ denote the corresponding width and height. Both $\hat{\mathbf{I}}_m$ and $\hat{\mathbf{I}}$ are fed into the vision encoder during training and inference.

Effectiveness of VPE. We conduct a simple experiment to verify the effectiveness of our visual prompt engineering. We use the textual prompt "*Please describe the clothing of pedestrian* {*in the bounding box [x, y, h, w]*}/{*in the green bounding box*}" to query the original LLaVA-1.6-34B [15] under the original frames and augmented or cropped frames, and compare the accuracy of outputs in different settings. As shown in Figure 3, the proposed paradigm enhances the fine-grained perception in specific objects and interested regions. The sub-figure on the top illustrates that the drawn green rectangle directs the model's attention to the interested pedestrian who is more likely to be collided with the ego-vehicle. These simple visual prompts prevent the model from being disturbed by irrelevant objects, building up the fine-grained visual-language alignment. The sub-figure on the bottom demonstrates the improvement in detail recognition during inference with a cropped view guided by bounding boxes. The experiments conducted indicate that datasets processed with the proposed visual prompt engineering exhibit enhanced alignment, which is conducive to the fine-tuning of the model.

3.2.3 Textual Prompt Engineering

A well-constructed prompt is essential for visual question answering and visual captioning, especially when detail-intensive descriptions are required. Through prompt engineering, we aspire to identify a question that not only encapsulates the content of the description accurately but also comprehensively, rather than just inputting "Please provide a detailed description of the pedestrian/vehicle in the video" into the model.

By conducting a dimensional analysis of the descriptions, we distill the key points such as height, clothing, line of sight, relative position, movement, and environment. We engage GPT-4v [20] to evaluate the alignment of the generated responses with ground truth, recognizing mismatches and areas for enhancement. The result is a set of prompts that are meticulously crafted to guide the model towards generating high-quality, context-relevant answers.

Prompt for pedestrian descriptions. This picture shows the relationship between the pedestrian in the green box and the vehicle in the blue box. Describe the pedestrian in the green box or the pedestrian closest to the vehicle based on age, height, clothing, line of sight, relative position to the vehicle, movement status, weather conditions, and road environment.

Prompt for vehicle descriptions. This picture shows the relationship between the vehicle in the blue box and the pedestrian in the green box. Describe the vehicle in the blue box or the vehicle closest to the pedestrian based on the relative position to the pedestrian, driving status, weather conditions, and road environment. And describe the age, height, and clothing of the pedestrian.

For the vehicle perspective, we utilize "ego-vehicle" in place of "the vehicle in the blue box" to enhance the contextual relevance and specificity of the prompt.

3.2.4 Short QA Construction

Verbose descriptions often hinder a model's alignment with pertinent content (e.g., *localization*, *attention*, *context attributes*). To tackle this issue, we have introduced a series of short question-answer (QA) pairs derived from the source descriptions to enhance dataset diversity. This approach aims to reduce model output style and template over-fitting during the fine-tuning process. By splitting each description into specific dimensions - attributes, location, motion state, and environment - we can construct targeted questions that elicit detailed and relevant responses from the model.

To ensure the preservation of the context and structure with the generated data, we develop a description splitting method that utilizes GPT-4 [19] to categorize each sentence of the descriptions into predefined dimensions. The sentences within the same dimension are concatenated to form a cohesive segment. This segment is then paired with the corresponding query for that dimension to construct a tailored question-answer pair. The prompt used for this process is as follows: "Please select the most appropriate label for each descriptive text from the following options, and format the output by providing the text index followed by the letter a, b, c, d, or e. Each selection should be on a new line."

Finally, we construct an image-text dataset including long QA and short QA pairs. Note that there are two different compositions of textual parts: (1) *Multi-round QA*, a multi-round conversation including both long QA defined in Sec. 3.2.3 and short QA pairs. (2) *Single-round QA*, a single-round conversation including just a QA pair. We compare the influence of these two manners in Sec. 4.4

3.3. Model Architecture

Initially, Qwen-VL-Chat [2]and Video-LLaVA [11] are selected as the candidate baseline model for this video understanding task due to their relatively good performance. Both of these models extract 8 frames uniformly from each stage video clip **V** as input during the data processing. Concerning the fine-tuning approach, inspired by LLaMA-Pro [30], we use the block expansion instead of LoRA to fine-tune the baseline model. In the block-expansion method, some zero linear Initialized decoder block layers, which are identity copied from the LLM module of VLM, are interleaved into the LLM backbone. During fine-tuning, only the parameters of these duplicate block layers are unfrozen. This method demonstrated enhanced learning capabilities, resulting in improved performance indicators. A detailed comparison between LoRA and block expansion can be found in Table 5. The result reveals that Qwen-VL-Chat significantly outperforms Video-LLaVA.

