Improving Event-Phase Captions in Multi-View Urban Traffic Videos via Prompt-Aware LoRA Tuning of Vision Language Models

Anonymous ICCV submission

Paper ID 13

Abstract

Understanding traffic scenes is a complex task for both humans and Vision Language Models (VLMs) due to the complexity of object interactions and event transitions. Accurate and informative captions on such scenes can support downstream tasks such as visual question answering and behavior recognition. In this paper, we present a captioning framework tailored for multi-view urban traffic videos, focusing on generative and informative video descriptions. Our approach leverages a vision language model specialized in spatial reasoning, which we then fine-tuned using LoRA for efficient adaptation to traffic-specific scenarios. To enhance descriptive accuracy, we employed bounding box information to guide a best-view selection, allowing the model to focus on salient regions such as vehicles and pedestrians. We further introduced prompt engineering strategies tailored to different camera perspectives-including vehicle, and overhead views, in order to optimize language grounding and scene specificity. The resulting captions demonstrate improved clarity and alignment with real-world traffic behaviors, offering valuable semantic context for downstream tasks such as visual question answering and event phase classification. Our experiments highlight the effectiveness of combining best-frame selection, prompt design, and lightweight fine-tuning in producing robust, multi-view-aware video captions under complex urban scenarios.

1. Introduction

Recent advances in multi-modal learning have produced vision-language models (VLMs) that can simultaneously interpret visual cues and generate fluent natural language. Flagship systems such as LLaVA [14], GPT-4V [16] and Qwen-VL [3] exhibit impressive general-purpose capability, but their performance tends to drop when faced with domain-specific imagery that differs from internet photo corpora. Urban traffic footage – shot from multiple vantage points, populated by fast-moving objects, obscured by occlusion and motion blur – presents exactly this kind of shift.

In this study, we focus on short video clips captured concurrently from overhead, dashboard and roadside cameras [15]. Crafting a concise description for each multi-view clip involves identifying the most informative camera angle, isolating key actors(e.g, vehicles, pedestrians), and articulating their interactions with spatial and temporal clarity. While possible to aggregate multiple views, our approach simplifies the problem by selecting a single best view per clip-guided by bounding box activity to reduce redundancy and computational load. This could be especially valuable in cooperative autonomous driving applications where the most informative streams of data need to be narrowed down from multiple sources. Although the dataset contains only about 20,000 annotated clips, the underlying SpaceLLaVA model has over 7 billion parameters. Fully fine-tuning such a model is both memory-intensive and prone to overfitting when data is limited. We tackle these constraints with **BestViewPrompt-Tuned SpaceLLaVA**, a lightweight pipeline that combines parameter-efficient adaptation, bounding-box-driven view selection, and perspective-aware prompt design:

1. **LoRA adaptation**: low-rank adapters update fewer than 2 % of SpaceLLaVA's weights, enabling overnight training on a single commodity GPU.
2. **Best-view extraction**: bounding boxes from object detectors guide a simple heuristic that picks the frame–camera pair richest in traffic activity, eliminating redundant or empty views.
3. **Perspective-aware prompts**: tokens that encode camera type (`[VEHICLE]`, `[OVERHEAD]`) supply geometric context and steer the model toward spatially grounded phrasing.

To measure caption quality, we report[1] BLEU [17], METEOR [4], ROUGE-L [13], and Consensus-based Image

[1]We use the AI City Challenge official evaluation script to compute these metrics.

Description Evaluation (CIDEr) [20]:

1. **BLEU** measures n-gram overlap between generated and reference captions.
2. **METEOR** accounts for synonym and stem matches.
3. **ROUGE-L** evaluates the longest common subsequence.
4. **CIDEr** scores based on TF-IDF–weighted n-gram similarity with consensus captions.

We compute these metrics separately for *pedestrian* and *vehicle* categories, then normalize the average to a 0–100 scale across both internal and external sets of the AI City Challenge 2025 Dataset. Our system improves the CIDEr score over a vanilla SpaceLLaVA baseline by $\Delta\text{CIDEr} = 0.08$, which although may seem modest, it indicates a measurable improvement in the model's ability to produce more informative human-aligned descriptions. Whom are especially important in structured, domain-specific tasks like traffic scene captioning.

Contributions. In addition to the framework described above, we contribute an ablation analysis that quantifies the individual impact of view selection, prompt design, and parameter-efficient fine-tuning.

2. Related Work

In this section, we review the foundational work relevant to our video captioning approach in the traffic safety domain. We begin by surveying major developments in vision–language models (VLMs), which serve as the backbone for multi-modal understanding. Following that, we highlight recent efforts to adapt such models to domain-specific applications in autonomous driving and traffic analysis, with a focus on large-scale, real-world datasets.

