SA-CLIP: Language Guided Image Spatial and Action Feature Learning

Guanlin Li[1]*, Wenhao Shao[2]*, Praboda Rajapaksha[1,3], Noël Crespi[1]
[1]Samovar, Telecom SudParis, Institut Polytechnique de Paris, France
[2]School of Computer, Electronics and Information, Guangxi University, China
[3]Department of Computer Science, Aberystwyth University, UK
{guanlin_li, noel.crespi}@telecom-sudparis.eu
wenhao.shao@gxu.edu.cn, prr16@aber.ac.uk

Abstract

We observed that Contrastive Language-Image Pretraining (CLIP) models struggle with real-world downstream tasks such as road traffic anomaly detection, due to their inability to effectively capture spatial and action relationships between objects within images. To address this, we propose a dependency parsing based method to compile and curate a dataset with 1M samples of images using language supervision provided by the common image caption dataset, in which each image is paired with subject-relationship-object descriptions emphasizing spatial and action interactions, and train a **S**patial and **A**ction relationship aware **CLIP** (**SA-CLIP**) model. We evaluated the proposed model on the Visual Spatial Reasoning (VSR) dataset and further verified its effectiveness on the Detection-of-Traffic-Anomaly (DoTA) dataset. Experiment results show that the proposed SA-CLIP demonstrates strong abilities in understanding spatial relationships while achieving good zero-shot performance on the traffic anomaly detection task.

1 Introduction

Vision-language models have demonstrated strong potential in real-world tasks and in producing explainable results (Lv and Sun, 2024; Zanella et al., 2024; Gu et al., 2024), among which the CLIP (Radford et al., 2021) based models gain particular attention (Pan et al., 2022; Wang et al., 2023; Fan et al., 2024). CLIP is pretrained to align image and text representations by minimizing their distance, and shows strong generalization ability for real-world image-text matching tasks. Such ability makes CLIP particularly attractive for tasks requiring a high degree of interpretability, as it allows the model to explain its decisions in human-understandable terms by linking specific visual elements with descriptive language, while being computationally effective, relying only on similarity computation at inference time.

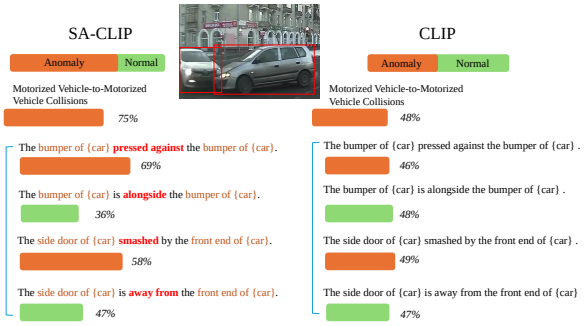


Figure 1: Comparison of image-text matching scores between SA-CLIP and CLIP in the traffic anomaly scene. SA-CLIP focuses on the objects and relationships between the objects both in image and text, and achieves better anomaly detection ability than the vanilla CLIP.

However, despite its strengths, the CLIP model shows limitations in complex scenarios. We observed that CLIP model is unable to distinguish between the normal and abnormal scenes through textual descriptions, as shown in the right part of Fig. 1. We hypothesize and confirm through experiments that the CLIP model is unable to accurately capture and reason about the spatial and action relationships between objects both in images and those described in texts, which are often critical in such real-world scenarios. For example, the relative positions of vehicles, pedestrians, and road signs could determine whether a situation is normal or potentially dangerous. Without a deep understanding of these spatial relationships and the action interaction between the objects, the model may misclassify events, leading to false positives or missed detections. To address this challenge, in this paper we propose to enhance the CLIP model's ability for spatial and action relationship understanding in both image scenes and textual descriptions. We first curate a dataset that focuses on the main objects and their spatial and action relationships

*Equal contribution.

in the image scenes, using the language description provided by the caption of the images, based on the existing large-scale image-caption datasets. We exploit the linguistic patterns and develop a rule-based method to extract subject-relationship-object triplets from the captions, and train a relation classification model to filter out the low-quality triplets. Leveraging the compiled dataset, we introduce **S**patial-**A**ction **CLIP** (**SA-CLIP**), which explicitly models the spatial and action relationships between objects in the images by learning to map the subject-relationship-object triplet in texts to the objects and their positions in images. Experiment results on the VSR dataset (Liu et al., 2023) demonstrate the effectiveness of the model in modelling spatial relationships. We further verify the model's ability to address real-world tasks by evaluating on the DoTA dataset (Yao et al., 2022) for the road traffic anomaly detection task.

