The STVchrono Dataset: Towards Continuous Change Recognition in Time (Supplementary Material)

The supplementary material includes further analysis of the STVchrono dataset, showcasing the distributions of word (I.1) and sentence (caption) lengths (I.2), time deltas (I.3), as well as all the cities encompassed in the STVchrono dataset (I.4). It also details the implementation of multi-modal LLM-based methods mentioned in the main paper, including OpenFlamingo and the combination of BLIP2 with GPT-4 (II.1). We also provide the analysis of regional performance (II.2) and image-based segmentation experiments (II.3), along with additional experimental results (II.4). A datasheet for the STRchrono dataset is provided in the last section (III).

I. Dataset Statistics

I.1. Word Distribution

We analyze the word distribution in the captions of two setups for the continual change captioning tasks: image pair (Figure 7, left) and image sequence (Figure 7, right) using WordCloud visualization [1]. Both setups feature a wide variety of words. The image pair task requires identifying differences between two images, leading to captions that include a greater number of comparative words, such as "brighter", "greener", "cleaner", and "different". Conversely, the image sequence task involves recognizing trends, superlatives, and similarities across image sequences. As a result, the dataset contains a higher frequency of relative terms like "newest", "thickest", "clearest", and "gradually".

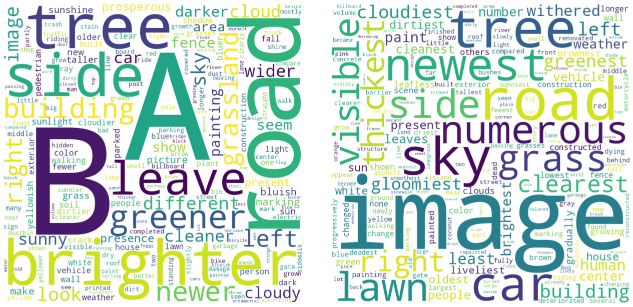


Figure 7. Wordcloud visualization of the continual change captioning (image pair) task (left) and the continual change captioning (image sequence) task (right) of the STVchrono dataset.

I.2. Sentence Length Distribution

The sentence length (change caption length per dataset instance) distribution for the two continual change captioning tasks is presented in Figure 8. Both dataset setups exhibit a long-tailed distribution. Specifically, the image pair task has an average sentence length of 35.98, while the image sequence task, involving more images, has an average sentence length of 50.65. Due to the minimum sentence number requirement set for sequences (three for 3-image and 4-image sequences, and five for 5-image and 6-image sequences), there are two distinct peaks in the sentence length distribution for the continual change captioning (image sequence). Additionally, both datasets feature a significant number of instances with longer sentences, offering a wide array of detailed changes in the image pairs and sequences for model training and evaluation.

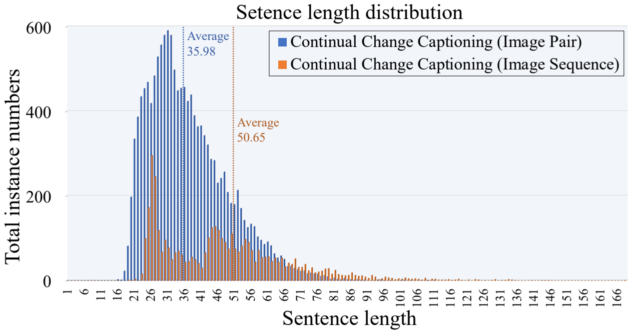


Figure 8. Sentence length distribution of two continual change captioning tasks of the STVchrono dataset.

I.3. Time Deltas Distribution

Figure 9 describes the distribution of time deltas (spanned years of each dataset instance) for the three tasks of the STVchrono dataset. All three tasks encompass instances with a wide range of time deltas.

[1] https://amueller.github.io/word_cloud

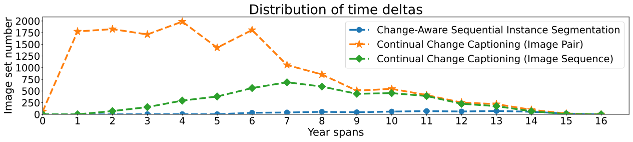


Figure 9. Distribution of the time deltas of the STVchrono dataset.

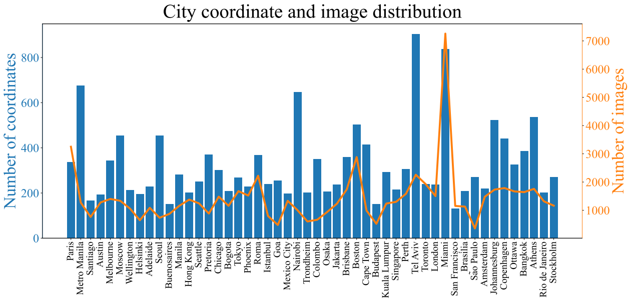


Figure 10. City distribution of the STVchrono dataset.

