EPiC Series in Technology

Volume 3, 2026, Pages 65–76

Proceedings of AI for Accelerated Research Symposium

Multi-Frame Grid Perspective for Traffic Video Captioning and Context-Aware VQA

Sanjita Prajapati[1], Ashutosh Dumka[1], Rajan Thakulla[2], Atmadip Goswami[3], Karo Ahmadi Dehrashid[4], and Anuj Sharma[1]

[1] Iowa State University, USA
sanjitap@iastate.edu, adumka@iastate.edu, anujs@iastate.edu
[2] National Institute of Technology Rourkela, India
[3] Indian Institute of Technology Madras, India
[4] Webster University, MO, USA

Abstract

"What really happened? Who was at fault? Did the pedestrian yield, or was the driver distracted?" In high stakes traffic incidents, understanding pedestrian vehicle interactions is essential for safety assessment, post crash analysis, and insurance decision making. We propose a novel vision language framework for traffic safety captioning and visual question answering (VQA), designed for the AI City Challenge 2025. Leveraging LLaVA 1.5 as our base vision language model, we introduce a multi frame collage input strategy to embed temporal context into image based architectures. We explored three input transformation techniques, Box Stitch, Blur Stitch, and Arrow Stitch, to emphasize semantic cues such as entity localization, contextual filtering, and motion trajectory. Structured captions are generated through a two stage process: LLaVA extracts fine grained semantic features via targeted question answering, which are then converted into narrative descriptions using Mistral 7B. For VQA, Mistral further reasons over structured scene features to identify the most contextually appropriate response. Our best performing configuration, Box Stitch, achieves an S_2 score of 33.93 on the official test set, demonstrating the effectiveness of structured prompting, modular caption pipelines, and strategic visual input augmentation in understanding pedestrian vehicle interactions. This work highlights the promise of combining static visual backbones with image based temporal fusion for traffic scenario comprehension.

1 Introduction

"A crash occurs at an urban intersection. Within seconds, insurance agents, city officials, and autonomous systems all need to understand what happened, who was involved, and whether any risky behaviors were observed." In such high stakes scenarios, generating accurate and structured captions from traffic video footage can play a pivotal role in safety monitoring, liability determination, and automated post incident analysis. Captioning provides a natural language summary of events, enhancing interpretability and supporting critical decisions for insurance workflows and transportation systems.

J. Masnec, H.R. Karimian, P. Kordjamshidi and Y. Li (eds.), AIAS 2025 (EPiC Series in Technology, vol. 3), pp. 65–76

Recent advancements in large visual language models (VLMs) have opened new opportunities for tackling this challenge. Research communities at the intersection of computer vision (CV) and natural language processing (NLP) have begun to harness these models for traffic video analysis, raising benchmarks for automated interpretation while revealing gaps when applied to domain specific tasks [25, 7, 19]. At the same time, the growing interest in smart city infrastructure has motivated the development of pedestrian vehicle interaction captioning methods, supported by richly annotated datasets that capture fine grained behavior and contextual cues [23, 10]. However, directly deploying pretrained VLMs often fails to capture the subtle semantics of traffic scenes, underscoring the need for domain aware fine tuning strategies that address both visual and linguistic nuances.

Our study introduces a comprehensive paradigm tailored for the AI City Challenge 2025. First, we perform an in depth decomposition of dense traffic descriptions into semantically coherent segments (e.g., appearance, location, environment, attention, action) and employ large language models to generate an expanded question answer bank that more effectively encapsulates domain specific queries. Second, we fine tune a state of the art VLM on both the Woven Traffic Safety (WTS) [10] and `BDD_PC_5K` [6] datasets, exploring three distinct input approaches: (1) a bounding box model, which overlays pedestrian and vehicle bounding boxes directly on video frames; (2) a blur model, which obfuscates peripheral content to emphasize the subject (pedestrian and vehicle); and (3) a trajectory model, which visualizes motion by connecting initial and terminal pedestrian and vehicle bounding box centroids with vectorized paths.

