This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

# The STVchrono Dataset: Towards Continuous Change Recognition in Time

Yanjun Sun[1,2*], Yue Qiu[1*], Mariia Khan[3], Fumiya Matsuzawa[1], Kenji Iwata[1]
[1]National Institute of Advanced Industrial Science and Technology (AIST), [2]Keio University,
[3]Edith Cowan University
{yanjun.son,qiu.yue,fumi8.matsuzawa,kenji.iwata}@aist.go.jp, mariiak@our.ecu.edu.au

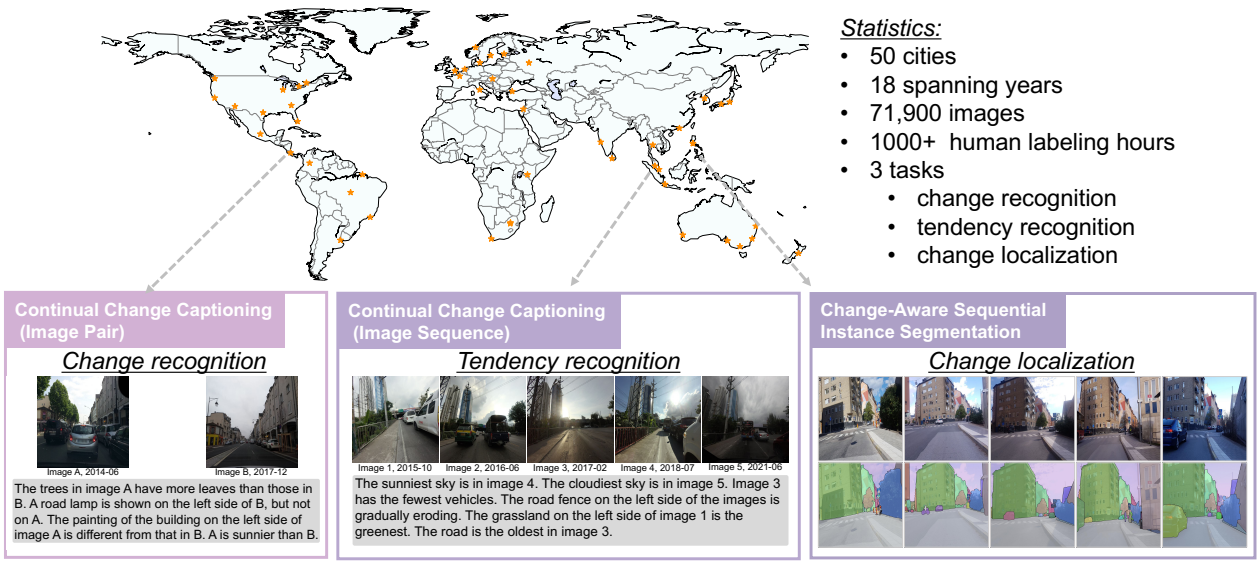


Figure 1. Overview of the proposed STVchrono dataset.

## Abstract

*Recognizing continuous changes offers valuable insights into past historical events, supports current trend analysis, and facilitates future planning. This knowledge is crucial for a variety of fields, such as meteorology and agriculture, environmental science, urban planning and construction, tourism, and cultural preservation. Currently available datasets in the field of scene change understanding primarily concentrate on two main tasks: the detection of changed regions within a scene and the linguistic description of the change content. Existing datasets focus on recognizing discrete changes, such as adding or deleting an object from two images, and largely rely on artificially generated images. Consequently, the existing change understanding methods primarily focus on identifying distinct object differences, overlooking the importance of continuous, gradual changes occurring over extended time intervals.*

*To address the above issues, we propose a novel benchmark dataset, STVchrono, targeting the localization and description of long-term continuous changes in real-world scenes. The dataset consists of 71,900 photographs from Google Street View API taken over an 18-year span across 50 cities all over the world. Our STVchrono dataset is designed to support real-world continuous change recognition and description in both image pairs and extended image sequences, while also enabling the segmentation of changed regions. We conduct experiments to evaluate state-of-the-art methods on continuous change description and segmentation, as well as multimodal Large Language Models for describing changes. Our findings reveal that even the most advanced methods lag human performance, emphasizing the need to adapt them to continuously changing real-world scenarios. We hope that our benchmark dataset will further facilitate the research of temporal change recognition in a dynamic world. The STVchrono dataset is available at STVchrono Dataset.*

*Equal contribution.