2 Methodology

To improve CLIP's ability to understand spatial and action relationships among objects in scenes, we focus on the Subject-Verb-Object (SVO) structure in both image descriptions and scene graphs. The SVO structure is a fundamental linguistic pattern that describes interactions (verbs) between subjects and objects, and we extend the verbal references to include prepositional phrases (PPs) that indicate spatial relationships. We train CLIP to map the SVO structures from language to visual representations. Due to the scarcity of datasets with diverse and detailed spatial and action annotations, we mine such information from abundant image-caption datasets and pretrain the proposed SA-CLIP on the curated data.

2.1 Dataset Curation

The approach follows a two-step process. First, we apply a rule-based parser to extract SVO triplets from a seed subset of image-caption data, sampled from cleaner sources such as MSCOCO (Lin et al., 2014) to reduce noise. The parser extracts subjects, objects, and associated verbal or prepositional phrases to form initial triplets. In the second step, we filter noisy triplets. The wrongly extracted triplets usually exhibit significant semantic inconsistencies between subjects and objects and their relations, which can be easily identified by LLMs. Therefore, we leverage LLMs to annotate the seed data and train a lightweight relation classifier on these annotations to identify and retain high-quality triplets. The overall process is illustrated in Fig. 2.

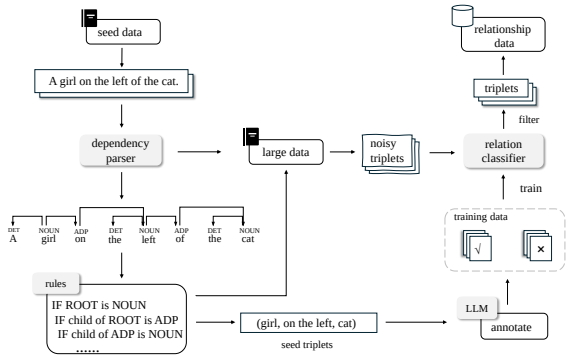


Figure 2: Dataset curation procedure.

Following the process, we compiled a dataset containing around 970,000 samples, filtered from MSCOCO (Lin et al., 2014), SBUCaption (Ordonez et al., 2011), CC3M (Sharma et al., 2018), CC12M (Changpinyo et al., 2021) and Visual Genome (VG) (Krishna et al., 2017).

2.2 Model Architecture

The proposed model employs a dual image text encoder architecture that incorporates both textual and visual inputs, which is illustrated in Fig.3. Specifically, a pretrained vision model is employed to encode image inputs and extract visual features, while a text encoder processes textual inputs. For the model to be aware of objects and their relationships in the scene, the vision model also takes object regions as inputs, which are extracted based on the triplet in the textual description, while the text encoder processes the SVO features. The SVO structure in the textual inputs is then mapped to the object interactions in the image through contrastive learning. We elaborate on each component below.

2.2.1 Textual Features

The textual input to the model consists of a sentence describing the scene and an extracted SVO triplet. To generate textual representations that emphasize the objects and their relationships, the SVO structure, composed of the subject s, relation v and object o, is embedded together with the input sentence. Using the positional information of the tokens corresponding to the subject idx_s, relationship idx_v, and object idx_o within the sentence, the model outputs individual representations for each part of the triplet, which are then aggregated using mean pooling followed by a linear projection to produce a unified representation of the SVO