Through rigorous evaluation on the AI City Challenge 2025 Track 2 test set [10], our Box Stitch approach achieves an S2 score of 33.93, demonstrating that targeted segment extraction combined with specialized input augmentations can substantially elevate traffic safety captioning performance. While our framework adopts open-source VLM backbones, the novelty lies in the temporal collage representation and structured two-stage prompting that enable lightweight temporal reasoning using image-based models.

Section 2 briefly reviews existing methods related to our problem and solution. Then, we present our proposed approaches and data preparation method in Section 3. In Section 4, we describe our model training and inference pipelines. Section 5 presents the evaluation, followed by conclusions and future work in Section 6.

2 Related Work

2.1 Image and Video Captioning

Image captioning has progressed from RNN-based models to advanced transformer architectures. Early works by Vinyals et al. [28] and Xu et al. [8] used LSTMs and attention to generate context-aware captions, while Anderson et al. [9] refined attention through stacked LSTMs. The introduction of Transformer decoders [5] enabled direct integration of image embeddings via cross-attention [20, 12, 13]. Large-scale vision–language pretraining further advanced performance, with Zhou et al. [28] proposing a unified encoder–decoder design, Zhang et al. [35] improving grounding through detection, and BLIP [16] filtering noisy image–text pairs. In traffic safety, dense video captioning is central, with PDVC enhancing temporal coherence [26, 11], streaming models supporting real-time analysis [38], and PR DETR improving event localization [37]. Domain-specific approaches leverage CLIP features on WTS and BDD 5K datasets [26], while frameworks like TrafficVLM [21], DRAMA [3], and multi-view transformers [18] address risk prediction and action recognition. Recent unsupervised and nested transformer methods continue

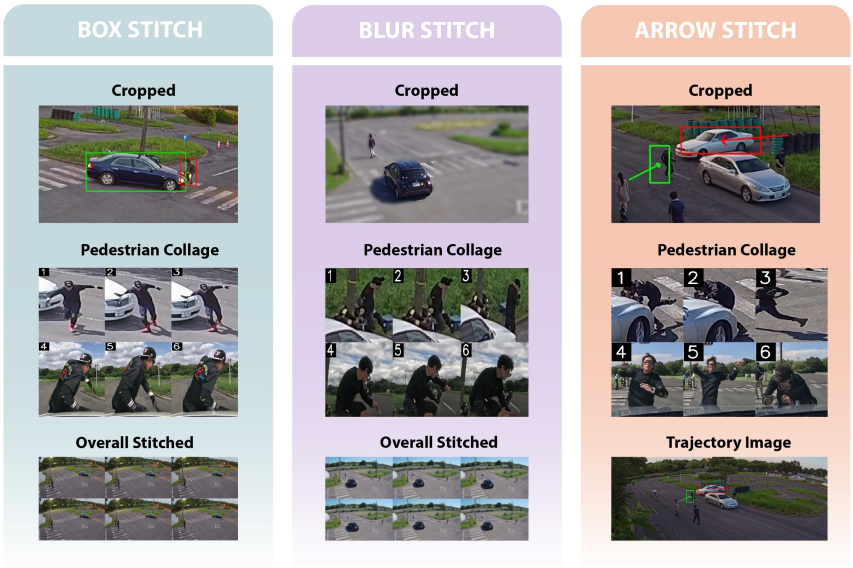

Figure 1: We construct three stitched image types to represent temporal and spatial cues: (a) Box Stitch highlights cropped bounding boxes for key agents; (b) Blur Stitch suppresses background to emphasize motion; (c) Arrow Stitch adds directional cues to illustrate intent and interaction

to enrich video-to-text generation [36, 32].

2.2 Vision Foundation Model

Transformer based image text encoders such as CLIP [24] and ALIGN [14] learn visual concepts via contrastive learning on massive, noisy web data, enabling strong zero shot transfer. BLIP [16] further boosts performance by pretraining on synthetic image caption pairs, achieving state of the art results on vision and image language benchmarks. However, these models lack motion and temporal modeling, limiting their applicability to video tasks [33]. Early video foundation approaches including BEVT [30], Masked Feat [31], and VideoMAE2 [29] extend masked autoencoding into spatio temporal domains. More recent methods pursue unified pretraining; LAVENDER [17] frames tasks as masked language modeling, and MERLOT Reserve [34] trains on 20M video text audio pairs using contrastive span matching, achieving top performance across diverse video benchmarks.