2.1. Vision-Language Models

Recent years have witnessed rapid progress in vision-language models (VLMs) that jointly process both image and text modalities. A foundational model in this space is CLIP [18], which aligns image-text pairs through contrastive pre-training, enabling robust zero-shot performance on downstream tasks. More advanced architectures such as Flamingo [1] interleave image and text tokens via gated cross-attention, allowing fluent, context-aware generation, but at the cost of enormous parameter counts. Blip-2 [11] introduces a lightweight Q-former module that bridges the modality gap between image embeddings and a frozen language model (Flan-T5 [7]), enabling efficient and effective fine-tuning. MiniGPT-4 [22] simplifies further by projecting visual features directly into Vicuna [6] an open-source, instruction-tuned derivative of LLaMA-through a single projection layer, but sacrifices spatial resolution. An even simpler pipeline is used in MiniGPT-v2 [5], which directly maps ViT-encoded tokens to the language model's embedding space, trading modeling flexibility for simplicity.

While these models offer increasingly general capabilities, most are trained on internet-scale image–text corpora and lack strong spatial or temporal grounding—two properties critical for real-world video understanding.

2.2. VLMs for Driving Scenarios

Applying VLMs to driving contexts presents unique challenges due to the need for spatial grounding, fine-grained temporal understanding, and multi-camera fusion. GPT-4V [16] shows promise in understanding urban scenes but struggles without domain adaptation [10], particularly when faced with motion blur or occlusion. Dolphins [21], based on OpenFlamingo [2], incorporates driving-specific pre-training to improve captioning in dynamic traffic videos but processes frames independently and lacks multi-view integration. DriveGPT-4 [12] further bridges perception and action by generating both descriptions and control signals from video–text input, though it requires reinforcement learning and significant training cost.

These works highlight the importance of tailoring VLMs to the structured demands of driving tasks. Yet, few have addressed the unique setting of multi-view captioning in urban traffic scenes—where camera selection, perception, spatial precision, and efficient training are all critical.

Our distinction. We build on **SpaceLLaVA**, a 7B-parameter VLM explicitly trained for spatial reasoning, and propose a novel *BBoxPrompt-SpaceLLaVA* pipeline that (i) selects the most informative view using bounding-box activity, (ii) injects geometric priors via perspective-aware prompts, and (iii) applies LoRA-based fine-tuning for domain adaptation under limited data.

3. Methodology

3.1. Overview

We build upon VLMS to generate fine-grained video captions in urban traffic environments. Our method includes selecting optimal camera views, applying visual preprocessing by cropping bounding box regions to focus the model's attention, constructing textual prompts with contextual tokens, and fine-tuning using Low-Rank Adaptation (LoRA) for efficiency. While the framework is compatible with most spatially-aware VLMs, we implement it using **SpaceLLaVA-7B**, a publicly available model trained for vision-language reasoning, due to its strong performance and open-source availability.

3.2. Dataset Preparation

Our dataset supplies synchronized overhead, vehicle-mounted, and roadside videos, each accompanied by frame-level bounding boxes for pedestrians and vehicles.

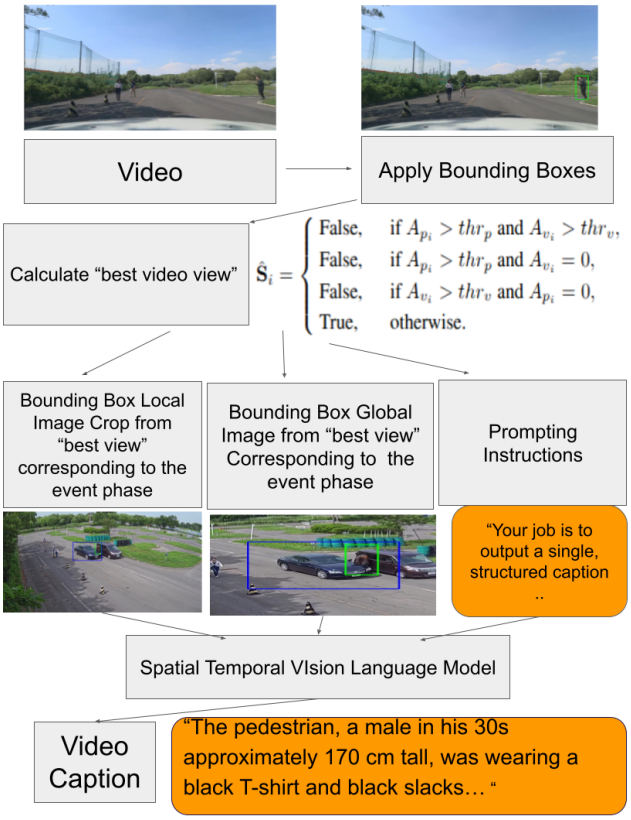


Figure 1. Our pipeline consists of using the generated bounding boxes to calculate the video in which the interested pedestrian and vehicle are most visible. Then passing in two image (full image, cropped image) corresponding to the event phase of our scenario. These images are then passed into the VLM, which then generates a descriptive caption of both the pedestrian and interested vehicle.

