# Unseen And Adverse Outdoor Scenes Recognition Through Event-based Captions

Hidetomo Sakaino
Visual Recognition Group, Weather Transportation Lab., Weathernews Inc.
sakain@wni.com

## Abstract

*This paper presents EventCAP, i.e., event-based captions, for refined and enriched qualitative and quantitative captions by Deep Learning (DL) models and Vision Language Models (VLMs) with different tasks in a complementary manner. Indoor and outdoor images are used for object recognition and captioning. However, outdoor images in events change in wide ranges due to natural phenomena, i.e., weather changes. Such dynamical changes may degrade segmentation by illumination and object shape changes. This increases unseen objects and scenes under such adverse conditions. On the other hand, single state-of-art (SOTA) DLs and VLMs work with single or limited tasks, Therefore, this paper proposes EventCAP with captions with physical scales and objects' surface properties. Moreover, an iterative VQA model is proposed to refine incomplete segmented images with the prompts. A higher semantic level in captions for real-world scene descriptions is experimentally shown compared to SOTA VLMs.*

## 1. Introduction

Segmentation has become an important task for real-world applications by Image Processing, and Computer Vision (CV), Deep Learning (DL) [26, 25, 44, 8, 5, 40, 13, 20, 22]. Since a number of cameras are increasingly implemented everywhere, segmented objects, i.e., things and stuff, are used for multi-purposes, i.e., surveillance, auto-driving, navigation, and mobile phone. Unlike indoor uses, more robustness and stability are required to segmentation models for outdoor uses. In outdoor scenes, objects, time-varying illumination, i.e., sunbeams, noise, and natural phenomena, i.e., weather conditions, are usually unknown. Moreover, normal events, i.e., traffic accidents and disaster events, are unpredictable. We call these dynamic changes. Elements of such dynamic changes have been dealt with by state-of-the-art (SOTA) methods, i.e., Derain, De-raindrops [29, 46], Defog, and Dehaze [14, 21, 45, 9, 27]; however, most element removers for rain streaks, fog, and snowfall are mainly effective in synthetic images. As shown in [33, 34], such SOTAs fail to deal with real heavy fog and snowfall events. In response to these events, road conditions are also impacted, showing dry, wet, and snow. Moreover, unpredicted disaster and traffic accident scenes can happen.

Human drivers have to pay attention to road scenes with various adverse conditions. On the other hand, auto-driving relies heavily on cameras and sensors, where segmentation plays an important role in visually determining the best routes. However, DL-based SOTA segmentation models are insufficient to recognize them, i.e., part of objects or no objects. Recently, Vision Language Models (VLMs) [1, 6] have come to improve previous performance of DL-based SOTA segmentation models [41, 4, 49]. It is known that unseen images that have not been pretrained have been recognized much better than only CV or DL models. VLMs are pre-trained on large datasets [30, 28, 19]. Since it is hard to train VLMs from scratch, fine-tunings on smaller datasets have been conducted. By this, VLMs can be utilized to save time and resources in various applications.

In VLMs, diverse and out-of-distribution data for pre-training and evaluation are used [10]. Prompt learning to adapt VLMs to new tasks without fine-tuning is also shown [12]. Contents of captions have been enhanced for better descriptions of real-world objects [6]. However, single VLMs are often insufficient for dynamic changes, i.e., disaster scenes, [39] even by fine-tunings. The main reason for this difficulty is due to camera image-based post-disaster object recognition for dirt, water, and rocks. Due to heavy rainfall and snowfall, traffic accidents are also caused on roads. Therefore, uncountable unseen objects can appear on roads. Domain adaptation segmentation [42, 11] may be another approach to cope with such post-disaster scenes. However, it requires a manual selection of the optimal pre-trained model.

Many efficient VL models without retraining for unseen images have been introduced [1, 6, 16, 18, 37, 38, 51, 30, 36, 35]. However, laborious and time-consuming tasks remain unsolved in pretraining VLMs.

Another approach would be to apply Visual ChatGPT [43]. It can produce acceptable results on the general scene and unseen classes. However, since Visual ChatGPT [43] has been trained on the limited data of the year 2021, it generates captions under older datasets. So far, Visual ChatGPT [43] is weak at generating dynamic scene descriptions like weather and road conditions.