2.3 Large Vision Language Models

Vision language models (VLMs) jointly exploit visual inputs and textual data to integrate cross modal knowledge and improve task performance. CLIP [24] pioneered this paradigm by pretraining on image caption pairs using a contrastive objective, yielding remarkable zero shot generalization across diverse applications. As large language models have advanced [7, 22, 27], a new strategy has emerged that pairs a visual feature extractor with an autoregressive language decoder, combining perceptual grounding with fluent text generation. Building on this, BLIP 2 [15] introduces the Q Former, a lightweight adapter module that transforms visual embeddings

into a space compatible with a pretrained language model such as Flan T5 [2]. MiniGPT v2 [1] further streamlines the process by mapping raw ViT [4] token outputs directly into the language decoder's embedding space. LLaVA [19] follows a similar projection based design but augments it with a two stage fine tuning regime that aligns modalities and then refines task specific capabilities. Each explores variations in projection mechanisms, training curricula, and adapter architectures to push the boundaries of unified vision language understanding.

3 Methodology

The Woven Traffic Safety (WTS) dataset provides an in-depth analysis of pedestrian-related traffic footage, focusing on spatial and temporal aspects. It includes two main tasks: generating captions for segments involving pedestrians and vehicles and selecting answers from a Visual Question Answering (VQA) test set. We explored three approaches to train Vision Language models by altering data input methods. The following section details our proposed methodology for these tasks.

3.1 Best View Selection

During our review of the recommended perspectives, we identified several problematic views due to poor visibility of pedestrians and suboptimal overall views. To address these issues, we used a bounding box ROI (Region of Interest) approach to assess each scenario. We calculate the area of the bounding boxes for the vehicle and pedestrian (if present), assigning zero if neither is found. We then select the camera view based on the maximum area calculated.

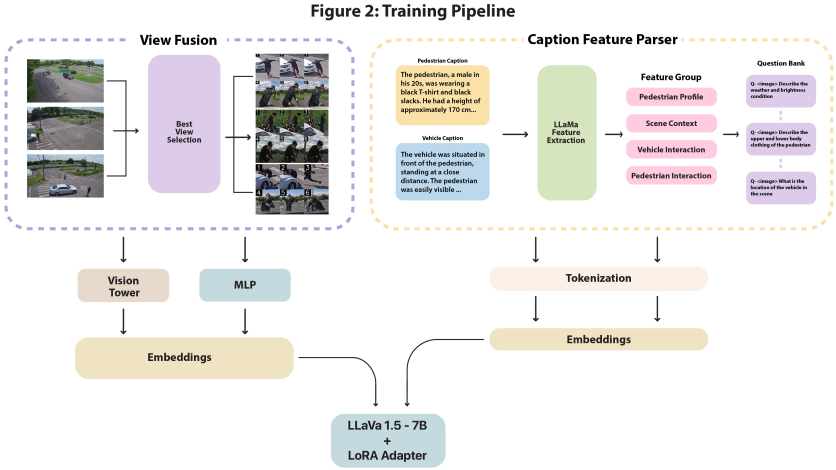

Figure 2: Overview of the training pipeline. Multi-view frames are fused into stitched grids capturing temporal dynamics. Structured captions are parsed into four semantic groups: Pedestrian Profile, Scene Context, Vehicle Interaction, and Pedestrian Interaction. Visual and textual inputs are fed into a LLaVA-1.5-7B model fine-tuned with LoRA adapters to align temporal context with scene semantics.

3.2 Caption Feature Extraction

During the review of training and validation caption files, we noted that each caption can be broken down into segments, consisting of short pieces of information like individual words. These segments fall into four categories: (1) Pedestrian Characteristics (age, height, clothing), (2) pedestrian-related events, (3) vehicle-related events, and (4) environmental information. Our goal is to segment captions into brief questions that support caption generation. Instead of doing this manually, we created a question bank using Llama to extract fixed questions for training our model. We recognize that this method might impact sentence structure and information accuracy, leading to a trade-off between precision and preservation. To mitigate this, we used prompt techniques to help the model extract accurate information from the captions in the identified segments.