I.4. City Distribution

We have amassed a collection of photographs from 50 cities across the globe. The distribution is as follows: Europe contributed a total of 17,061 images, Asia represented 15,967 images, Oceania with 6,893 images, North America with 21,749 images, South America with 5,626 images, and Africa with 4,604 images. Figure 10 shows the distribution of the number of coordinates per city alongside the number of images.

We split the STVchrono dataset by cities into train and test sets for the three tasks. In detail, for the two continual change captioning (image pair and image sequence) tasks, cities including Johannesburg, Nairobi, Pretoria, Colombo, Hong Kong, Kuala Lumpur, Manila, Metro Manila, Osaka, Seoul, Singapore, Tokyo, Adelaide, Melbourne, Perth, Wellington, Athens, Copenhagen, Helsinki, Moscow, Paris, Roma, Stockholm, Trondheim, Austin, Boston, Miami, Mexico City, Phoenix, San Francisco, Seattle, Toronto, Goa, Brasília, Buenos Aires, Rio de Janeiro, Santiago, São Paulo were used for training, and cities including Cape Town, Bangkok, Istanbul, Jakarta, Brisbane, Amsterdam, Budapest, London, Chicago, Ottawa, Tel Aviv Yafo, Bogotá were for testing. For the change-aware sequential instance segmentation task, cities used in training consist of Johannesburg, Jakarta, Kuala Lumpur, Manila, Osaka, Singapore, Tokyo, Melbourne, Perth, Wellington, Moscow, Paris, Roma, Stockholm, Austin, Mexico City, Miami, Seattle, Toronto, Brasília, Buenos Aires, Santiago, and cities used in testing consist of Bangkok, Hong Kong, Istanbul, Brisbane, Amsterdam, London, Ottawa, Bogotá.

Prompt for OpenFlamingo

Example

*<image>*This is image A and taken at {prompt_imgA_year} .
<image> This is image B and taken at {prompt_imgB_year}.
The changes of A and B are: {prompt_cap} *<|endofchunk|>*.

Query

<image> This is image A and taken at {query_imgA_year} .
<image> This is image B and taken at {query_imgB_year}.
The changes of A and B are:

(a) Prompt design for OpenFlamingo (image pair).

Prompt for OpenFlamingo

Example

<image> This is image 1 and taken at {prompt_img1_year} .
<image> This is image 2 and taken at {prompt_img2_year}.
<image> This is image 3 and taken at {prompt_img3_year}.
...
The changes tendency of 1, 2, 3 … are: {prompt_cap} *<|endofchunk|>*.

Query

<image> This is image 1 and taken at {query_img1_year} .
<image> This is image 2 and taken at {query _img2_year}.
<image> This is image 3 and taken at {query _img3_year}.
...
The changes tendency of 1, 2, 3 … are:

(b) Prompt design for OpenFlamingo (image sequence).

Figure 11. Prompt designs for OpenFlamingo.

II. Experiment

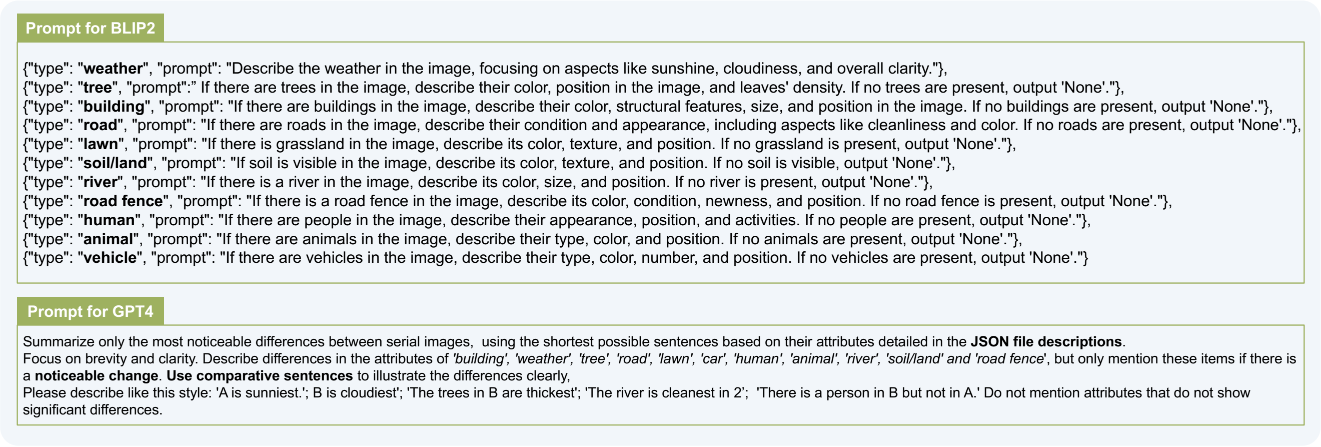
II.1. Prompt Design

In this section, we present the detailed prompt designs used with OpenFlamingo and BLIP2 + GPT4 in continual change captioning tasks.