Bounding Box-Guided Crops. To improve spatial focus and reduce background noise, we preprocess each frame using bounding boxes to highlight localized views of key subjects (e.g., pedestrians, vehicles). This allows the model to generate more precise, context-aware captions by attending directly to the relevant visual regions.

Best-View Selection Following the guided view-selection strategy of CityLLaVA [8], we treat every clip as a tuple $S = \{V, T, B\}$, where V is the three-view video, T the ground-truth caption, and B the set of 2-D detections. For each view we compute the *average vehicle area* A_v and *average pedestrian area* A_p (in pixels) over its bounding boxes. A view is **discarded** if

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

with thresholds $\tau_p = 3{,}000$ px and $\tau_v = 5{,}000$ px (0.3 % and 0.5 % of a 1920×1080 frame). Intuitively, we keep views that contain *either* clear vehicles *or* clear pedestrians but down-weight views dominated by one class to the exclusion of the other, which often correspond to occlusions or distracting foreground objects.

For the surviving view we perform two visual crops:

1. Global frame — the full 1920×1080 image.
2. Local crop — the union of all bounding boxes, iso tropically enlarged by $1.5\times$ to retain context.

Both images are fed to the VLM by padding them side-by-side and encoding as a single visual token sequence; this gives the model simultaneous access to global layout and fine-grained object detail.

3.3. Prompt Engineering

To help the model generate context-aware captions, we construct textual prompts that encode scene-specific information. One key element is the inclusion of an **event phase token**, which reflects the semantic category of the current video clip — such as `[AVOIDANCE]`, `[JUDGMENT]`, `[RECOGNITION]` or `[PRECOGNITION]`. These labels correspond to different types of pedestrian or vehicle behavior.

These tokens are then prepended to a fixed prompt template that instructs the model to describe the clip in a structured format. For example:

```
[AVOIDANCE]: Your job is to output a
single, structured caption wrapped
in <answer>...</answer> tags.  The
caption should briefly describe
the subject's appearance, position,
motion, and surroundings, using
short declarative sentences.  Avoid
conversations or extra commentary.
```

This prompt helps the model align its language generation with the high-level activity depicted in the scene, encouraging more relevant and behavior-specific descriptions. To empirically validate the impact of including event phase tokens, we compare caption outputs generated by the model with and without the token prepended to the prompt. Figure 2 shows the input image used for this test.

Table 1 presents the resulting captions. The inclusion of the `[AVOIDANCE]` token leads to a caption that captures the subject's behavior during the avoidance phase more precisely, whereas the caption generated without the token is more generic and less descriptive.

3.4. Fine-tuning with LoRA

Fine-tuning the entire SpaceLLaVA model can be computationally expensive. We adopt **LoRA** [9], which introduces lightweight trainable *adapter* modules into the attention layers of the transformer. These adapters consist of a

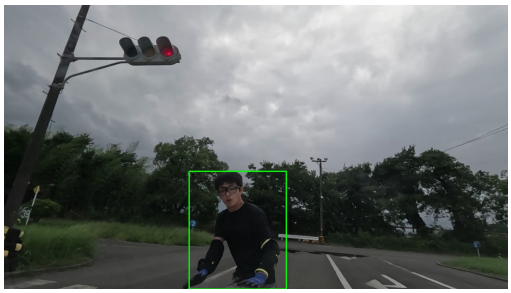


Figure 2. Input traffic scene used for prompt comparison experiments.

Table 1. Comparison of captions generated using original prompt versus prompt with event phase token.

Original Prompt Caption	**Prompt with Event Phase Token Caption**
The pedestrian is wearing dark clothing and standing still.	The pedestrian wearing a black shirt is crossing the street in front of the vehicle during the avoidance phase.