In order to better understand post-disaster and traffic accident scenes, captions from SOTA VLMs cannot describe complicated scene changes only by the combination of segmented objects. Since the natural phenomenon is impacted, physical scales cannot be ignored. For example, physical scales are helpful for rescuing people and recovering damaged regions. Geometric reasoning or depth estimation to infer 3-D information from 2-D images [47, 48] is shown using 3D point-cloud data and indoor scenes. However, few papers report captions with physical scales in outdoor scenes.

To this end, this paper proposes EventCAP with complementary DLs and VLMs under adverse conditions using single images. EventCAP consists of seven modules, i.e., Deep Visual Language Classification (Dvlc), Deep Visual Language Segmentation (Dvls), Deep Road conditions (Droad), Deep anomaly (Danomal), Deep snowfall (Dsnow), and VQA. The branched architecture allows us to maintain and upgrade each of multiple modules efficiently. Contributions of this paper are fourfold:

1. Multiple vision language and Transformer-based Deep Learning (DL) models with branched structures for efficiency in light of memory, training, and maintenance. Danomal excludes difficult images, i.e., lens reflection, to stabilize the overall system. Due to enormous datasets of VLMs, Dvls, and Dvlc are fine-tuned VLMs from SOTA models for segmentation and classification, respectively.

2. Refined and enriched captions are generated from single images. It is the first time for such captions to contain dynamic changes under adverse weather conditions, i.e., weather conditions by Dsnow, and road conditions by Droad. Unseen images like adversarial weather and disaster conditions can be dealt with. Moreover, more specified scene descriptions of disaster events are shown.

3. Iterative VQA is proposed to enhance answers and segmented images as compared with one-time VQA. This looping at VQA is effective whenever adverse images are used.

4. Many experimental results show the superiority of the proposed EventCAP over SOTA DL models and VLMs. The proposed EventCAP will help notify specified scene descriptions, i.e., more quantitative texts, to drivers, auto-driving, and rescue workers from camera images.

## 2. Proposed Method

This section describes the proposed EventCAP method/system to refine and enrich captioning and classes from a single image input. In particular, this paper introduces a dynamic caption by a physical scale that cannot be pre-trained in a vision-language model.

To realize this, SOTAs in segmentation and vision-language models face their limits. Therefore, instead of using only vision models or a single vision-language model, this paper proposes a new architecture that integrates multiple Deep Learning and vision-language modules. Figure 1 shows an overview of the proposed EventCAP.

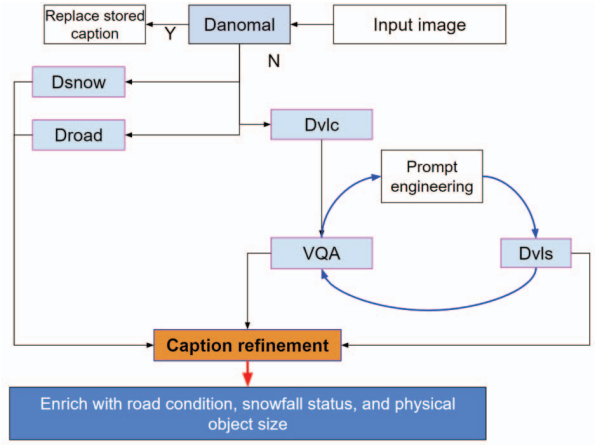



Figure 1. Overview of the proposed EventCAP model.

Since this paper deals with many challenging scenes with disasters and car accidents, adversarial conditions are considered. And a Danomal-like DeepReject in [34, 33, 31] is proposed to avoid the degradation of the cascaded other recognition modules. Further detailed explanations of the multiple modules will be given in Sections 2.1 to 2.4.

### 2.1. Proposed Dvlc and Dvls

Dvlc is a vision-language model trained on image and text pairs that can predict the most relevant text given an image. It does not need to be directly optimized for this task and can perform "zero-shot" learning like GPT-2 and -3. Dvlc matches the performance of the original ResNet50 on ImageNet "zero-shot" without using any of the original 1.28M labeled examples, which is a significant accomplishment in Computer Vision.

Dvlc utilizes the input texts of five distinct disaster categories: car crashes, flooding, fog, landslide, and rain. Tailored textual input descriptions are employed for each disaster category to enhance natural language processing techniques in analyzing disaster-related data. These scenes are

associated with domain-specific terms such as pedestrian, airplane, debris flow, and eruption to improve the accuracy of automated disaster detection and classification.