However, it was observed that many stage video clips contain fewer than 8 frames, sometimes only one. It implies that we can deal with this task with an image model. Considering the first frame of each video clip **V** is manually annotated, and the annotation quality surpassed that of the remnant tracking data, we deploy the LLaVA-1.6-34B [15], which is the state-of-the-art VLM. We only use the first frame of each video clip **V** as the input and also apply the block expansion for efficient fine-tuning. Table 1 shows the number of model's parameter with block-expansion. We find that despite the LLaVA-1.6-34B with single frame input losing some temporal information possibly, its larger parameter size aided in a more fine-grained understanding of the image.

Furthermore, we attempted to enhance the model's performance by implementing Reinforcement Learning from Human Feedback (RLHF). Previous work [26] shows that LLaVA can benefit from RLHF by reducing hallucinations. However, the original RLHF is based on the PPO algorithm [24], which consumes large computational resources. We choose the Direct Preference Optimization (DPO) algorithm [23] as an alternative since it is more computationally efficient. Following previous work [33, 34], our target is using DPO to alleviate hallucinations and improve performance. We leverage the ground truth of the phase descriptions as positive samples and the outputs from the SFT model as negative samples, then feed them into the DPO. Unfortunately, DPO leads to a decline in performance indicators. Upon analysis, we find two possible reasons for the degraded performance. Firstly, the description in this task is relatively longer than the captions or responses in other datasets, making DPO harder to align. Secondly, various annotation templates have been applied in the dataset, which might confuse DPO such that the model does not know which template needs to be aligned. Therefore, we remove DPO training from this challenge.

Model	Blocks	Dim	Heads	Parameters
Qwen-VL-Chat	32	4096	32	7B
Video-LLaVA	32	4096	32	7B
LLaVA-34B	60	7168	56	34B
Qwen-VL-Chat+BE	40	4096	32	9B
Video-LLaVA+BE	40	4096	32	9B
LLaVA-34B+BE	72	7168	56	41B

Table 1. the model's parameter w/o block-expansion. **BE** refers to the block expansion [30].

3.4. Harnessing Sequential Questioning

We have investigated the impact of sequential questioning on the performance of the model trained exclusively on single-round QA instances. Despite the absence of multi-round QA pairs in the training dataset (i.e., the items in `train` set have no sequential question-answer pairs containing both of the "*pedestrian*" prompts and "*vehicle*" prompts defined in Sec. 3.2.3 at the same time), our findings reveal an improvement in response accuracy when the model is subjected to a series of questions in a specific order during inference.

A general phenomenon has been discovered that "*vehicle*" scores are higher than "*pedestrian*" scores on average. This comparison indicates that the model's output for the "*vehicle*" prompt contains a more precise description of context attributes, localization, and attention. Therefore, a reasonable approach is to insert the "*vehicle*" description into the "*pedestrian*" prompt providing enhanced contexts for a more precise "*pedestrian*" description. The above strategy can be summarized into "Firstly asking *vehicle*, then asking *pedestrian*."

By simply leveraging the sequential questioning, the model conducts outputs with higher evaluation metrics, which can be regarded as a prediction augmentation. The detailed experimental results for the sequence of the questions are analyzed in Table 7. Note that only a specific order of questions can improve model performance.

Rank	Team ID	Team Name	Score
1	208	**AliOpenTrek (ours)**	**33.4308**
2	28	AIO_ISC	32.8877
3	68	Lighthouse	32.3006
4	87	VAI	32.2778
5	184	Santa Claude	29.7838
6	34	LTDT	29.4070

Table 2. **Leaderboard of Traffic Safety Description and Analysis**. Our method ranks first place in the 2024 AI City Challenge Track 2.

4. Experiments

In this section, we explain the datasets and metrics firstly. Subsequently, we introduce our implementation details. We also provide the results on the ablation study.