## 1. Introduction

The world around us is constantly changing over time due to environmental changes, human activities, and technological progress. Change recognition in real-life outdoor scenes is crucial for applications such as meteorology and agriculture, environmental science, urban planning and construction, tourism, and cultural preservation.

Real-world changes may include different spatial and temporal changes in the natural landscape (*e.g*. water volume in the river), urban infrastructure (*e.g*. road width), weather conditions (*e.g*. season change), or population dynamics (*e.g*. type of human activities). What matters the most is the continuous and dynamic nature of all these change types. Recognition of continuous changes can provide valuable insights into past historical events, support current trend analysis, and facilitate future planning.

Currently available tasks related to scene change understanding focus on change detection and change description. While the target of change detection is to find changed regions within a scene, change description deals with the generation of language captions for the detected changes. The existing change datasets [1–11] mostly focus on recognizing discrete changes between paired images or 3D point clouds, overlooking the importance of continuous, gradual changes, occurring over long time periods. Additionally, these datasets either include synthetic data [7–10] (artificially generated images from simulated environments) or concentrate on simplified real-world scenes (like tabletop rearrangement in [1]). Thus, these datasets are not suitable for understanding the real-world continuous changes.

To address the above-mentioned limitations, we propose a novel benchmark STVchrono (STreet View chrono) dataset. STVchrono is designed to facilitate the understanding of long-term continuous changes in the real world. To capture continuous outdoor changes, we utilize the Google Street View API [*] for data collection. Specifically, we collected 71,900 photographs of 50 different cities over a span of 18 years (2006 to 2023). The chosen 50 cities vary in location (spread across various continents) and encompass different landscape types (urban and rural areas).

The STVchrono dataset is suitable to facilitate three change understanding tasks (Figure 1): continual change captioning for image pairs and image sequences, and change-aware sequential instance segmentation (for change recognition). The aim of continual change captioning for an image pair is to describe the content of the change between a pair of images, taken in the same location but at two different times. These changes may include variations in color, age, volume, or condition for 10 object types (Table 2). Another type of continual change captioning task deals with the longer image sequences (3-6 images) taken over a span of several years. This task involves evaluating the degree of the change, its progression over time, and visible trends (Table 2). The primary objective of the change-aware sequential instance segmentation task is to identify and track object instances within a set of 5 images, taken over different time intervals in the same location.

We further evaluate the effectiveness of the state-of-the-art methods for continuous change detection and captioning within the STVchrono dataset. In various experiments, we compared the performance of traditional methods and multimodal Large Language Models (LLMs) for change description, considering both image pairs and sequences. Our experimental results indicate that multimodal LLMs demonstrate superior accuracy in change description when compared to conventional non-LLM approaches. In addition, we assessed the state-of-the-art instance segmentation methods within the change-aware sequential instance segmentation task using the STVchrono dataset. Our findings reveal that even the most advanced methods still lag behind human performance, emphasizing the need to adapt these methods to continuously changing real-world scenarios.

## 2. Related Works

### 2.1. Change Understanding Datasets

Currently, available change understanding datasets primarily concentrate on two main tasks: the detection of changed regions within a scene and the linguistic description of the change content. The KTH Meta-rooms [12] and tvtable [1] datasets facilitate change detection in the robotics field between pairs of 3D point clouds of indoor rooms and tabletop surfaces, correspondingly. The Change3D [3], Panoramic Change Detection [4], and SOCD [5] datasets are suitable for the street-view scene recognition. While [3] consists of 3D point cloud pairs, [4] and [5] works in 2D and use semantic masks and bounding boxes for change detection, correspondingly. Another set of four datasets was recently proposed for 2D change detection: COCO-Inpainted, Synthtext-Change, Kubric-Change, and VIRAT-STD [6]. The 3DCD [2], EGY-BCD [13], and ChangeNet [14] datasets, aim for change detection in satellite remote sensing.