Dvls is proposed to obtain semantic segmentation of these scenes. Dvls is finetuned from OvSeg [23] by adding a new physical constraint to the loss function. To obtain descriptions of disasters for the Dvlc, a classification task is performed using keywords corresponding to each disaster scene. These texts are used to generate text descriptions of the disasters that are fixed for each type of scene.

Therefore, since Dvlc and Dvls recognize texts and segmented objects from a single image, this paper proposes to combine respective outputs.

### 2.2. Droad and Proposed Dsnow

This section discusses Droad and proposed Dsnow. Unlike SOTA papers in DL models and VLMs, this paper aims to generate dynamic scene changes with the weather conditions, i.e., rain, snow, and fog, and road conditions, i.e., dry, wet, and snow. Droad [33] is applied for further detailed classes of segmented objects. Dscene [33, 34] is also applied to ensure snow conditions.

In Droad [33], and Dscene [33, 34, 32], Swinformer [25] is trained from over 7500 winter road images. It is noted that since publicly available annotation datasets are insufficient, various weather and road scenes from different countries under adversarial conditions have been collected and used to train. The proposed Dsnow employs a transformer-based classifier trained on images captured during adverse weather conditions to estimate the level of snowfall. Our experts captured and labeled all the images used in the aforementioned DL models.

### 2.3. Proposed Iterative Caption Refinement

This section describes the VQA loop, which refines captions based on segmentation results from a VLM. As shown in Figure 2, for the same event, using different prompts leads to different segmentation results. Therefore, looping through all prompts from pre-defined Dvlc output is needed. The process of the loop between VQA and Dvls is depicted in Figure 3. The prompts of objects/events are from a list of synonym words. Among these words, the answer with the highest cosine similarity score is selected. Figure 3 shows an example with a step-by-step of the VQA loop process. The templates used is *"Are there O?"*. Where *O* are objects in the scene.

### 2.4. Caption Refinement

The caption refinement process involves utilizing a large language model (LLM) which incorporates the segmentation outcomes from Dvls and the captions generated by VQA. The output of Dvls comprises semantic segmentation along with corresponding locations and descriptions,

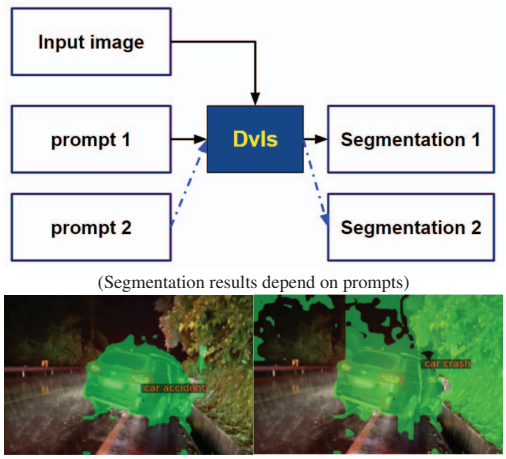



Figure 2. Results of proposed Dvls with different prompt: (a) "car accident" . (b) "car crash"

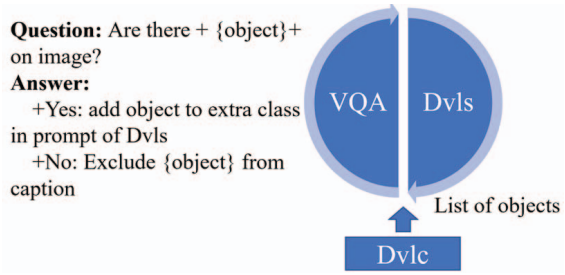



Figure 3. Overview of VQA loop.

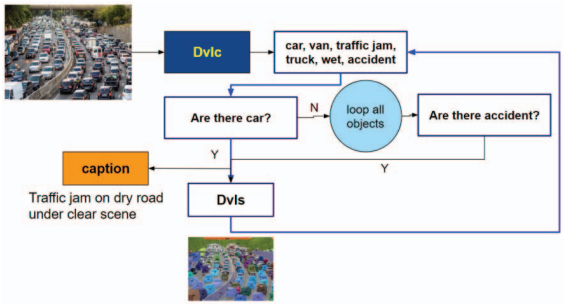



Figure 4. Query loop VQA and segmentation.