OpenFlamingo: We emulated the image caption generation prompts released by the official OpenFlamingo, designing our own prompts. Figure 11 illustrates our prompt designs, where we tested different numbers of examples provided to OpenFlamingo and observed its change description performance. The top side of Figure 11 shows the prompt for image pairs, and the bottom side shows the prompt for image sequences.

During the implementation of OpenFlamingo, we focused on two primary strategies. Firstly, we standardized the number of examples input into OpenFlamingo, selecting sets of 3, 5, 10, 15, and 20 examples for different experimental scenarios. Secondly, we experimented with the modifications of the model's output format. We designed two output formats as follows:

- complete sentences (*e.g.*, *"B building is clearer than A. B grassland is greener than A. Road B is newer than Road A. B is darker than A."*).
- itemized structures (*e.g.*, *'building': ['item': 'Old and new', 'answer': 'B building is clearer than A.'], 'human': [], 'grassland': ['item': 'Color', 'answer': 'B grassland*



Prompt for BLIP2

```
{"type": "weather", "prompt": "Describe the weather in the image, focusing on aspects like sunshine, cloudiness, and overall clarity."},
{"type": "tree", "prompt":” If there are trees in the image, describe their color, position in the image, and leaves' density. If no trees are present, output 'None'."},
{"type": "building", "prompt": "If there are buildings in the image, describe their color, structural features, size, and position in the image. If no buildings are present, output 'None'."},
{"type": "road", "prompt": "If there are roads in the image, describe their condition and appearance, including aspects like cleanliness and color. If no roads are present, output 'None'."},
{"type": "lawn", "prompt": "If there is grassland in the image, describe its color, texture, and position. If no grassland is present, output 'None'."},
{"type": "soil/land", "prompt": "If soil is visible in the image, describe its color, texture, and position. If no soil is visible, output 'None'."},
{"type": "river", "prompt": "If there is a river in the image, describe its color, size, and position. If no river is present, output 'None'."},
{"type": "road fence", "prompt": "If there is a road fence in the image, describe its color, condition, newness, and position. If no road fence is present, output 'None'."},
{"type": "human", "prompt": "If there are people in the image, describe their appearance, position, and activities. If no people are present, output 'None'."},
{"type": "animal", "prompt": "If there are animals in the image, describe their type, color, and position. If no animals are present, output 'None'."},
{"type": "vehicle", "prompt": "If there are vehicles in the image, describe their type, color, number, and position. If no vehicles are present, output 'None'."}
```

Prompt for GPT4

Summarize only the most noticeable differences between serial images, using the shortest possible sentences based on their attributes detailed in the **JSON file descriptions**.
Focus on brevity and clarity. Describe differences in the attributes of *'building', 'weather', 'tree', 'road', 'lawn', 'car', 'human', 'animal', 'river', 'soil/land' and 'road fence'*, but only mention these items if there is a **noticeable change**. **Use comparative sentences** to illustrate the differences clearly,
Please describe like this style: 'A is sunniest.'; B is cloudiest'; 'The trees in B are thickest'; 'The river is cleanest in 2'; 'There is a person in B but not in A.' Do not mention attributes that do not show significant differences.

Figure 12. Prompt design for BLIP2 + GPT4.

Number of examples	**image pair** BLEU4/CIDEr	**3-image sequence** BLEU4/CIDEr	**4-image sequence** BLEU4/CIDEr	**5-image sequence** BLEU4/CIDEr	**6-image sequence** BLEU4/CIDEr
3	-	-	-	11.0/6.7	**8.8/8.8**
5	7.7/31.3	11.5/30.6	**11.8/32.4**	**9.8/12.2**	4.8/7.0
10	6.5/19.3	**14.4/40.0**	9.7/17.9	7.5/5.0	-
15	9.4/27.2	12.9/29.4	11.7/24.4	-	-
20	**7.8/37.3**	11.5/30.6	-	-	-

Table 5. Change description evaluation on continual change captioning tasks using OpenFlamingo.

is greener than A.'], 'road': ['item': 'Old and new', 'answer': 'Road B is newer than Road A.'], 'road fence': [], 'tree': [], 'weather': ['item': 'Light and darkness', 'answer': 'B is darker than A.']).