3.3 Dataset Preparation

We tried three different approaches using different styles of data input for our training process. In this section, we will discuss the dataset for the three approaches, i.e., (1) Box Stitch, (2) Blur Stitch, and (3) Arrow Stitch were prepared.

3.3.1 Multi-frame Collage Input Strategy

Unlike prior methods that use 8-frame video sequences [7, 20], we treat captioning as a multi-frame visual reasoning task. Six temporally diverse frames are stitched into a grid, as shown in Figure 1, to create a single image compatible with image-based vision–language models like LLaVA while capturing richer temporal context. This design is motivated by two observations: (1) image-based models often outperform video-based ones on captioning and reasoning tasks when provided with dense visual context, and (2) selecting frames with the greatest visual movement or scene change retains temporal diversity while simplifying the input modality to image only.

To train using a single image per segment instead of full videos, we select six frames that best capture variation over time. Starting from an initial frame, Gaussian divergence measures visual difference between candidates, prioritizing those with higher divergence (e.g., pedestrian motion or vehicle approach) to ensure key transitions. For short clips under one second, limited frame diversity may restrict dynamic coverage. For real-time use, Gaussian divergence can be replaced by a rolling-window motion-saliency or optical-flow estimator, providing similar diversity at negligible compute cost (under 10 ms per clip).

3.3.2 Input Transformation Strategies

To enhance the model's ability to focus on event relevant visual cues, we designed three input transformation approaches corresponding to the segment categories shown in Figure 2, View Fusion block.

1. **Box Stitch**: For each extracted frame, we overlaid bounding boxes around pedestrians and vehicles. This localization highlights key entities, helping the model attend to appearance, attention, and spatial context more effectively during caption generation.

2. **Blur Stitch**: In this strategy, the background and all non-essential regions were blurred, preserving only the subject(s) within the bounding box. Reducing visual noise helps isolate critical interactions and supports reasoning over event-specific segments, such as pedestrian and vehicle actions or potential conflicts.

3. **Arrow Stitch**: We visualized motion by drawing a trajectory vector from the first frame's bounding box's centroid to the last frame's box for each subject. This representation encodes movement patterns, aiding the model's understanding of dynamic interactions, such as approach angles and collision risk. Arrow Stitch currently uses first-order straight vectors for efficiency. Future work will extend this with curved or optical-flow-based trajectories to capture non-linear motion.

Figure 3: Interface Pipeline

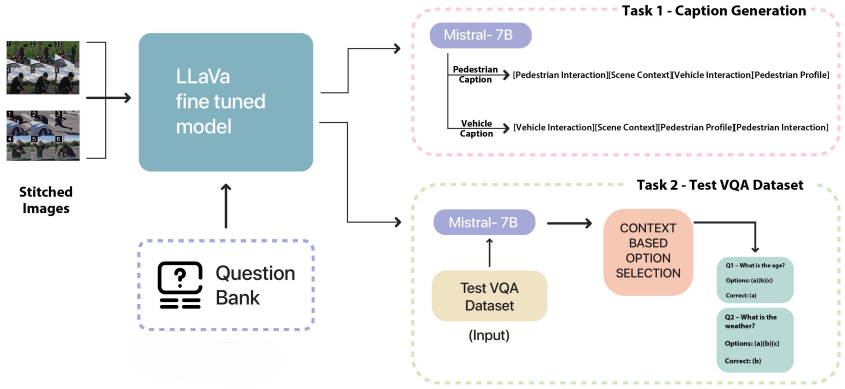

Figure 3: The interface pipeline performs two tasks: (a) Caption Generation—LLaVA processes stitched images with semantic prompts, and Mistral-7B organizes outputs into structured captions based on four feature groups: Pedestrian Interaction, Vehicle Interaction, Scene Context, and Pedestrian Profile; (b) VQA Resolution—Mistral-7B interprets the query, aligns it with relevant features from LLaVA captions, and selects the most contextually accurate answer.