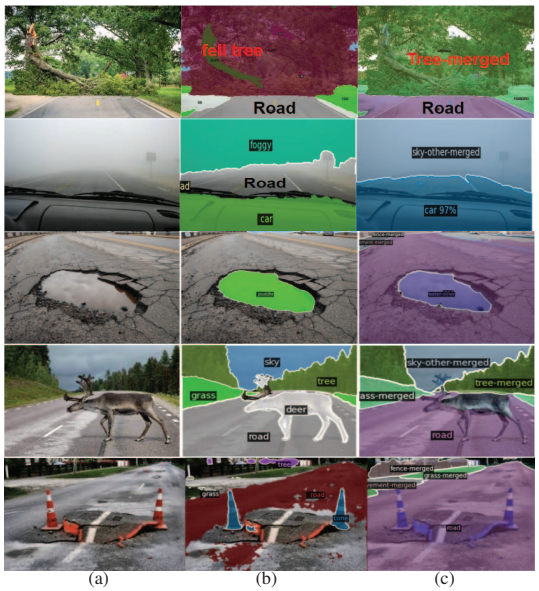
expressed in a language-based segmentation format as a list of {object description: bounding box of the object in pixels}. VQA contributes additional descriptions that capture the overall dynamic conditions, including adverse weather conditions, to provide contextual information for the LLM. The final result of caption refinement is an enriched caption that encompasses information about road conditions, and object size. The LLM used in caption refinement can determine the physical size of a standard object based on the segmentation results, even with just the object's name.

## 3. Experiments and Discussion

### 3.1. Refined Semantic Segmentation by Prompt Engineering

This section denotes the proposed Dvls and how to obtain the final refined captions using prompt engineering. The prompt for each scene is pre-defined as a list of words, i.e., (1) **car crashes**: ["pedestrian","car", "car crash", "road", "bike", "tree"]; (2) **flooding**: ["water", "car", "person", "tree", "sky"]; (3) **fog**: ["foggy", "mountain", "road", "car", "wet"]; (4) **landslide**: ["landslide", "debris flow", "rocks", "road", "dirt"]; (5) **rain**: ["water", "rain", "umbrella", "road", "person"]. Prompts for Dvls model are selected based on the classification results from Dvlc and the aforementioned pre-defined texts.

Figure 5 illustrates the effectiveness of our approach on images with foggy and traffic accident scenes. (a) shows the input images, while (c) displays the segmentation results generated by the transformer-based SOTA segmentation model, i.e., Mask2former [2], which shows generic classes, i.e., "sky-other-merged", and "water". (b) presents improved segmentation results and achieved prompt engineering, which provides more detailed semantic segmentation results, i.e., more detail from sky-other-merged" to "foggy" for the foggy scene and from "tree-merged" to "fell tree" for the disaster scene. It has been demonstrated that prompt tuning for Dvlc is helpful for achieving precise segmentation results under dynamic conditions.




(a) (b) (c)

Figure 5. Results of segmentation by SOTA and proposed EventCAP: (a) Original image. (b) Proposed refined semantic segmentation. (c) Mask2Former [2].

### 3.2. Dynamic Captions with Weather and Road Conditions by Droad and Proposed Dsnow

This section conducts experiments to create more intricate captions by considering various weather conditions along with traffic and disaster scenes. The comparison involves the utilization of the proposed Dsnow and Droad models in contrast with a SOTA VL captioning model, BLIP [17].

Figure 6 shows six scenes. As a result, road conditions by Droad (1)-(6) are wet in blue and snow in yellow. Dsnow's indicators (3)-(6) present light to heavy snowfall. Dvlc recognizes overall scene objects like mountains, rivers, rocks, sky, and trees. Therefore, the road condition and visibility distance have been included in the captions of Dvlc.

Table 1 shows a comparison of the refined captions and a SOTA BLIP [17] result using six scenes of Figure 6. The comparison results show that a refined caption is detailed about the scene by adding road conditions, snowfall status, location of objects, and exact visibility in meters. Besides, the caption from BLIP lacks a description. The result has proven that the proposed method integrating Droad, and Dsnow outperforms single VLM, i.e., BLIP [17].