During the experiments, we found that among the two output formats, complete sentences slightly outperformed the itemized output. Therefore, in our main experiments, we used complete sentences as the output. Moreover, initially, there was a concern that OpenFlamingo might not handle multiple images effectively. This led to trials, in which images were concatenated horizontally, before being input into OpenFlamingo. However, we found that inputting images separately yielded more effective results, and we adopted this approach for all experiments.

The results of experiments with varying example numbers are shown in Table 5, where the best results are reported in the main paper. In shorter sequences (like image pairs or 3/4-image sequences) the model often replicated prompt language, leading to minimal BLEU4 and CIDEr scores. Thus, we excluded 3-example experiments for these sequences. Experiments with larger example numbers for longer sequences were also omitted due to the input token length limitations.

BLIP2 + GPT4: The prompt design for BLIP2 and GPT4 includes in-context examples for BLIP2 (as shown in the top part of Figure 12) and system messages for GPT4 (as shown in the bottom part of Figure 12). Since GPT4 cannot directly process images, we first utilized BLIP2 to identify different subject attributes in the images, focusing on characteristics such as color and age, as highlighted in the main paper. The recognized results for each subject were then saved as JSON files. In the final step, we input these JSON files into GPT4, enabling it to effectively summarize the differences between images by parsing the JSON data.

Base model	**image pair** BLEU4/CIDEr	**3-image sequence** BLEU4/CIDEr	**4-image sequence** BLEU4/CIDEr	**5-image sequence** BLEU4/CIDEr	**6-image sequence** BLEU4/CIDEr
Blip2-opt-2.7b	**4.2/16.1**	**5.1/12.4**	5.0/4.3	**5.2/6.3**	**4.1/6.8**
Blip2-flan-t5-xl	3.9/8.6	4.9/7.6	**5.2/4.6**	4.1/3.8	3.1/2.8

Table 6. Change description evaluation on continual change captioning tasks using BLIP2+GPT4.

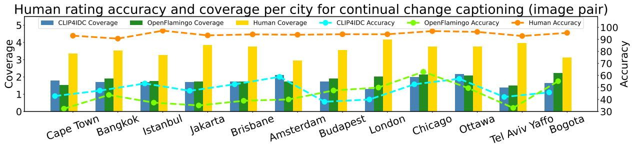


Figure 13. Region-based performance for the continual change captioning (image pair) task.

Method	Backbone	AP	AP50	AP75
Mask2Former (image-based) + SORT	ResNet50	3.78	5.06	3.81
	ResNet101	3.94	5.20	3.96

Table 7. Evaluation of the image-based methods on the change-aware sequential instance segmentation task.

We experimented with two different BLIP2 base models (BLIP2-opt-2.7b, BLIP2-flan-t5-xl), setting the number of BLIP2 tokens at 10, 15, 20, 25, and 30. Additionally, we adjusted the GPT4 prompts based on the output from BLIP2 to generate grammatically similar sentences to ground truths. Through experiments with continual change caption (image pair) and continual change caption (image sequence) 3-image sequence data, we finalized the design of the prompts for BLIP2 and GPT4. We found that a BLIP2 token count of 25 was optimal. A comparative analysis of the opt- and flan-based models, detailed in Table 6, revealed that the opt-based model generally outperformed the flan model. Consequently, the main paper presents the average results derived from the opt-based model.

II.2. Analysis of Region Performance

Figure 13 presents city-wise human-rated scores for the continual change captioning (image pair) task. Humans showed consistent accuracy across cities, while two automated methods struggled in certain cities, indicating varying levels of difficulty for these methods. For coverage, humans identified more changes in certain cities, likely noticing significant changes more easily in some cities.

II.3. Image-based Segmentation Experiments

In the main paper, we mainly evaluated video-based methods for change-aware sequential instance segmentation. Here, we treat the images in the time series independently and implement an adaptive matching method for connecting results within each image sequence. Specifically, we used Mask2Former [1] for independent image segmentation and SORT [2] for matching. The results (Table 7) show slightly lower performance compared to the video-based methods, highlighting the challenges and the need for improvement in the image-and-matching-based methods.

II.4. Additional Experimental Results

Continual change captioning (image pair): Two result examples utilizing existing methods are presented in Figure 14. These examples demonstrate that most existing methods accurately identified one to two changes in the image pairs. In contrast to human-annotated ground truth, which includes comprehensive change details (like “a bicycle on the left side of A” in Figure 14 (a), and “a red mailbox on the left side of B” in Figure 14 (b)), the majority of methods offered less precise descriptions (such as “road B is newer than A” or “A has more leaves than B”). Remarkably, the BLIP2+GPT4 combination yielded more nuanced change descriptions: example in Figure 14 (b) specifically highlighting the busyness level of the road.