The CLEVR-Change [7] and CLEVR-Multi-Change [8] datasets focus on captioning single and multiple changes in synthetic image pairs, whereas the TRANCE [9] and OVT [10] datasets represent changes and their temporal orders using triples and graphs. Real-world datasets include Spot-the-Diff [11] (surveillance) and LEVIR-CC [15] (aerial imagery). Research also covers change detection in multi-view images [16] and 3D point clouds [17, 18]. Additionally, Weihs *et al*. [19] introduced a Visual Room Rearrangement task, where agents rearrange a room to its original layout by interacting with changed objects.

[*]https://developers.google.com/maps/documentation/streetview

| Dataset | Environment | # change pair | # city | Time span | Sequence length | Real image | Discrete change | Continuous change | Human-labeled caption | Change detection |
|---|---|---|---|---|---|---|---|---|---|---|
| Meta-rooms [12] | indoor | 588 | - | days | 2 | ✓ | ✓ | ✗ | ✗ | ✓ |
| Change3D [3] | outdoor | 866 | 1 | 4 years | 2 | ✓ | ✓ | ✓ | ✗ | ✓ |
| SOCD [5] | outdoor | 15,000 | - | - | 2 | ✗ | ✓ | ✗ | ✗ | ✓ |
| COCO-Inpainted [6] | in- & out-door | 60,000 | - | - | 2 | ✓ | ✓ | ✗ | ✗ | ✓ |
| Synthtext-Change [6] | outdoor | 5,000 | - | - | 2 | ✗ | ✓ | ✗ | ✗ | ✓ |
| Kubric-Change [6] | outdoor | 1,605 | - | - | 2 | ✓ | ✓ | ✗ | ✗ | ✓ |
| VIRAT-STD [6] | in- & out-door | 1,000 | - | hours | 2 | ✓ | ✓ | ✗ | ✗ | ✓ |
| 3DCD [2] | satellite | 472 | 1 | 7 years | 2 | ✓ | ✓ | ✓ | ✗ | ✓ |
| EGY-BCD [13] | satellite | 6,091 | 1 | 8 years | 2 | ✓ | ✓ | ✓ | ✗ | ✓ |
| ChangeNet [14] | satellite | 31,000 | 100 | 9 years | 6 | ✓ | ✓ | ✓ | ✗ | ✓ |
| CLEVR-Change [7] | table | 79,606 | - | - | 2 | ✗ | ✓ | ✗ | ✗ | ✗ |
| CLEVR-Multi-Change [8] | table | 60,000 | - | - | 2 | ✗ | ✓ | ✗ | ✗ | ✓ |
| Spot-the-Diff [11] | outdoor | 13,192 | 1 | hours | 2 | ✓ | ✓ | ✗ | ✓ | ✗ |
| LEVIR-CC [15] | satellite | 10,077 | 1 | 15 years | 2 | ✓ | ✓ | ✓ | ✓ | ✗ |
| STVchrono (our) | outdoor | 19,400 | 50 | 18 years | 2-6 | ✓ | ✓ | ✓ | ✓ | ✓ |

Table 1. Comparision of the STVchrono against existing change detection (top ten rows) and change description (four middle rows) datasets.

Existing datasets for change understanding often focus solely on detecting or describing discrete changes in static image pairs. In contrast, our STVchrono dataset captures continuous, gradual changes over time using sequences of 2-6 images and is created from historical photographs of 50 different cities around the world (Table 1).

### 2.2. Image Sequence Recognition Datasets

Alongside change understanding, various datasets support image pair or image sequence recognition tasks: NLVR [20] and NLVR2 [21] for difference reasoning, and GeneCIS [22] and VisualDNA [23] for image similarity. Similar to STVchrono, Mapillary [24] and [25] datasets utilize image sequences taken over different time periods for place recognition and robust aerial place representation, correspondingly. SatlasPretrain [26] is a temporal and spatial remote sensing dataset for remote sensing image analysis. In contrast, STVchrono focuses on identifying and describing regions of long-term continuous change. In this paper's figures, we employ images from the Mapillary dataset [24] rather than the STVchrono, due to Google Street View API restrictions.