Table 1. Comparison between the enhanced captions and BLIP's caption result.

| | Proposed method | BLIP [17] |
|---|---|---|
| (1) | muddy road with fallen trees without snowfall | a fallen tree sitting on top of a muddy road |
| (2) | Two people are crossing street under snowfall | two people walking on the snow covered road |
| (3) | cars on snow road under light snowfall | a car is driving down a snow covered road |
| (4) | two people walking under snowfall | two people walking down a snow covered street |
| (5) | a SUV on side of road under heavy snowfall | a snow covered SUV driving down a snow covered road |
| (6) | snow covered road under heavy snowfall | a black and white photograph of a snowy city |

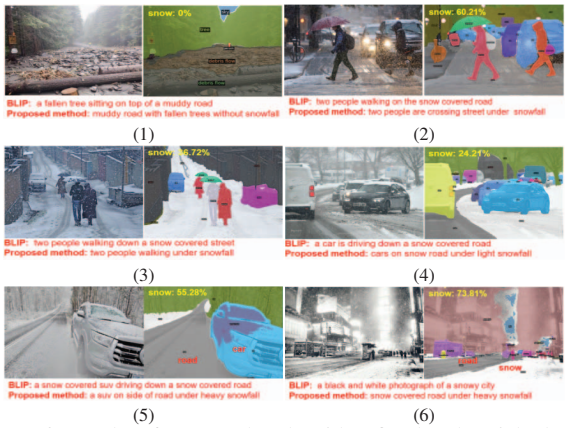



Figure 6. Results of proposed Dvls with refined and enriched captions in dynamic scenes: (1) Muddly road. (2) Heavy snowfall at night. (3), (4) Daytime light snowfall. (5) Car on the side road under heavy snowfall. (6) Heavy snowfall on the highway.

# 4. Ablation study

## 4.1. Caption Refinement by Dvls

To show the usefulness of refined Dvls, many unseen disaster scenes that have not been pre-trained are used to segment with classes. As shown in Figure 7 (a), images present disaster events. Two SOTAs of (c) MaskDINO [15]. (d) OVSeg [23] are compared.

As a result, Table 2 summarizes classes of (b) proposed Dvls and (c), (d) two SOTAs. In (1), a track (c) or boat (d) has been annotated, whereas the proposed Dvls have refined to "car crash" over water (b). In (2)-(5), snow to water, landslide to rocks, pavement to rain, and tree to strong wind have been annotated by (b) the proposed Dvls, respectively.

Therefore, refined texts from SOTAs' texts could enhance original to higher semantic texts. In particular, (5) tree (c) is normal segmentation, but strong wind (b), (d) stands for intuitive weather conditions as humans may announce. When combined with location prompts, Dvls can label segmented objects more semantically. Therefore, it has been proven that the proposed Dvls with texts will play an important role in messaging heavy disaster events more clearly than SOTAs' texts.

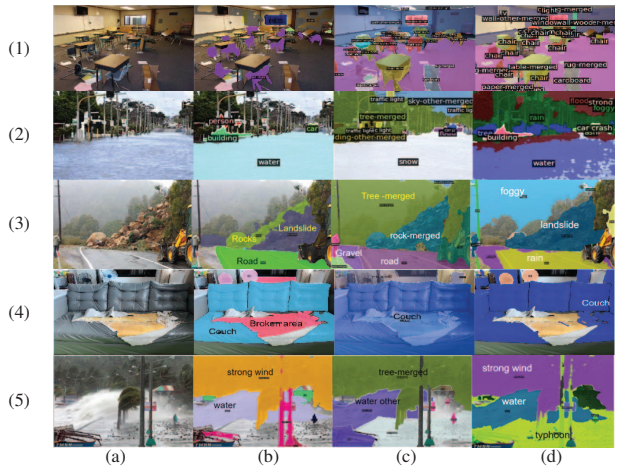



Figure 7. Comparison of the proposed method, MaskDINO [15], and OVSeg [23] (a) Input image. (b) Proposed Dvls. (c) MaskDINO [15] (d) OVSeg [23].

Table 2. Comparison of classes by SOTAs and proposed Dvls.

| Image | SOTA | Proposed |
|---|---|---|
| (1) | table, chair | fell chairs |
| (2) | snow, rain | water |
| (3) | rock-merged, rain | landslide |
| (4) | couch | couch, broken area |
| (5) | tree-merged, typhoon | strong wind |

## 4.2. The Comparison of Refined Caption and SOTA Image Captioning Model

This section describes a comparison of these integrated models to online image captioning API, i.e., visual ChatGPT [43], Midjourney, Img2prompt, and ClipIntegrator. Figure 8 presents a comparison between two approaches: (a) visual ChatGPT [43] caption results and (b) finally refined segmentation of the proposed with consideration of the relative size and location of objects. The comparison reveals that visual ChatGPT [43] only provides an overview of scene descriptions with no physical scales. On the other hand, the proposed model presents more detailed physical scales of sizes and locations for accident events. Thus, the proposed method has proven capable of handling dynamic captions. More refined and enriched captions by the proposed model have been generated for such traffic accident scenes than visual ChatGPT [43].