Continual change captioning (image sequence): Four result examples are presented in Figure 15. In this task, most existing methods were only able to retrieve zero or one accurate description. The challenge of continual change captioning from image sequences lies in the need for comparisons and correlations, across these sequences, to identify tendencies, superlatives, and similarities. This remains a difficult task for the current methodologies. We hope that the STVchrono dataset will serve as a valuable resource for the future improvements in understanding and correlating image sequences.

Change-aware sequential instance segmentation: Additional experimental results for Mask2Former and CTVIS are illustrated in Figure 16. In the visualizations of the ground truth images and the results from these two methods, the same instance (*e.g.*, a specific building or tree) in sequential images is consistently marked with the same color mask, regardless of appearance changes such as a tree growing or a building being constructed. The results show that both methods accurately segment larger object instances (like “sky” and “road”) with high consistency, compared to the ground truth. However, both tend to overlook smaller objects (*e.g.*, small cars and trees in (a) and (b)). They also struggle with maintaining consistent instance labels for objects that have undergone significant appearance changes (*e.g.*, buildings in (b)). The STVchrono dataset, encompassing a diverse range of object types and their appearance changes, as well as shifts in camera viewpoint, presents a novel challenge in correlating scenes and objects in image sequences to achieve a human-level understanding.

III. Datasheet

We follow the framework defined by [3] and provide the datasheet for the STVchrono dataset in this section.

III.1. Motivation

For what purpose was the dataset created? Who created this dataset?

The STVchrono dataset was created by researchers to confidently recognize changes in the long-term serials from street view images, both in terms of what has changed and where the changes have occurred.

Any other comments?
None.

III.2. Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?
All instances of the dataset are the text, images, and segmentation masks.

How many instances are there in total (of each type, if appropriate)?
There are 71,900 images, 5.3MB texts, and 120.9MB segmentation masks in total.

Does the dataset contain all possible instances, or is it a sample (not necessarily random) of instances from a larger set?
The dataset contains all possible instances.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features?
The data in STVchrono has different types for three tasks. For continual change captioning (image pair), each instance is made up of an image pair (2 images) and a corresponding text; for continual change captioning (image sequence) each instance is made up of an image sequence (3, 4, 5, 6 images) and a corresponding text; for the change-aware sequential instance segmentation, we have an image sequence (5 images) and the segmentation masks. (see Section 3 in the main paper)

Is there a label or target associated with each instance?
Yes, there is a label associated with each instance.

Is any information missing from individual instances?
No.

Are relationships between individual instances made explicit?
No.

Are there recommended data splits (e.g., training, development/validation, testing)?
We provide the data split on our GitHub Page.

Are there any errors, sources of noise, or redundancies in the dataset?
Due to the human labeling process, there is some noise in several images, which does not obstruct the tasks.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?
Yes. The dataset needs to access images from Google Street View.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?
No.

Does the dataset relate to people?
Yes. Some images in the STVchrono dataset present people. The annotations of image pairs and image sequences are human-made.

Does the dataset identify any subpopulations (e.g., by age, gender)?
No.

Does the dataset contain data that might be considered sensitive in any way?
The dataset likely contains some location information in the street view images.

Any other comments?
No.

III.3. Collection Process

How was the data associated with each instance acquired?
The data is available on our GitHub Page.

What mechanisms or procedures were used to collect the data?
We used Google Street View API to access the images and ask humans to annotate the data.

Who was involved in the data collection process and how were they compensated?
Students and researchers.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances?
The dataset was collected over a period of several months in 2023.

Were any ethical review processes conducted?
No.

III.4. Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done?
The details of preprocessing are discussed in the main paper (Section 3).

III.5. Uses

Has the dataset been used for any tasks already?
Yes. We used the dataset for three tasks we designed: continual change captioning (image pair), continual change captioning (image sequence), and change-aware sequential instance segmentation. For the details, please see Section 3 in the main paper.

Is there a repository that links to any or all papers or systems that use the dataset?
STVchrono Dataset

Any other comments?
No.

III.6. Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?
Yes.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?
GitHub. We shared the link above.

When will the dataset be distributed?
We will release the dataset when submitting the camera-ready paper.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?
Yes. The dataset is for research-only purpose. The dataset was created based on Google Street View. Google Street View images are protected under copyright laws, and utilizing them to create a dataset involves adhering to Google's terms of service and ensuring compliance with the copyright laws.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances?
Google.

III.7. Maintenance

Who will be supporting/hosting/maintaining the dataset?
The authors of this paper.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?
You can find the email address on the GitHub Page.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?
Yes, we will update the dataset later by adding new instances.