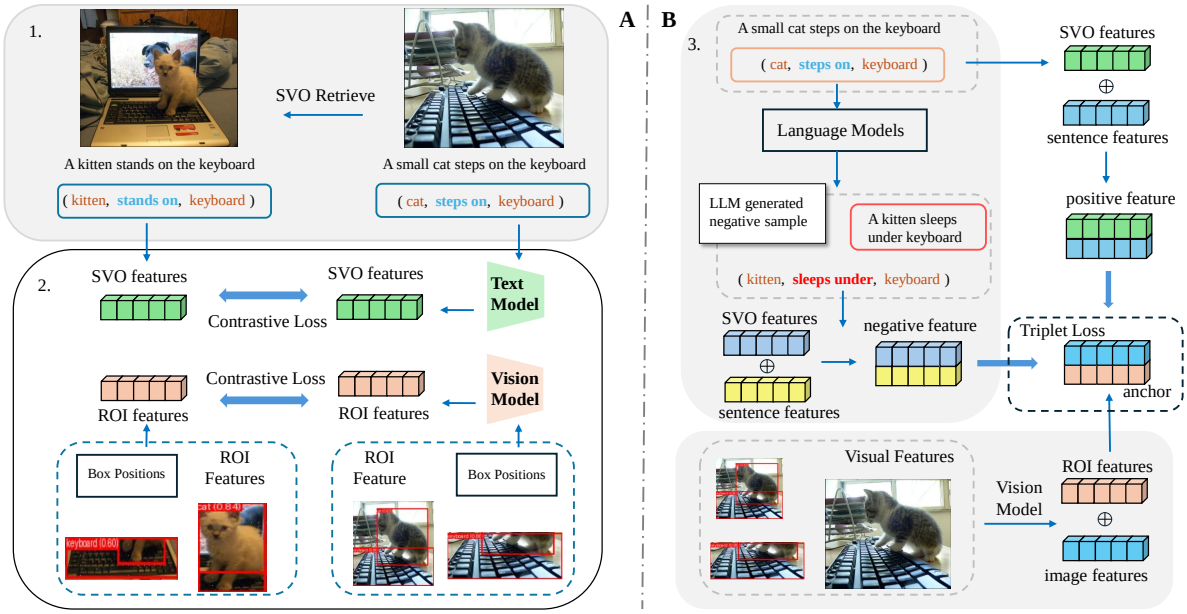


Figure 3: The overall architecture of the proposed method. Part 1 shows positive sample construction process using SVO textual similarity; part 2 shows inter-sample contrastive loss between positive samples; part 3 shows negative sample construction for triplet loss. Altogether, part A shows inter-sample training process, and part B shows intra-sample training.

relationship within the scene, denoted by r_{svo}, as shown in following equations and the upper part of part 2 in Fig.3.

$$r_{\text{sent}} = f_u(\text{sent}) \tag{1}$$

$$r_s, r_v, r_o = r_{\text{sent}}[idx_s, idx_v, idx_o] \tag{2}$$

$$\hat{r}_{svo} = \text{MeanPooling}(r_s, r_v, r_o) \tag{3}$$

$$r_{svo} = W\hat{r}_{svo} + b \tag{4}$$

Where f_u denotes the text model and sent refers to the sentence description. The textual representation is the concatenated representation of the sentence features and the SVO features, as shown in Eq.(5) and the upper right of part B in Fig.3.

$$u = \text{Concat}(\text{Pool}(r_{\text{sent}}), r_{svo}) \tag{5}$$

2.2.2 Visual Features

For each image input, the Regions of Interest (ROI) corresponding to the subject and object in the image are first extracted, based on the object and subject in the SVO structure of the textual description. Each ROI is then encoded by the vision model f_v to obtain an ROI feature. Similar to (Tan and Bansal, 2019), positional information is incorporated into the ROI features by combining the embeddings of their box position, as described in Eq.(6), Eq.(7) and lower part of part 2 in Fig.3.

$$\hat{r}_{roi} = f_v(ROI) + \text{Embed}(Position_{ROI}) \tag{6}$$

$$r_{roi} = W\hat{r}_{roi} + b \tag{7}$$

The model's overall image representation is the concatenated representation of the image features and the ROI features, described in Eq.(8) and the lower part of part B in Fig.3.

$$v = \text{Concat}(f_v(image), r_{roi}) \tag{8}$$

2.3 Training Objectives

2.3.1 Inter-sample Contrastive Learning

We employ inter-sample contrastive learning to improve the model's representation of SVO structures in both textual and visual spaces. The inter-sample contrastive learning aims to pull the embedding of similar SVO structures in different samples close in the representation space, improving the model's ability to recognize similar object relationships in various scenarios. The inter-sample contrastive loss is shown in Eq.(9) and described in part 2 of Fig.3.

$$\mathcal{L}_u = -\sum_{i=1}^{N}\sum_{j=1}^{N} \log \frac{\exp(\text{sim}(u_i, u_j)/\tau)}{\sum_{k=1}^{2N-1} \exp(\text{sim}(u_i, u_k)/\tau)} \tag{9}$$

Where u is the textual features described in Eq.(5). The inter-sample loss for the visual features is calculated similarly.