Figure 3. Training/loss (cross-entropy) curve during fine-tuning of the vision-language model using LoRA adapters. The loss decreases steadily over epochs, indicating stable convergence without overfitting.

pair of low-rank matrices that project the input to a smaller subspace and then back to the original dimension. This allows the model to learn task-specific behavior with minimal updates to the full model weights. Key hyper parameters include:

- **Rank** ($r = 128$): Controls the complexity of the adapter's bottleneck layer.
- **Alpha** (256): Scales the adapter updates relative to the base model.
- **Dropout** (0.05): Adds regularization to prevent overfitting.

We utilize **DeepSpeed ZeRO-2** optimization [19] to manage memory usage during training on an NVIDIA A100 GPU with 40GB VRAM. Additional techniques such as **gradient checkpointing**, which recomputes activations during backpropagation to save memory, and **lazy preprocessing**, which defers image-to-token conversion until just-in-time execution, help further reduce peak memory consumption.

We use the **cross-entropy loss** function to train the model, which measures the difference between predicted caption tokens and ground truth captions. This objective encourages the model to generate fluent and accurate descriptions aligned with annotated labels.

3.5. Inference

At inference time, the model receives a selected video frame along with its cropped bounding-box views and a corresponding prompt. We generate captions using **nucleus sampling** with top-$p = 0.9$, a decoding strategy that restricts sampling to the smallest set of most probable next tokens whose cumulative probability exceeds p. This balances output diversity with fluency by avoiding both highly unlikely and overly generic completions. The caption length is capped at 512 tokens to reduce verbosity and minimize hallucinated content.

Final captions are post-processed by extracting the text inside the first `<answer>...</answer>` block, then removing any surrounding prompt artifacts (e.g., `USER:`, `ASSISTANT:`) if present. If no structured tags are found, we fall back to returning the original model output. This ensures that only the intended caption content is retained for evaluation.

4. Experiments

4.1. Evaluation Metrics

To evaluate caption quality, we use standard natural language generation metrics: BLEU-4, METEOR, ROUGE-L, and CIDEr. These metrics assess aspects such as n-gram precision, semantic overlap, and syntactic fluency between generated captions and reference annotations.

Following common practice in multi-view traffic captioning benchmarks, we report a composite score computed as:

$$\text{Score} = \frac{\text{BLEU-4} + \text{METEOR} + \text{ROUGE-L} + 0.1 \cdot \text{CIDEr}}{4} \cdot 100.$$

All results are computed on the AI City Challenge 2025 validation set.

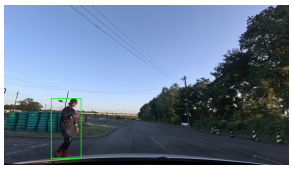
4.2. Ablation Study

To isolate the contribution of each component, we vary one factor at a time and hold the others fixed.

graphicx

Table 2. Captioning performance comparison across model variants and input settings.

Model / Setting	BLEU-4	METEOR	ROUGE-L	CIDEr
Base (Zero-shot)	0.017	0.187	0.136	0.000
LoRA Fine-tuned	0.033	0.265	0.213	0.000
LoRA + Prompt Engineering	0.038	0.280	0.225	0.000
LoRA + BBox Cropping	0.041	0.290	0.230	0.080
LoRA + Prompt + BBox Cropping	**0.045**	**0.300**	**0.240**	**0.080**



With BBox: The color of the pedestrians shoes are red.

Without BBox: The color of the pedestrians shoes are black.

Figure 4. Comparison of SpaceLLaVA-7B captions with bounding boxes (left) and without bounding box-guided input (right). Bounding box crops help the model focus on relevant actors and generate more informative descriptions. Both tests were given the same prompt "What color is the color of the interested pedestrians shoes?

Bounding-Box Crops. Removing cropped local views lowers fine-grained detail recognition, decreasing CIDEr by ≈ 0.08 points.

Effect of Bounding Box Cropping. We investigate the impact of bounding box-guided cropping on caption generation quality. As shown in Figure 4, the model produces more specific and semantically rich captions when the subject is isolated using bounding boxes. In contrast, global views often lead to vague or contextually ambiguous descriptions, especially in cluttered scenes or when the subject occupies a small area.

Why Bounding Boxes Help. Bounding box cropping provides critical spatial focus by isolating the key subjects—pedestrians and vehicles—within each frame. This focused view reduces background clutter and irrelevant details that can confuse the vision-language model, especially in busy urban traffic scenes. By presenting the model with a focused image centered on the actor of interest, it is better able to extract fine-grained features such as clothing color, posture, or subtle motions that are essential for accurate captioning.

Moreover, bounding box crops improve the model's ability to distinguish overlapping actors or small objects that would otherwise be lost or misinterpreted in a global view. This precise localization enhances the semantic relevance of generated captions, leading to improvements in metrics like CIDEr and BLEU-4.