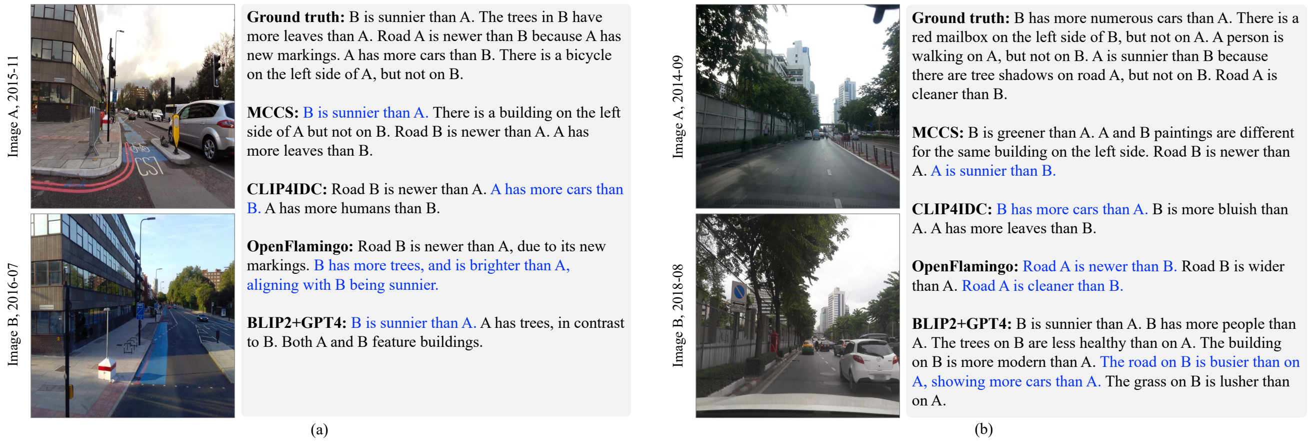
Any other comments?
No.

References

[1] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2022. 4

[2] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE international conference on image processing (ICIP)*, pages 3464–3468, 2016. 4

[3] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *CoRR*, abs/1803.09010, 2018. 4



Ground truth: B is sunnier than A. The trees in B have more leaves than A. Road A is newer than B because A has new markings. A has more cars than B. There is a bicycle on the left side of A, but not on B.

MCCS: B is sunnier than A. There is a building on the left side of A but not on B. Road B is newer than A. A has more leaves than B.

CLIP4IDC: Road B is newer than A. A has more cars than B. A has more humans than B.

OpenFlamingo: Road B is newer than A, due to its new markings. B has more trees, and is brighter than A, aligning with B being sunnier.

BLIP2+GPT4: B is sunnier than A. A has trees, in contrast to B. Both A and B feature buildings.

(a)

Ground truth: B has more numerous cars than A. There is a red mailbox on the left side of B, but not on A. A person is walking on A, but not on B. A is sunnier than B because there are tree shadows on road A, but not on B. Road A is cleaner than B.

MCCS: B is greener than A. A and B paintings are different for the same building on the left side. Road B is newer than A. A is sunnier than B.

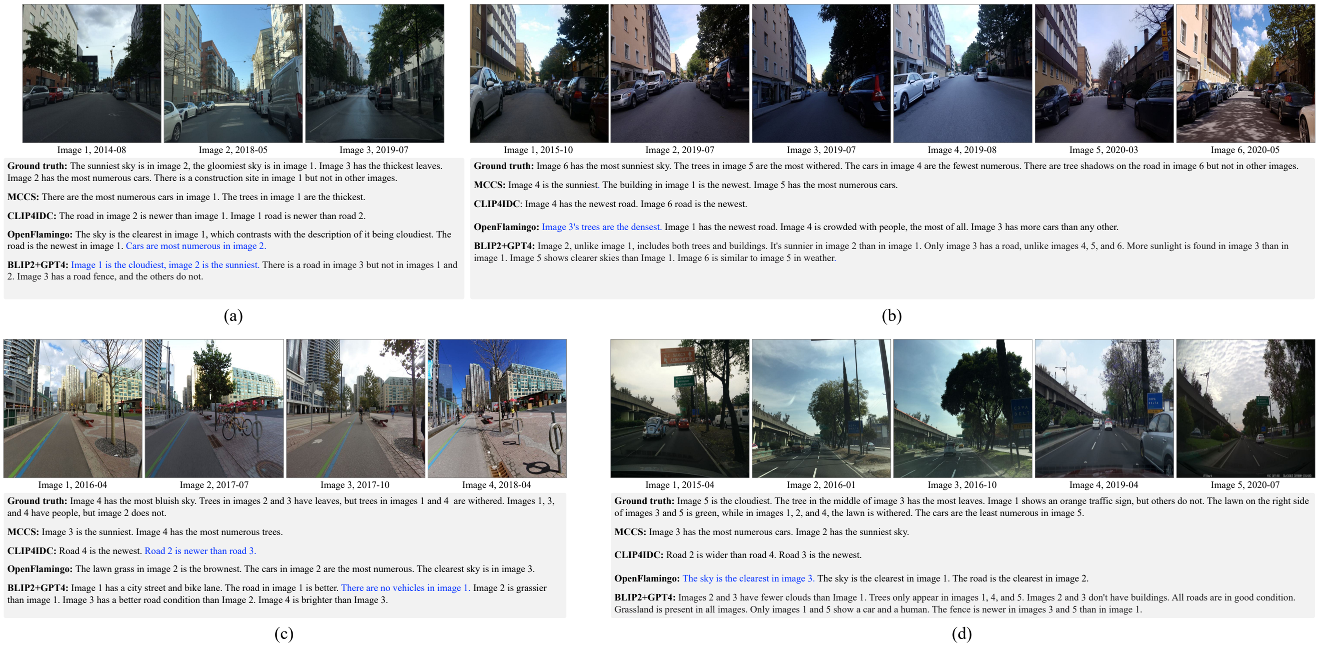
CLIP4IDC: B has more cars than A. B is more bluish than A. A has more leaves than B.