As depicted in Figure 9, the novel approach results in captions that encompass augmented information. This includes specific details , i.e., the road condition being either wet or dry, the presence of road damage due to debris flow or traffic jam, and the scene without snowfall. In contrast, when compared to alternative online API tools these enriched captions offer more specialized insights beyond just general descriptions.

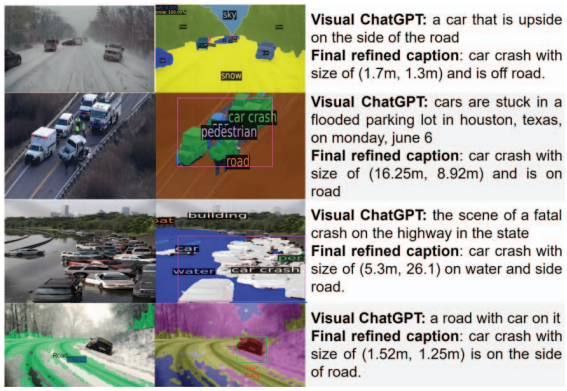



Figure 8. Comparison of image captioning between visual ChatGPT and the proposed method: (a) Input image. (b) Refined segmentation by the proposed EventCAP. (c) Captions from visual ChatGPT and the proposed EventCAP.

## 4.3. Overall Evaluation of the Proposed Dvls

This section describes an experiment that evaluates the performance of Dvls using the Intersection Over Union (IoU), a metric for evaluation of segmentation, and is conducted on ADE20k dataset [50], and COCO dataset [24]. The results is compared with SOTA VL models for segmentation, i.e., ZegFormer [3], and OpenSeg [7]. Table 3 show that Dvls outperforms ZegFormer [3], and OpenSeg [7].

Table 3. The performance of Dvls and compare with ZegFormer [3], and OpenSeg [7].

| *model dataset* | ADE20K [50] | COCO [24] |
|---|---|---|
| **Dvls** | 22.3 | 27.1 |
| **OpenSeg [7]** | 21.1 | 26.5 |
| **ZegFormer [3]** | 16.4 | 25.1 |

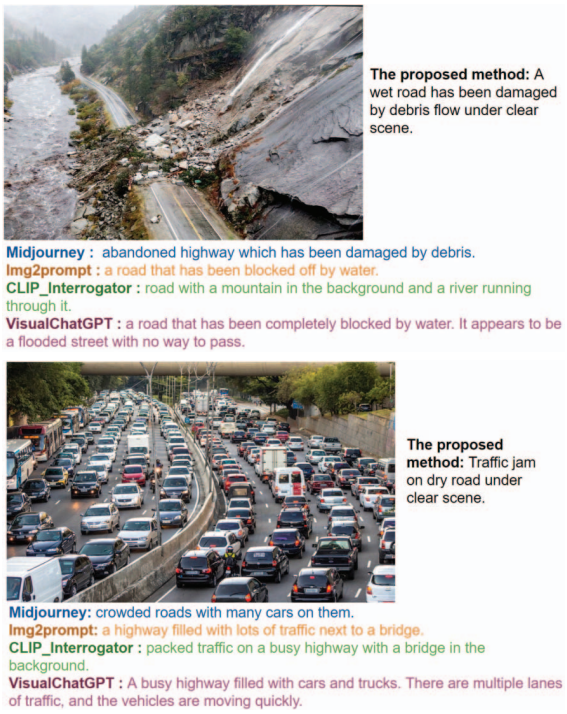



Figure 9. Comparison of vision-language models between online API tools and the proposed method.

## 5. Conclusion

This paper proposed EventCAP, a novel approach utilizing multiple complementary DL and VLM models featuring branched structures for enhanced efficiency in terms of memory, training, and maintenance aspects. This work marks the first instance of incorporating dynamic changes into captions under adverse weather conditions, encompassing factors like weather conditions and road conditions. EventCAP holds the potential to offer detailed scene depictions to drivers, autonomous driving systems, and rescue workers from camera images.

A 2D physics-based loss function will be applied to enhance captions further semantically. Moreover, additional modules such as water level estimation, enhanced road condition estimation, and traffic jam detection will be integrated for more detailed captions.