4 Model Architecture and Training

Our methodology consists of three core components, vision language model fine tuning, structured caption generation, and contextual VQA reasoning. Each module is built upon stitched multi-frame inputs and parsed semantic features to enable a rich understanding of pedestrian-vehicle interactions. We adopt a modular two-stage design (LLaVA - Mistral) rather than joint end-to-end training to maintain interpretability and allow independent optimization, an essential property for safety-critical traffic applications.

4.1 Vision Language Model Fine-Tuning

We use `llava-1.5-7b` as our base vision-language model (VLM) due to its strong multimodal reasoning capabilities and efficient support for adapter-based tuning. To encode temporal context without relying on video models, we convert each video segment into a stitched 6-frame collage (2×3 grid), capturing motion cues across time in a single image.

This approach is motivated by two insights: (1) image-based VLMs perform competitively on structured reasoning tasks when provided with dense visual input, and (2) selecting frames

with maximum scene change helps retain temporal variation. Each collage is paired with a structured instruction and answer, forming a causal language modeling (CLM) training triplet. Since the stitched layout differs significantly from the CLIP style images seen during pretraining, we finetune the vision tower's final self attention layer and insert LoRA adapters in the language model. This helps the model adapt to the spatial structure and reason over visual changes across frames.

To enable efficient tuning, we apply low rank adaptation (LoRA) to selected layers of the model. Specifically, LoRA adapters are inserted into the language model's `q_proj`, `k_proj`, `v_proj`, and `o_proj` layers, as well as the vision tower's final block, `encoder.layer.23`, which includes self attention components `q_proj`, `k_proj`, `v_proj`, and `out_proj`. The LoRA configuration used is: $r = 8$, $\alpha = 32$, and dropout rate = 0.05. Only these injected adapter weights are trainable, while the rest of the model remains frozen. We use `bfloat16` mixed precision for efficient GPU utilization.

4.2 Training and Implementation

Each training instance includes a stitched image I, a prompt x prefixed with `<image>`, and an answer y. The training objective is defined using the causal language modeling (CLM) loss in Equation 1:

$$\mathcal{L}_{\text{CLM}} = -\sum_{t=1}^{T} \log P(y_t \mid y_{<t}, x, I; \theta) \tag{1}$$

Training is conducted using HuggingFace's `Trainer` with a batch size of 4 (via gradient accumulation), 15 epochs, a learning rate of 2×10^{-4} with 5% warmup, and a cosine learning rate scheduler.

To avoid overfitting this large scale dataset with structured prompts and stitched multi-frame inputs, we limited training to 15 epochs, a choice driven by the risk of memorization in repetitive or constrained scene templates. Training was conducted on NSF provided infrastructure using 8× NVIDIA A100 GPUs (80 GB each), enabling efficient fine-tuning of the 7B-parameter `llava-1.5-7b` backbone with approximately 60 million LoRA parameters applied to both the language model and the final vision encoder layer.

4.3 Structured Caption Generation

The captioning process begins by running `LLaVA-1.5` on each stitched 6-frame image, using structured questions from the Caption Feature Parser focused on pedestrian profile, scene context, vehicles, and interactions. `LLaVA` generates concise, interpretable answers that capture scene level semantics. These question-answer pairs are then fed into `Mistral-7B-Instruct` for paragraph style caption generation. We use single shot prompting, combining system instructions, a few shot example, and the full Q&A set to guide the model as shown in Figure 3. This structured approach helps `Mistral` generate coherent, factual captions that align closely with the visual content, ensuring narrative consistency across diverse traffic scenes and improving interpretability in downstream tasks. To maintain linguistic diversity and avoid rigid templates, multiple paraphrased question formats and randomized prompt orders were used during caption generation.