OpenFlamingo: Road A is newer than B. Road B is wider than A. Road A is cleaner than B.

BLIP2+GPT4: B is sunnier than A. B has more people than A. The trees on B are less healthy than on A. The building on B is more modern than A. The road on B is busier than on A, showing more cars than A. The grass on B is lusher than on A.

(b)

Figure 14. Two examples of the continual change captioning (image pair) setup, along with the results using existing methods. The correctly retrieved changes are highlighted in blue.



Ground truth: The sunniest sky is in image 2, the gloomiest sky is in image 1. Image 3 has the thickest leaves. Image 2 has the most numerous cars. There is a construction site in image 1 but not in other images.

MCCS: There are the most numerous cars in image 1. The trees in image 1 are the thickest.

CLIP4IDC: The road in image 2 is newer than image 1. Image 1 road is newer than road 2.

OpenFlamingo: The sky is the clearest in image 1, which contrasts with the description of it being cloudiest. The road is the newest in image 1. Cars are most numerous in image 2.

BLIP2+GPT4: Image 1 is the cloudiest, image 2 is the sunniest. There is a road in image 3 but not in images 1 and 2. Image 3 has a road fence, and the others do not.

(a)

Ground truth: Image 6 has the most sunniest sky. The trees in image 5 are the most withered. The cars in image 4 are the fewest numerous. There are tree shadows on the road in image 6 but not in other images.

MCCS: Image 4 is the sunniest. The building in image 1 is the newest. Image 5 has the most numerous cars.

CLIP4IDC: Image 4 has the newest road. Image 6 road is the newest.

OpenFlamingo: Image 3's trees are the densest. Image 1 has the newest road. Image 4 is crowded with people, the most of all. Image 3 has more cars than any other.

BLIP2+GPT4: Image 2, unlike image 1, includes both trees and buildings. It's sunnier in image 2 than in image 1. Only image 3 has a road, unlike images 4, 5, and 6. More sunlight is found in image 3 than in image 1. Image 5 shows clearer skies than Image 1. Image 6 is similar to image 5 in weather.

(b)

Ground truth: Image 4 has the most bluish sky. Trees in images 2 and 3 have leaves, but trees in images 1 and 4 are withered. Images 1, 3, and 4 have people, but image 2 does not.

MCCS: Image 3 is the sunniest. Image 4 has the most numerous trees.

CLIP4IDC: Road 4 is the newest. Road 2 is newer than road 3.

OpenFlamingo: The lawn grass in image 2 is the brownest. The cars in image 2 are the most numerous. The clearest sky is in image 3.

BLIP2+GPT4: Image 1 has a city street and bike lane. The road in image 1 is better. There are no vehicles in image 1. Image 2 is grassier than image 1. Image 3 has a better road condition than Image 2. Image 4 is brighter than Image 3.

(c)

Ground truth: Image 5 is the cloudiest. The tree in the middle of image 3 has the most leaves. Image 1 shows an orange traffic sign, but others do not. The lawn on the right side of images 3 and 5 is green, while in images 1, 2, and 4, the lawn is withered. The cars are the least numerous in image 5.

MCCS: Image 3 has the most numerous cars. Image 2 has the sunniest sky.