## References

[1] Feilong Chen, Duzhen Zhang, Minglun Han, Xiu-Yi Chen, Jing Shi, Shuang Xu, and Bo Xu. VLP: A survey on vision-language pre-training. *Int. J. Autom. Comput.*, 20(1):38–56, 2023. 1

[2] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 1280–1289. IEEE, 2022. 4

[3] Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. 2022. 5

[4] Jian Ding, Nan Xue, Gui-Song Xia, Bernt Schiele, and Dengxin Dai. Hgformer: Hierarchical grouping transformer for domain generalized semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15413–15423, June 2023. 1

[5] 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. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 1

[6] Jonathan Francis, Nariaki Kitamura, Felix Labelle, Xiaopeng Lu, Ingrid Navarro, and Jean Oh. Core challenges in embodied vision-language planning. *CoRR*, abs/2106.13948, 2021. 1

[7] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels, 2022. 5

[8] Jiaqi Gu, Hyoukjun Kwon, Dilin Wang, Wei Ye, Meng Li, Yu-Hsin Chen, Liangzhen Lai, Vikas Chandra, and David Z. Pan. Multi-scale high-resolution vision transformer for semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 12084–12093. IEEE, 2022. 1

[9] Chunle Guo, Qixin Yan, Saeed Anwar, Runmin Cong, Wenqi Ren, and Chongyi Li. Image dehazing transformer with transmission-aware 3d position embedding. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5802–5810, 2022. 1

[10] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 328–339. Association for Computational Linguistics, 2018. 1

[11] Lukas Hoyer, Dengxin Dai, Haoran Wang, and Luc Van Gool. MIC: masked image consistency for context-enhanced domain adaptation. *CoRR*, abs/2212.01322, 2022. 1

[12] Jingjing Jiang, Ziyi Liu, and Nanning Zheng. Correlation information bottleneck: Towards adapting pretrained multimodal models for robust visual question answering, 2023. 1

[13] Dahun Kim, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Video panoptic segmentation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 9856–9865. Computer Vision Foundation / IEEE, 2020. 1

[14] Sohyun Lee, Taeyoung Son, and Suha Kwak. FIFO: learning fog-invariant features for foggy scene segmentation.

In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 18889–18899. IEEE, 2022. 1

[15] Feng Li, Hao Zhang, Huaizhe xu, Shilong Liu, Lei Zhang, Lionel M. Ni, and Heung-Yeung Shum. Mask dino: Towards a unified transformer-based framework for object detection and segmentation, 2022. 5

[16] Feng Li, Hao Zhang, Yi-Fan Zhang, Shilong Liu, Jian Guo, Lionel M. Ni, PengChuan Zhang, and Lei Zhang. Vision-language intelligence: Tasks, representation learning, and large models. *CoRR*, abs/2203.01922, 2022. 1

[17] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR, 2022. 4

[18] Manling Li, Ruochen Xu, Shuohang Wang, Luowei Zhou, Xudong Lin, Chenguang Zhu, Michael Zeng, Heng Ji, and Shih-Fu Chang. Clip-event: Connecting text and images with event structures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16420–16429, June 2022. 1

[19] Yi Li, Yi Chang, Yan Gao, Changfeng Yu, and Luxin Yan. Physically disentangled intra- and inter-domain adaptation for varicolored haze removal. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 5831–5840. IEEE, 2022. 1

[20] Yanwei Li, Xinze Chen, Zheng Zhu, Lingxi Xie, Guan Huang, Dalong Du, and Xingang Wang. Attention-guided unified network for panoptic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 7026–7035. Computer Vision Foundation / IEEE, 2019. 1

[21] Yu Li, Shaodi You, Michael S. Brown, and Robby T. Tan. Haze visibility enhancement: A survey and quantitative benchmarking. *Computer Vision and Image Understanding*, 165:1–16, 2017. 1

[22] Yanwei Li, Hengshuang Zhao, Xiaojuan Qi, Liwei Wang, Zeming Li, Jian Sun, and Jiaya Jia. Fully convolutional networks for panoptic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 214–223. Computer Vision Foundation / IEEE, 2021. 1

[23] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted CLIP. *CoRR*, abs/2210.04150, 2022. 3, 5

[24] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. 5

[25] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 9992–10002. IEEE, 2021. 1, 3

[26] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 3431–3440. IEEE Computer Society, 2015. 1