4.1. Dataset

This paper uses the WTS dataset for the model training and evaluation. WTS is the largest dataset for spatial-temporal fine-grained video understanding in the traffic domain, aiming at describing detailed behaviors of both vehicles and pedestrians within a variety of staged traffic events including accidents. WTS features over 1,200 video events from more than 130 distinct traffic scenarios, combining perspectives from both ego-vehicle and fixed overhead cameras within a vehicle-infrastructure cooperative environment. It offers detailed textual descriptions for each event, covering observed behaviors and contexts. Additionally, for broader research applications, detailed textual annotations are also available for 4,861 publicly accessible pedestrian-centric traffic videos from BDD100K.

Because of the large number of `val` set in the WTS dataset and the large computational resources usage, we conduct the ablation experiments on a selected subset of the original `val` set, which contains 82 samples, a total 301 entries.

4.2. Evaluation Metrics

For CityLLaVA, we use BLEU-4, METEOR, ROUGE-L, and CIDEr as evaluation indicators to compare the predicted descriptions against the ground truth. More specifically, these 4 metrics are used to calculate a final score:

$$\text{Score} = \frac{\text{BLEU-4} + \text{METEOR} + \text{ROUGE-L} + 0.1 \times \text{CIDEr}}{4} \times 100 \tag{7}$$

4.3. Implementation Details

Training. We use the pretrained LLaVA-1.6-34B as our backbone. The whole reproducible training process is only composed of the SFT stage. During the SFT stage, we focus on training the aided blocks with other parts freezing. Table 3 shows the hyperparameters we use to finetune the CityLLaVA model. For CityLLaVA, the whole SFT phase takes 7.8 hours utilizing NVIDIA 8×A100-80G GPUs.

Hyperparameters	Value
LR	2e-4
Epoch	1
BatchSize	64
MaxLength	2048
ZeRO3	True

Table 3. The Hyperparameters for CityLLaVA.

Precision	Memory (GB)
Float16	78.2
INT8	41.8
INT4	22.6

Table 4. GPU resource usage during inference. The statistic is from the single GPU.

Model	frames	BLEU-4	METEOR	ROUGE-L	CIDEr	Score
Qwen-VL-Chat+BE	8	0.243	0.451	0.439	0.692	30.03
VideoLLaVA+BE	8	0.221	0.419	0.426	0.867	28.81
LLaVA-34B+LoRA	1	0.263	0.464	0.455	1.039	32.15
LLaVA-34B+BE	1	**0.278**	**0.477**	**0.470**	**1.130**	**33.43**

Table 5. The performance on different backbone and SFT method on `test` set.

	BLEU-4	METEOR	ROUGE-L	CIDEr	Score
	View Combination (text part under single-round QA)				
Global Only	0.275	0.471	0.464	0.997	32.72
Local Only	**0.289**	**0.484**	**0.481**	1.044	33.91
Global + Local	0.287	0.483	0.477	**1.186**	**34.12**
	Training QA manner (vision part under local view only)				
Multi-round QA	0.252	0.452	0.442	0.928	30.94
Single-round QA	**0.289**	**0.484**	**0.481**	**1.044**	**33.91**

Table 6. **The performance on different visual inputs and text inputs on `val` set**. Note that the ablation study concerning **View Combination** is carried out with text inputs in a single-round QA format, while the ablation study about the **Training QA manner** is executed with visual inputs from only a locally cropped view. *Global/Local Only* refers to the performance of the model that is solely trained on global/local views. *Single-round* and *multi-round QA* refer to the definition at Sec. 3.2.4

Inference. We use the prompts defined in Sec. 3.2.3 to query the model for captions of the pedestrian and vehicle. We implement INT4 quantization for LLaVA-1.6-34B during inference. Model quantization can significantly reduce the GPU memory usage without obvious performance degradation. The statistics of GPU resource usage during inference are summarized in Table 4. All evaluations are executed on NVIDIA 8×A100-80G GPUs. The time required to evaluate the `test` set is approximately 1.7 hours. Our method takes first place in 2024 AI City Challenge Track 2 with a score of 33.4308, as shown in Table 2.