2.3.2 Intra-sample Contrastive Learning

To maintain cross-modal alignment during inter-sample contrastive learning, we adopt an intra-

sample contrastive loss. Instead of using an in-batch contrastive loss with the sample's caption as the positive and other captions in batch as negatives, we incorporate a triplet loss to allow the model to better distinguish between different relationships of the same objects in the scene. The triplet loss is shown in Eq.(10) and indicated in part B of Fig.3.

$$\mathcal{L}_{\text{triplet}} = \max\left(0, \|v-u\|_2^2 - \|v-u'\|_2^2 + \alpha\right) \quad (10)$$

where v is the visual features of the sample, u is the textual feature of the sample's caption, u' is the textual feature of the sample's negative feature. The overall loss for a sample is a linear addition of $\mathcal{L}_u$, $\mathcal{L}_v$ and $\mathcal{L}_{\text{triplet}}$.

2.3.3 Positive and Negative Sample Construction

In inter-sample learning, positive pairs are image samples retrieved based on the textual similarity of their SVO triplets; negative samples are other in-batch samples. For the triplet loss in intra-sample learning, the image feature is used as the anchor, the positive sample is its corresponding text, while the negative is a textual description with similar objects and scenes but an opposite relationship. To obtain negative samples that satisfy this constraint, simply retrieving from the dataset is often insufficient. Therefore, given the textual description of the anchor image and its SVO triplet, we use LLM to generate an opposing description that retains the same subject and object but conveys an inverse relationship. The opposite description together with its triplet are then used as the negative sample.

3 Experiments

3.1 Datasets

CLEVR Dataset (Johnson et al., 2017): To test our hypothesis that CLIP struggles with spatial reasoning in both text and image, we use a controlled dataset inspired by CLEVR. We generate scenes and textual descriptions containing two objects of controlled visual features and consistent shapes (cube and sphere), placed in one of eight spatial configurations: front, behind, left, right, front-left, front-right, behind-left, and behind-right. We input the image-text pairs into the model and visualize the resulting embeddings using t-SNE (Van der Maaten and Hinton, 2008). Sample images are shown in Fig.4b.

VSR Dataset (Liu et al., 2023): To further examine the proposed model's spatial understanding ability quantitatively, we use the Visual Spatial Reasoning (VSR) dataset, which includes 6,940 images and 66 distinct spatial relations. The task is to classify whether a caption correctly describes the spatial arrangement in the image. We evaluate the model on the **zero-shot test set** (1,222 samples) using accuracy as the metric.

DoTA (Yao et al., 2022): For real-world evaluation, we test the model's anomaly detection ability on road traffic videos from the DoTA dataset, focusing on non-ego (non-self-induced) anomalies. From the test set, 597 relevant samples are selected and grouped into three anomaly interaction types: vehicle-vehicle (VV), vehicle-person (VP), and vehicle-obstacle (VO). ROC-AUC is used as the evaluation metric.

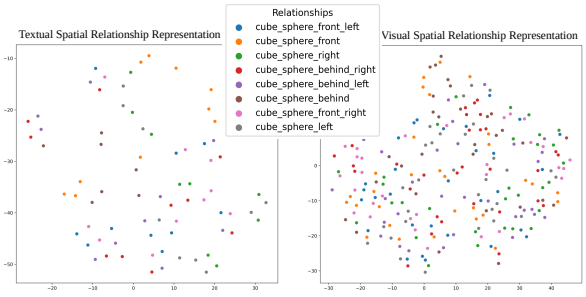
Model	Params	Accuracy
Finetuned		
VisualBERT (Li et al., 2019)	111 M	51.0
ViLT (Kim et al., 2021)	111 M	63.0
LXMERT (Tan and Bansal, 2019)	208 M	61.2
Full zero-shot		
CLIP-ViT-H-14 (w/ prompting)	1 B	54.5
SA-CLIP	151 M	57.8

Table 1: Evaluation results on the VSR **zero-shot split**.