4.4 Visual Question Answering (Context Refiner)

We implement a visual question answering (VQA) module using `Mistral-7B-Instruct` to enable fine-grained reasoning over complex traffic scenes. Each stitched image is paired with a high-level multiple-choice question assessing situational understanding, such as pedestrian profiles, scene context, and vehicle-pedestrian interactions (Figure 2). Structured answers previously inferred by `LLaVA` and parsed by the Caption Feature Parser module are used for contextual grounding. When presented with a VQA question and its answer options, the `Mistral` model first interprets the question's intent, compares it with parsed caption features, and identifies the most semantically aligned Q&A pair. Leveraging this contextual understanding (Figure 3), the model evaluates all options and selects the one best matching the scene interpretation. This reasoning process strengthens answer accuracy by linking visual semantics with language-based inference, allowing the model to handle ambiguity and better capture pedestrian–vehicle dynamics.

5 Evaluation

This section presents the evaluation methodology, including metrics and experimental results. We focus on analyzing the impact of different visual input transformation strategies through ablation studies.

5.1 Evaluation Metrics

For evaluation, we use BLEU-4, METEOR, ROUGE-L, and CIDEr for sub-task 1 of caption generation. These metrics gauge the textual overlap and semantic correspondence between the generated and ground truth captions. The S_1 score (Equation 2) is obtained as an average of the four metrics for the internal and external datasets. For sub-task 2, which involves correctly answering the question, accuracy is used (Equation 3). Averaging both the scores gives us the final S_2 score.

$$S_1 = \frac{\text{BLEU-4} + \text{METEOR} + \text{ROUGE-L} + \text{CIDEr}}{4} \times 100 \tag{2}$$

$$\text{Accuracy} = \frac{\text{Correct Answers}}{\text{Total Questions}} \tag{3}$$

Table 1 summarizes the comparative performance of the proposed visual transformation strategies. The Box Stitch approach delivered the best results, with an S_1 score of 20.39, an S_2 of 35.4, accuracy of 51.58%, and a composite score of 35.79. The explicit use of bounding boxes helped the model localize pedestrians and vehicles effectively, improving its understanding of spatial relationships, attention dynamics, and visual context. The Blur Stitch method, designed to suppress irrelevant background regions, achieved an S_1 of 17.81 and S_2 of 33.41 with 49.01% accuracy, indicating moderate improvement. In contrast, the Arrow Stitch strategy, intended to encode motion through trajectory vectors between object centroids, reached an S_1 of 19.86, S_2 of 34.68, and 49.53% accuracy. Although this abstraction may affect spatial clarity, it assists reasoning about dynamic interactions such as pedestrian-vehicle motion, confirming that localization remains the most consistent cue across descriptive and inferential tasks.

For reference, a pretrained LLaVA-Video-7B baseline (no fine-tuning) evaluated on the same validation set achieved markedly lower performance, with an S_1 score of 11.03, S_2 of 13.30, and an accuracy of 22.75%, yielding a composite score of 15.69. Despite using full-frame video inputs, it also exhibited slower inference (0.37 fps) compared to our image-collage Box Stitch

model (2.2 fps). These results highlight the efficiency–accuracy trade-off and demonstrate that our approach achieves better reasoning performance while operating with significantly lower computational cost. Consistent trends observed across the WTS and BDD-PC5K validation sets further confirm that the proposed collage representation generalizes well beyond the AI City dataset.

No.	Models	S_1 [i]	S_2 [e]	Acc (%)	Score
1	Box Stitch	20.39	35.42	51.58	35.79
2	Blur Stitch	17.81	33.41	49.01	33.41
3	Arrow Stitch	19.86	34.68	49.53	34.69
4	Pre-trained LlaVA-Video-7B	11.03	13.3	22.75	15.69

Table 1: Performance of different approaches on the validation set.

Our proposed approach achieves a competitive result of 33.93 S_2 score on the global data on the AI City leaderboard, demonstrating its promising performance in describing and analyzing real-world traffic safety scenarios.

6 Conclusion

In this work, we proposed a modular vision-language framework for traffic safety captioning and question answering, addressing the need for interpretable and temporally aware analysis of pedestrian-vehicle interactions. Centered around structured prompting, stitched multi-frame inputs, and three distinct visual transformation strategies, our approach demonstrates that lightweight, image-based VLMs can effectively reason about dynamic traffic scenes without relying on expensive full video modeling. Among the evaluated methods, the Box Stitch approach proved most effective, achieving a final S_2 score of 33.93 by leveraging explicit bounding boxes for enhanced localization of critical scene elements and improved contextual understanding. In contrast, the Blur Stitch method focused model attention by suppressing background clutter, while the Arrow Stitch technique introduced motion-aware trajectory cues, each offering complementary insights into spatial focus, temporal reasoning, and efficiency. Overall, these results highlight a scalable and resource-efficient strategy for traffic scenario understanding and demonstrate that careful design of visual inputs and structured prompts can unlock the full potential of VLMs for safety-critical, interpretable, and adaptive traffic analysis.