4.4. Ablation Study

We perform important ablation experiments to validate the effectiveness of the proposed modules. Note that the evaluation of LLaVA-1.6-34B requires large computational resources, we can hardly conduct the complete experiments over all different settings, and some experimental results are yielded on the `test` set rather than selected `val` set.

Besides the ablation experiments about backbones and SFT methods, the other experiments are performed in LLaVA-1.6-34B equipped with block expansion.

Effects of backbone and SFT method. Table 5 shows the results of different backbones and SFT methods in the test set. For Qwen-VL-Chat, we simply concatenate 8 frame images as the inputs for the model training and inference. Beyond our expectations, Video-LLaVA yields the poorest score even though it was the only base model we chose that was trained with video data during both the pretraining and instruction tuning stages. TempCompass [16] also finds that most open-source video LLMs do hardly understand videos. We believe that Qwen-VL-Chat outperforms Video-LLaVa because Qwen-VL-Chat has been pretrained in more diverse and abundant data. In addition, it can be found that the increasing number of parameters can achieve better results. For LoRA and block expansion, we set the LoRA parameter of LLaVA-1.6-34b to r=256 and α=512. Experimental results show that the block expansion performs better than LoRA.

Effects of Prompt Engineering. Table 6 shows the effects of the proposed visual and textual prompt engineering. For the ablation study about view combination, the model based on global and local joint views outperforms models that rely solely on either global or local cues. Note that models with *Global only* and *Local only* are trained in **AnyRes** [15] manner while yielding the suboptimal results. This comparison verifies the perspective proposed in Sec. 3.2.2, unsupervised multi-cropped views contain much visual redundancy disturbing the model's attention. The local cropped view, guided by bounding boxes, provides the model with precise visual information contributing to the generation of high quality.

The comparison between *Multi-round QA* and *Single-round QA* reveals that the latter format contributes to increasing the diversity of the existing dataset, an aspect that is advantageous for model training. As shown in Figure 4, the training loss for *Multi-round QA* converges to approximately 0.30, whereas for *Single-round QA*, it converges to around 0.45. Comprehensively considering the evaluation results and loss values, we can infer that the dataset consisting of single-round QA pairs alleviates the over-fitting during model training.

Effects of Sequential Questioning. Table 7 shows the effects of different question sequences during inference. The best performance is produced with *Vehicle-Pedestrian* manner. Furthermore, the caption of the pedestrian is more precise in this manner, compared to the independent QA manner. The improvement indicates that informative and precise history or prompts are beneficial to the model generation. Conversely, the hallucinatory and incorrect one prevents the model from producing high-quality responses.

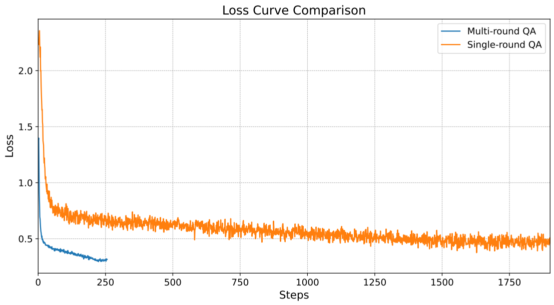


Figure 4. Training loss curves of models with multi-round and single-round QA.

	BLEU-4	METEOR	ROUGE-L	CIDEr	Score
			Pedestrian Statistics		
Independent QA	0.214	0.418	0.360	0.937	27.14
Pedestrian-Vehicle	0.214	0.418	0.360	0.937	27.14
Vehicle-Pedestrian	**0.215**	**0.425**	**0.366**	**0.989**	**27.62**
			Vehicle Statistics		
Independent QA	0.340	0.531	0.578	1.175	39.16
Pedestrian-Vehicle	0.330	0.521	0.566	1.076	38.12
Vehicle-Pedestrian	**0.340**	**0.531**	**0.578**	**1.175**	**39.16**

Table 7. **The performance on different question sequences during inference on `val` set**. **Pedestrian-Vehicle** indicates that firstly asking *pedestrian*, then asking *vehicle* in a sequential conversation. Similarly, **Vehicle-Pedestrian** implies the reverse order of queries. **Independent QA** denotes separate queries for the *pedestrian* and the *vehicle* in distinct conversations.