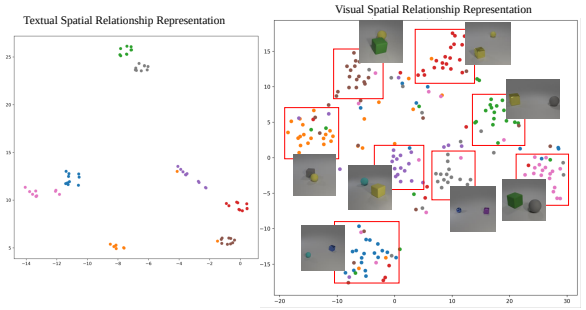
3.2 Evaluation Results

Spatial Understanding Ability In CLEVR scenes, where the visual features of the objects are controlled, the only distinguishing factor is their spatial relationship. Therefore, the degree to which a model forms distinct clusters in its representations reflects its ability to capture spatial differences. As shown in Fig.4, CLIP fails to distinguish between these relationships. In contrast, SA-CLIP, pretrained on the curated dataset, effectively captures spatial distinctions, producing clear clusters in both visual and textual embeddings. This is further supported by results on the VSR test set in Table 1, where SA-CLIP surpasses CLIP-ViT-H-14 in zero-shot performance without careful prompting and with significantly fewer parameters, and performs comparably to fully finetuned baselines. The contributions of each model component are evaluated through an ablation study, as shown in Table 3 in Appendix B.

Traffic Anomaly Detection Results The evaluation results on the DoTA testset are shown in Table 2. The proposed model achieves superior zero-shot



(a) Visualization results for the CLIP model.



(b) Visualization results for the SA-CLIP model.

Figure 4: The visualization of spatial relationships in the generated CLEVR scenes.

Methods	**VV**	**VP**	**VO**
Finetuned			
AnoPred (Liu et al., 2018)	64.9	64.9	64.2
AnoPred +Mask (Liu et al., 2018)	66.0	64.0	58.8
FOL-STD (Yao et al., 2019)	70.8	69.7	63.8
FOL-Ensemble (Yao et al., 2022)	**73.2**	71.2	65.2
TTHF (Liang et al., 2024)	71.3	64.3	**69.9**
SA-CLIP	71.6	**72.1**	58.0
Zero-shot			
CLIP-ViT-B-32 (Radford et al., 2021)	49.7	49.8	50.0
SA-CLIP	**56.8**	**58.2**	**52.1**

Table 2: AUC on the DoTA testset. Best performance is marked in **bold**; second-best is underlined.

anomaly detection ability over the vanilla CLIP model and remains competitive with SOTAs. Notably, it performs well in detecting VP anomalies but worse in VO cases. Further analysis reveals that this is due to the reliance on the object detector, as the model's performance is closely related to the quality of region-of-interest.

4 Conclusion

In this work, we address CLIP's limitations in spatial and action understanding by introducing SA-CLIP, pretrained on a curated dataset focused on these relationships in both images and text. SA-CLIP outperforms vanilla CLIP in spatial action understanding and also proves effective in the real-world task of road traffic anomaly detection.

Limitations

This work has three main limitations. First, due to the contrastive training objective, the model lacks generative capabilities and thus relies on predefined textual knowledge in downstream tasks. While this limits flexibility, it also enables efficient inference. Second, it remains unclear whether the model truly learns to understand the semantic relationships between regions of interest in the images and the corresponding textual SVO triplets, or if it simply relies on mapping representations of seen instances for inference. We plan to conduct a qualitative study on the curated data to enable a more thorough understanding of the model's behavior. Finally, the model heavily depends on an off-the-shelf object detector to identify subjects and objects, making its performance sensitive to detection accuracy.

References

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3558–3568.

Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. 2024. Improving clip training with language rewrites. *Advances in Neural Information Processing Systems*, 36.

Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. 2024. Anomalygpt: Detecting industrial anomalies using large vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 1932–1940.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, pages 5583–5594. PMLR.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al.

2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Rongqin Liang, Yuanman Li, Jiantao Zhou, and Xia Li. 2024. Text-driven traffic anomaly detection with temporal high-frequency modeling in driving videos. *IEEE Transactions on Circuits and Systems for Video Technology*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Fangyu Liu, Guy Emerson, and Nigel Collier. 2023. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651.

Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. 2018. Future frame prediction for anomaly detection–a new baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6536–6545.

Hui Lv and Qianru Sun. 2024. Video anomaly detection and explanation via large language models. *arXiv preprint arXiv:2401.05702*.

Vicente Ordonez, Girish Kulkarni, and Tamara Berg. 2011. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24.