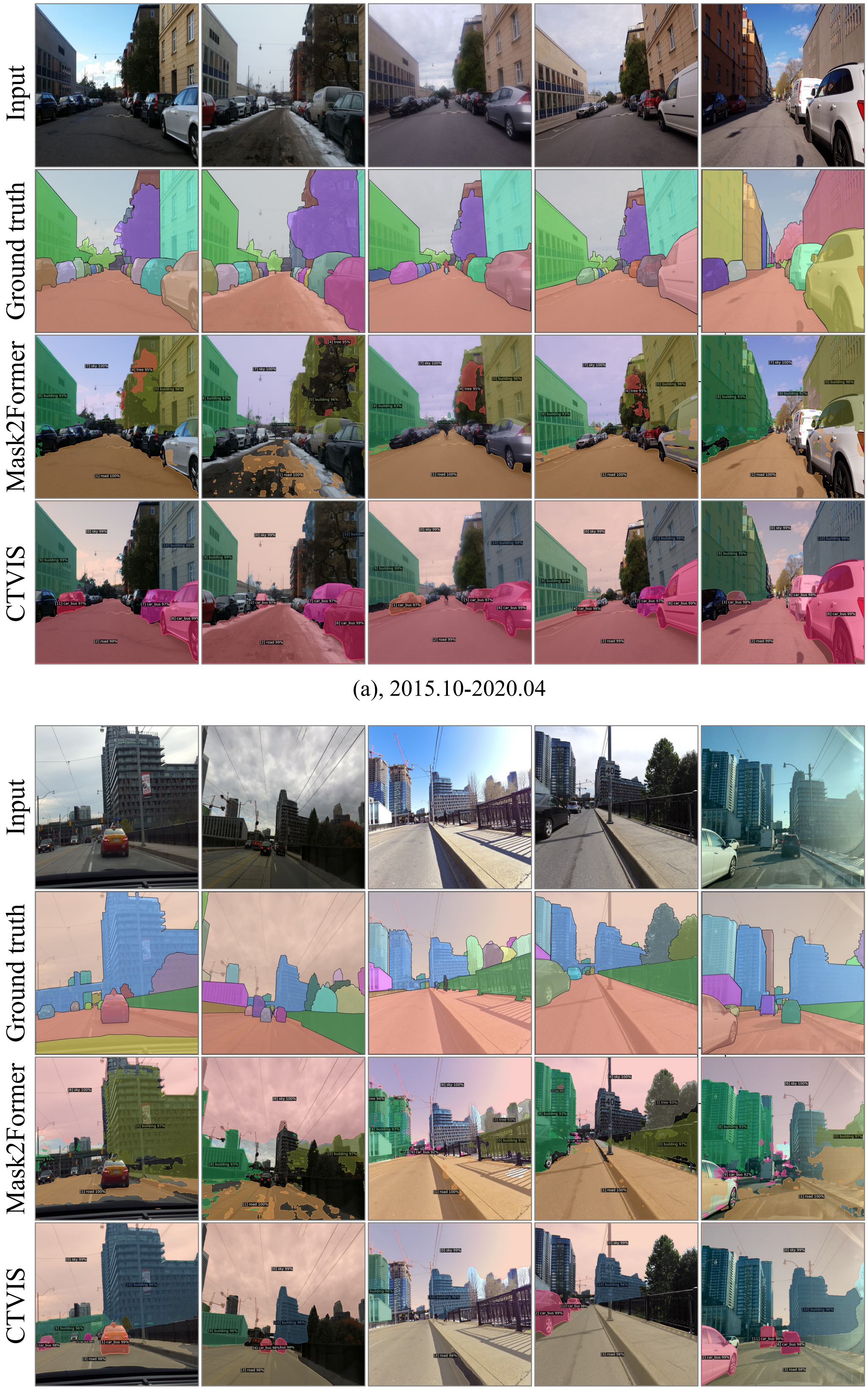
CLIP4IDC: Road 2 is wider than road 4. Road 3 is the newest.

OpenFlamingo: The sky is the clearest in image 3. The sky is the clearest in image 1. The road is the clearest in image 2.

BLIP2+GPT4: Images 2 and 3 have fewer clouds than Image 1. Trees only appear in images 1, 4, and 5. Images 2 and 3 don't have buildings. All roads are in good condition. Grassland is present in all images. Only images 1 and 5 show a car and a human. The fence is newer in images 3 and 5 than in image 1.

(d)

Figure 15. Four examples of the continual change captioning (image sequence) setup, along with the results using existing methods. The correctly retrieved changes are highlighted in blue.



(a), 2015.10-2020.04

(b), 2015.11-2019.12

Figure 16. Two dataset examples and results on the existing methods: Mask2Former and CTVIS. Objects with the consistent IDs share the same mask colors within each sequence.