5. Conclusion

This paper proposes CityLLaVA, an efficient fine-tuning for VLMs in city scenarios. Based on the modules of bounding-box guided view selection, visual prompt engineering, textual prompt engineering, short QA construction, block expansion SFT, and prediction augmentation for LLaVA. Our method obtains the best score on the leaderboard.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. 2

[2] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan

Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 1, 2, 6

[3] Mu Cai, Haotian Liu, Siva Karthik Mustikovela, Gregory P. Meyer, Yuning Chai, Dennis Park, and Yong Jae Lee. Making large multimodal models understand arbitrary visual prompts. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2024. 3

[4] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechu Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023. 2

[5] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90 2

[6] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022. 2

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, 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. *International Conference on Learning Representations*, 2021. 2

[8] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. 2

[9] Quan Kong, Yuki Kawana, Rajat Saini, Ashutosh Kumar, Jingjing Pan, Ta Gu, Yohei Ozao, Balazs Opra, David C. Anastasiu, Yoichi Sato, and Norimasa Kobori. Wts: A pedestrian-centric traffic video dataset for fine-grained spatial-temporal understanding. 2024. 1

[10] 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 *International Conference on Machine Learning*, pages 19730–19742. PMLR, 2023. 2

[11] Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023. 2, 6

[12] Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, Jiaming Han, Siyuan Huang, Yichi Zhang, Xuming He, Hongsheng Li, and Yu Qiao. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. *arXiv preprint arXiv:2311.07575*, 2023. 4

[13] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. 2

[14] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 1, 2

[15] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 2, 3, 4, 5, 6, 8

[16] Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos? *arXiv preprint arXiv: 2403.00476*, 2024. 8

[17] Gen Luo, Yiyi Zhou, Yuxin Zhang, Xiawu Zheng, Xiaoshuai Sun, and Rongrong Ji. Feast your eyes: Mixture-of-resolution adaptation for multimodal large language models. *arXiv preprint arXiv:2403.03003*, 2024. 2

[18] Yingzi Ma, Yulong Cao, Jiachen Sun, Marco Pavone, and Chaowei Xiao. Dolphins: Multimodal language model for driving, 2023. 2

[19] OpenAI. Gpt-4. 2023. 5

[20] OpenAI. Gpt-4v(ision). 2023. 1, 2, 5

[21] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744, 2022. 2

[22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2

[23] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024. 6

[24] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. 6

[25] Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. What does clip know about a red circle? visual prompt engineering for vlms. *arXiv preprint arXiv:2304.06712*, 2023. 3

[26] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*, 2023. 2, 6

[27] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 2

[28] Shuo Wang, David C. Anastasiu, Zheng Tang, Ming-Ching Chang, Yue Yao, Liang Zheng, Mohammed Shaiqur Rahman, Meenakshi S. Arya, Anuj Sharma, Pranamesh

Chakraborty, Sanjita Prajapati, Quan Kong, Norimasa Kobori, Munkhjargal Gochoo, Munkh-Erdene Otgonbold, Ganzorig Batnasan, Fady Alnajjar, Ping-Yang Chen, Jun-Wei Hsieh, Xunlei Wu, Sameer Satish Pusegaonkar, Yizhou Wang, Sujit Biswas, and Rama Chellappa. The 8th AI City Challenge. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 2

[29] Licheng Wen, Xuemeng Yang, Daocheng Fu, Xiaofeng Wang, Pinlong Cai, Xin Li, Tao Ma, Yingxuan Li, Linran Xu, Dengke Shang, Zheng Zhu, Shaoyan Sun, Yeqi Bai, Xinyu Cai, Min Dou, Shuanglu Hu, and Botian Shi. On the road with gpt-4v(ision): Early explorations of visual-language model on autonomous driving. *arXiv preprint arXiv: 2311.05332*, 2023. 2

[30] Chengyue Wu, Yukang Gan, Yixiao Ge, Zeyu Lu, Jiahao Wang, Ye Feng, Ping Luo, and Ying Shan. Llama pro: Progressive llama with block expansion. *arXiv preprint arXiv:2401.02415*, 2024. 3, 6

[31] Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kwan-Yee. K. Wong, Zhenguo Li, and Hengshuang Zhao. Drivegpt4: Interpretable end-to-end autonomous driving via large language model, 2024. 2

[32] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 2

[33] Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. *arXiv preprint arXiv: 2312.00849*, 2023. 6

[34] Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization, 2024. 6

[35] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 2