Prompt Engineering. As show in table 2 dropping the camera/phase tokens degrades BLEU-4 and CIDEr, confirming that structured prompts steer the model toward more context-aware descriptions.

5. Concluding Remarks

5.1. Limitations

Our current system relies on a single representative frame at training time; richer temporal modeling may further improve motion-related descriptions. In addition, BLEU-4 is sensitive to exact word overlap and can undervalue semantically correct paraphrases, so future work should consider alternative metrics.

5.2. Future Directions

There are several promising avenues to extend this work:

- **Temporal Modeling.** While our current system processes frames independently, incorporating a lightweight temporal encoder (e.g., a causal transformer or temporal convolution) could allow the model to capture motion patterns and inter-frame dependencies directly, improving coherence and action recognition.
- **Prompt Optimization.** We aim to explore automated prompt search using instruction-tuned LLMs to systematically optimize prompt phrasing and structure across different event types and views.
- **Enhanced Use of Detection Outputs.** Beyond cropping around bounding boxes, future work may leverage class labels, detection confidence scores, and object trajectories (e.g., from YOLO or ByteTrack) to provide richer grounding cues, helping the model better understand spatial relationships and evolving scene dynamics.
- **Best-View Selection Strategies.** Our current view selection heuristic is based on bounding-box area thresholds; future work could investigate alternative criteria (e.g., visibility metrics or learned scoring functions), as well as adaptive threshold tuning based on scene complexity.
- **Broader VLM Evaluation.** While we used SpaceLLaVA due to its strong spatial reasoning capabilities, applying our pipeline to other spatially-aware VLMs would help assess generalization and robustness across architectures.

5.3. Conclusion

We presented a parameter-efficient SpaceLLaVA-7B captioning pipeline that combines bounding-box–guided view selection, view-aware prompts, and LoRA fine-tuning. Our approach improves both language quality and event-phase alignment on a benchmark dataset, demonstrating the value of targeted prompt engineering and lightweight adaptation for multi-view traffic video captioning.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, et al. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, 2022. 2

[2] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023. 2

[3] Jinze Bai, Yunfei Chu, Fei Huang, et al. Qwen-vl: A versatile vision-language model. https://huggingface.co/Qwen/Qwen-VL, 2023. Accessed: 2025-07-05. 1

[4] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, pages 65–72. ACL, 2005. 1

[5] Jun Chen, Deyao Zhu, Xiaoqian Shen, et al. Minigpt-v2: Large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023. 2

[6] Wei-Lin Chiang, Zhuohan Li, Zi Lin, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90% chatgpt quality. *arXiv preprint arXiv:2304.11264*, 2023. 2

[7] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022. 2

[8] Yiyang Duan, Zhongzheng Wang, Yifei Meng, Li Zheng, Weiyao Chen, Minghui Xu, and Liang Wang. Cityllava: Efficient fine-tuning for vlms in city scenario. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1577–1586, 2024. 3

[9] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Chen, and Weizhu Li. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 3

[10] Chen Li et al. Evaluating gpt-4v in the domain of autonomous driving. In *CVPR Workshops*, 2024. 2

[11] Junnan Li, Dongxu Zhang, Xue Li, Ming Tan, Yusuke Sugano, Yongxin Yang, Xiaokang Hu, Zicheng Wang, Lei Li, et al. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 2

[12] Qing Li et al. Drivegpt4: A multimodal model for text-based driving simulation. In *ECCV*, 2024. 2

[13] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop*, pages 74–81. ACL, 2004. 1

[14] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 1

[15] Milind Naphade, Zeyu Wang, Ming-Yu Chang, Liang Zheng, et al. Ai city challenge 2025: Traffic safety description and analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2025. AI City Challenge Track 2. 1

[16] OpenAI. Gpt-4 with vision. https://openai.com/research/gpt-4v, 2023. Accessed: 2025-07-05. 1, 2

[17] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318. ACL, 2002. 1

[18] 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. *arXiv preprint arXiv:2103.00020*, 2021. 2

[19] Jesse Rasley, Samyam Rajbhandari, Oscar Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3505–3506, 2020. 4

[20] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 4566–4575, 2015. 2

[21] X Zhang et al. Dolphins: A vision-language model for driving scenarios. In *CVPR Workshops*, 2024. 2

[22] Deyao Zhu, Jun Chen, Xiaoqian Shen, et al. Minigpt-4: Enhancing vision-language understanding with gpt-4. *arXiv preprint arXiv:2304.10592*, 2023. 2