Limitations and Qualitative Insights. While the proposed Box Stitch approach consistently outperforms other variants, certain limitations remain. Qualitative inspection shows that performance degrades when bounding boxes are missing or imprecise, leading to incomplete spatial context and weaker reasoning in crowded or fast-moving scenes. These errors primarily occur under occlusion, motion blur, or low illumination, where object detectors struggle to capture all agents accurately. Addressing these cases through improved detection stability and temporal association remains an important direction for future work.

Future Directions. Exploring hybrid models that dynamically combine blur, box, and motion cues depending on scene complexity could further enhance generalization. Integrating temporal transformers, uncertainty modeling, or reinforcement-based feedback may refine caption quality and decision accuracy. Additionally, expanding the question–answer bank using retrieval-augmented generation or domain-specific large language models could improve VQA robustness in complex real-world deployments. We are currently integrating the model into a smart-city dashboard for live captioning and VQA, bridging benchmark-stage research with real-world

traffic monitoring.

Acknowledgments

We thank the AI City Challenge organizers for providing access to the dataset used in this study. We also acknowledge the National Science Foundation (NSF) for providing computational resources and GPU access used for model training and inference.

References

[1] J. Chen et al. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. 2023. URL: https://arxiv.org/pdf/2310.09478.

[2] H. W. Chung et al. Scaling instruction-finetuned language models. 2022. URL: https://arxiv.org/pdf/2210.11416.

[3] Q. M. Dinh, M. K. Ho, A. Q. Dang, and H. P. Tran. Trafficvlm: A controllable visual language model for traffic video captioning. 2024. Accessed: Jun. 29, 2025. URL: https://arxiv.org/pdf/2404.09275.

[4] A. Dosovitskiy et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR 2021 - 9th International Conference on Learning Representations*, 2020. URL: https://arxiv.org/pdf/2010.11929.

[5] A. Vaswani et al. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[6] F. Yu et al. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2633–2642, 2018. doi:10.1109/CVPR42600.2020.00271.

[7] J. Bai et al. Qwen technical report. https://arxiv.org/pdf/2309.16609, 2023. Accessed: Jun. 29, 2025.

[8] K. Xu et al. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning (ICML)*, volume 3, pages 2048–2057, 2015. URL: https://arxiv.org/pdf/1502.03044.

[9] P. Anderson et al. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6077–6086, 2017. doi:10.1109/CVPR.2018.00636.

[10] S. Wang et al. The 8th ai city challenge. https://arxiv.org/pdf/2404.09432, 2024. Accessed: Jun. 29, 2025.

[11] S. Ge et al. Implicit location-caption alignment via complementary masking for weakly-supervised dense video captioning, 2024. Accessed: Jun. 29, 2025. URL: https://arxiv.org/pdf/2412.12791.

[12] L. Guo, J. Liu, X. Zhu, P. Yao, S. Lu, and H. Lu. Normalized and geometry-aware self-attention network for image captioning. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 10324–10333, 2020. doi:10.1109/CVPR42600.2020.01034.

[13] S. Herdade, A. Kappeler, K. Boakye, and J. Soares. Image captioning: Transforming objects into words. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[14] C. Jia et al. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the International Conference on Machine Learning*, volume 139, 2021. URL: https://proceedings.mlr.press/v139/jia21b.html.

[15] J. Li et al. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *Proc Mach Learn Res*, 202:20351–20383, 2023. URL: https://arxiv.org/pdf/2301.12597.

[16] J. Li, D. Li, C. Xiong, and S. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the International Conference on Machine Learning (PMLR)*, 2022. Accessed: Jun. 29, 2025. URL: https://proceedings.mlr.press/v162/li22n.html.