Xuran Pan, Tianzhu Ye, Dongchen Han, Shiji Song, and Gao Huang. 2022. Contrastive language-image pre-training with knowledge graphs. *Advances in Neural Information Processing Systems*, 35:22895–22910.

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

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. 2023. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 2555–2563.

Yu Yao, Xizi Wang, Mingze Xu, Zelin Pu, Yuchen Wang, Ella Atkins, and David Crandall. 2022. Dota: unsupervised detection of traffic anomaly in driving videos. *IEEE transactions on pattern analysis and machine intelligence*.

Yu Yao, Mingze Xu, Yuchen Wang, David J Crandall, and Ella M Atkins. 2019. Unsupervised traffic accident detection in first-person videos. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 273–280. IEEE.

Luca Zanella, Willi Menapace, Massimiliano Mancini, Yiming Wang, and Elisa Ricci. 2024. Harnessing large language models for training-free video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18527–18536.

A Training Details

Dataset Curation We used spaCy[1] to get the dependency tree from the image captions. To extract spatial and action relationships, we composed rules focusing on the prepositional and verbal phrases. The rules start by checking the POS tag of the root node and its child nodes to determine the subject for the sentence, and recursively looking for POS patterns that satisfy the desired phrases from the root node. We selected captions from the MSCOCO dataset as seed data to obtain seed triplets, which contain less noise compared to web-collected data. The seeds were annotated by GPT-4o[2] to obtain binary labels for the parsing correctness, and a binary classifier was trained on the annotated data using bert-base-uncased model which was then used to filter high-quality extractions.

Training Settings We initialize the weights of SA-CLIP text and vision models using CLIP-ViT-B-32. During training, we adopt the locked tuning strategy and freeze the vision model. We use a learning rate of 5e-4 with AdamW optimizer and set batch size to 512, which corresponds to 1,024

[1] https://spacy.io/

[2] https://openai.com/index/hello-gpt-4o/

image-text pairs in each step, and train the model for 30,000 steps on the curated dataset using one NVIDIA A100 GPU with 80GB memory. The pretrained SA-CLIP was then used in the following downstream tasks and evaluations.

B Ablation Details

We evaluate the contribution of model components to relationship understanding using the VSR dataset and examine the impact of pretraining on the curated dataset for anomaly detection. The results are shown in Table 3. SVO indicates inter-sample training using textual SVO triplets; ROI indicates inter-sample training using visual ROI features; Triplet loss indicates intra-sample training using both textual and visual features.

Pretrain	SVO	ROI	Triplet Loss	Accuracy
w/o	✗	✓	✓	53.5
w/o	✓	✗	✓	54.5
w/o	✗	✗	✓	53.1
w/o	✓	✓	✗	51.1
w/o	✓	✓	✓	54.7
w/ zero-shot	✓	✓	✓	57.8
w/ + finetune	✓	✓	✓	57.9

Table 3: Ablation results on VSR dataset. w/o denotes without pretraining on the curated relationship dataset; finetune denotes training on the target dataset.

Table 3 shows that incorporating inter-sample contrastive learning improves the model's spatial relationship understanding by improving the accuracy score on the VSR dataset compared to only using triplet loss. However, without the triplet loss, the model could not match texts with images, leading to a drastic drop of accuracy.

C Traffic Anomaly Detection Evaluation Details

We use a text-driven method to detect anomalies based on the SA-CLIP model. Specifically, we first summarize the domain knowledge into a comprehensive hierarchy of traffic anomaly descriptions, under the assistance of domain experts.

Then, we compute the similarity between the textual features from the anomaly description and the visual features of the frame to be evaluated using the model to decide if a frame contains anomaly. Specifically, given the structured description D of traffic anomaly, the anomaly scenes D_a and normal scenes D_b are represented by the subcategories in the hierarchical knowledge, and the binary classification problem for the anomaly detection task is decomposed into matching the visual features of a frame to the textual features of the structure node describing the abnormal and normal scenes. The final anomaly score for a frame f_n is calculated through the additive decomposition over all the descriptions as follows:

$$score(f_n) = \frac{1}{|D(a)|} \sum_{d \in D(a)} (sim(u_d, \hat{z}_n)) \quad (11)$$

Where d is a description from abnormal scene descriptions, and u is the textual representation of the description. By inspecting the similarity score between the frame and each description, descriptions that best match the scene can be identified, thereby providing insight into the model's decision for anomaly.