[27] Xianzheng Ma, Zhixiang Wang, Yacheng Zhan, Yinqiang Zheng, Zheng Wang, Dengxin Dai, and Chia-Wen Lin. Both style and fog matter: Cumulative domain adaptation for semantic foggy scene understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 18900–18909. IEEE, 2022. 1

[28] Antoine Miech, Jean-Baptiste Alayrac, Ivan Laptev, Josef Sivic, and Andrew Zisserman. Rareact: A video dataset of unusual interactions. *CoRR*, abs/2008.01018, 2020. 1

[29] Ruijie Quan, Xin Yu, Yuanzhi Liang, and Yi Yang. Removing raindrops and rain streaks in one go. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 9147–9156. Computer Vision Foundation / IEEE, 2021. 1

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021. 1

[31] Hidetomo Sakaino. Deepreject and deeproad: Road condition recognition and classification under adversarial conditions. In *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 382–389, 2022. 2

[32] Hidetomo Sakaino. Deepscene, deepvis, deepdist, and deepreject: Image-based visibility estimation system for uav. In *2023 IEEE Aerospace Conference*, pages 1–11, 2023. 3

[33] H. Sakaino. Panopticroad: Integrated panoptic road segmentation under adversarial conditions. *in CVPR Workshop*, 2023. 1, 2, 3

[34] H. Sakaino. Panopticvis: Integrated panoptic segmentation for visibility estimation at twilight and night. *in CVPR Workshop*, 2023. 1, 2, 3

[35] H. Sakaino. Physicscap: Dynamic captions for natural scene changes. In *ACM International Conf. Machine Learning (ICML), Workshop on Data-centric Machine Learning Research (DMLR)*, 2023. Nonarchival. 1

[36] H. Sakaino. Refined and enriched physics-based captions for unseen dynamic changes. In *ACM International Conf. Machine Learning (ICML), Workshop on the 2nd New Frontiers In Adversarial Machine Learning (ADVML FRONTIERS)*, 2023. Nonarchival. 1

[37] Aditya Sanghi, Hang Chu, Joseph G. Lambourne, Ye Wang, Chin-Yi Cheng, Marco Fumero, and Kamal Rahimi Malekshan. Clip-forge: Towards zero-shot text-to-shape generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18603–18613, June 2022. 1

[38] Hengcan Shi, Munawar Hayat, Yicheng Wu, and Jianfei Cai. Proposalclip: Unsupervised open-category object proposal generation via exploiting clip cues. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9611–9620, June 2022. 1

[39] S. Sreelakshmi and S.S. V. Chandra. Machine learning for disaster management: insights from past research and future implications. *In Proc. IEEE Int. Conf. Comput. Communication, Security, and Intelligent Systems (IC3SIS)*, 2022. 1

[40] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR, 2019. 1

[41] Wei Wang, Zhun Zhong, Weijie Wang, Xi Chen, Charles Ling, Boyu Wang, and Nicu Sebe. Dynamically instance-guided adaptation: A backward-free approach for test-time domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24090–24099, June 2023. 1

[42] Wenguan Wang, Tianfei Zhou, Fisher Yu, Jifeng Dai, Ender Konukoglu, and Luc Van Gool. Exploring cross-image pixel contrast for semantic segmentation. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 7283–7293. IEEE, 2021. 1

[43] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *CoRR*, abs/2303.04671, 2023. 2, 5

[44] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 12077–12090, 2021. 1

[45] Wending Yan, Aashish Sharma, and Robby T. Tan. Optical flow in dense foggy scenes using semi-supervised learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13256–13265, 2020. 1

[46] Wenhan Yang, Robby T. Tan, Shiqi Wang, Yuming Fang, and Jiaying Liu. Single image deraining: From model-based to data-driven and beyond. *CoRR*, abs/1912.07150, 2019. 1

[47] Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19313–19322, June 2022. 2

[48] Haotian Zhang, Pengchuan Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Harold Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. Glipv2: Unifying localization and vision-language understanding. *CoRR*, abs/2206.05836, 2022. 2

[49] Zhen Zhao, Sifan Long, Jimin Pi, Jingdong Wang, and Luping Zhou. Instance-specific and model-adaptive supervision for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23705–23714, June 2023. 1

[50] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset, 2018. 5

[51] Xiangyang Zhu, Renrui Zhang, Bowei He, Ziyao Zeng, Shanghang Zhang, and Peng Gao. Pointclip V2: adapting CLIP for powerful 3d open-world learning. *CoRR*, abs/2211.11682, 2022. 1