[17] L. Li et al. Lavender: Unifying video-language understanding as masked language modeling. pages 23119–23129, 2022. doi:10.1109/cvpr52729.2023.02214.

[18] Y. Li, S. Zhou, Z. Qin, and L. Wang. Pr-detr: Injecting position and relation prior for dense video captioning. 2025. Accessed: Jun. 29, 2025. URL: https://www.arxiv.org/pdf/2506.16082.

[19] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916, 2023. Accessed: Jun. 29, 2025. URL: https://llava-vl.github.io.

[20] Y. Luo et al. Dual-level collaborative transformer for image captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2286–2293, 2021. doi:10.1609/AAAI.V35I3.16328.

[21] S. Malla, C. Choi, I. Dwivedi, J. H. Choi, and J. Li. Drama: Joint risk localization and captioning in driving. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1043–1052, 2023. doi:10.1109/WACV56688.2023.00110.

[22] L. Ouyang et al. Training language models to follow instructions with human feedback. *Adv Neural Inf Process Syst*, 35, 2022. URL: https://arxiv.org/pdf/2203.02155.

[23] E. Qasemi, J. M. Francis, and A. Oltramari. Traffic-domain video question answering with automatic captioning. https://arxiv.org/pdf/2307.09636, 2023. Accessed: Jun. 29, 2025.

[24] A. Radford et al. Learning transferable visual models from natural language supervision. *Proc Mach Learn Res*, 139:8748–8763, 2021. URL: https://arxiv.org/pdf/2103.00020.

[25] O. September. GPT-4V(ision) System Card, 2023. https://openai.com/research/gpt-4v-system-card.

[26] M. Shoman, D. Wang, A. Aboah, and M. Abdel-Aty. Enhancing traffic safety with parallel dense video captioning for end-to-end event analysis, 2024. Accessed: Jun. 29, 2025. URL: https://arxiv.org/pdf/2404.08229.

[27] H. Touvron et al. Llama 2: Open foundation and fine-tuned chat models. 2023. URL: https://arxiv.org/pdf/2307.09288.

[28] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164, 2015. doi:10.1109/CVPR.2015.7298935.

[29] L. Wang et al. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2023-June, pages 14549–14560, 2023. doi:10.1109/CVPR52729.2023.01398.

[30] R. Wang et al. Bevt: Bert pretraining of video transformers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2022-June, pages 14713–14723, 2021. doi:10.1109/CVPR52688.2022.01432.

[31] C. Wei et al. Masked feature prediction for self-supervised visual pre-training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2022-June, pages 14648–14658, 2021. doi:10.1109/CVPR52688.2022.01426.

[32] A. Yang et al. Vid2seq: Large-scale pre-training of a visual language model for dense video captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2023-June, pages 10714–10726, 2023. doi:10.1109/CVPR52729.2023.01032.

[33] L. Yuan et al. Videoglue: Video general understanding evaluation of foundation models. 2023. URL: https://arxiv.org/pdf/2307.03166.

[34] R. Zellers et al. Merlot reserve: Neural script knowledge through vision and language and sound. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume

2022-June, pages 16354–16366, 2022. doi:10.1109/CVPR52688.2022.01589.

[35] P. Zhang et al. Vinvl: Revisiting visual representations in vision-language models, 2021. Accessed: Jun. 29, 2025. URL: https://github.com/pzzhang/VinVL.

[36] Q. Zhang, Y. Song, and Q. Jin. Unifying event detection and captioning as sequence generation via pre-training. In *Lecture Notes in Computer Science*, volume 13696, pages 363–379, 2022. doi:10.1007/978-3-031-20059-5_21.

[37] X. Zhou et al. Streaming dense video captioning. In *CVPR*, 2024. doi:10.1109/CVPR52733.2024.01727.

[38] W. Zhu, B. Pang, A. V. Thapliyal, W. Y. Wang, and R. Soricut. End-to-end dense video captioning as sequence generation. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, volume 29, pages 5651–5665, 2022. Accessed: Jun. 29, 2025. URL: https://arxiv.org/pdf/2204.08121.