### 2.3. Change Understanding Methods

State-of-the-art change captioning methods, such as DDLA [11], DUDA [7], MCCFormers [8], M-VAM [27] and CLIP4IDC [28] compute differences either at the pixel- [11] or feature-level [7] or use transformers [8, 27, 28] to correlate image pairs. Models like VARD-Trans [29] and SCORER [30] focus on identifying consistent features in images with viewpoint shifts. In the field of change detection task, two recent studies [6, 31] target identification of change regions with viewpoint differences. [6] introduces a co-attention-based approach for identifying correspondences between image pairs, while [31] relies on depth map generation for image correlations. Despite numerous existing methods, most of them focus on 2D image pairs or 3D data pairs and overlook serial-image change recognition. Recent studies highlight the potential of LLMs in context reasoning, but their application in change recognition remains unexplored. Our paper delves into change recognition in serial images, encompassing captioning, change region detection, and the usage of LLMs in this realm.

### 2.4. Instance Segmentation Methods

Instance segmentation aims to identify and outline distinct objects in visual content through pixel masks. The Mask R-CNN [32] method improved upon Faster R-CNN [33] by adding mask prediction. Subsequent methods like MaskFormer [34] incorporated transformer technology to enhance accuracy. Recent studies, such as Mask2Former [35], Mask DINO [36], and UNINEXT [37] have merged instance, semantic, and panoptic segmentation into unified models for simultaneous segmentation across various levels. Mask2Former has been adapted to 3D masked attention for video instance segmentation [38], while Ying *et al*. [39] propose the CTVIS method by adding a memory bank to maintain consistency across frames. Alternatives like Seq2Former [40] and DVIS [41] have developed trackers to preserve temporal continuity in image-level segmentation results. Our work introduces a change-aware instance segmentation for image sequences, tracking the evolution of natural scenes over years, thus extending beyond the typical short-term focus of existing video segmentation methods.

## 3. The STVchrono Dataset

The STVchrono dataset uniquely localizes and describes details of ongoing, extensive changes across space and time, going beyond the discrete changes (such as add, delete,

(a) Seasonal changes (b) Vegetation growth (c) Building exterior changes

(d) Building construction stages (d) Road and building maintenance and development

Figure 2. Different change types contained in the STVchrono dataset. This dataset encompasses a wide range of changes, including natural changes (*e.g.* (a) and (b)), as well as changes related to infrastructure and construction (*e.g.* (c), (d), and (e)).

or move) identified by current datasets (Table 1). It addresses both easily labeled discrete changes and complex continuous gradual shifts, which are hard to quantify with labels. Encompassing shifts in weather patterns, seasonal transitions, vehicular movement, and city architecture, the STVchrono dataset captures the dynamics of the real-world environments (Figures 1 and 2), supporting three distinct tasks related to these changes:

- **Continual Change Captioning (Image Pair)** aims at the recognition of the change details between 2 images taken at 2 distinct time periods (Figure 1, left). Examples of such changes can include the appearance of new cars, the removal of road signs, or a change in a building color.
- **Continual Change Captioning (Image Sequence)** focuses on the change tendencies over a sequence of 3-6 images taken at different time periods (Figure 1, middle). It offers insights into patterns, progressions, and trends over an extended time period, like the growth of plants.
- **Change-Aware Sequential Instance Segmentation** is suitable for the detection, understanding, and tracking of the change regions (Figure 3), ensuring a comprehensive analysis of the change dynamics for the specific object instances over a long time period.

### 3.1. Image Collection

The STVchrono dataset was collected using the Google Street View API. We chose Google Street View for its repository of images from diverse global locations captured over many years, enabling an in-depth analysis of temporal historical changes. Specifically, we selected images for 50 different cities, spanning 18 years: from 2006 to 2023 (Figure 1). We employed OpenStreetMap [†] to determine the boundaries of each city and then randomly sampled 300 to 1,000 latitude and longitude coordinates within these city limits. In the preliminary phase of the dataset creation, we retrieved all the images available for these coordinates (each with a resolution of 640x640 pixels). Subsequently, we excluded any images that contained projection-related distortions that made it hard for annotators to recognize changes and any coordinates that yielded fewer than 2 images. The resulting dataset comprises 71,900 photographs. Depending on the specific caption or detection task, we then handpicked and manually annotated the relevant images from the preliminary collection.

[†]https://nominatim.openstreetmap.org/search

### 3.2. Continual Change Captioning (Image Pair)

The goal of this task is to describe in detail the visual differences between two street view images taken at 2 different time periods. For this purpose, we selected 15,000 image pairs (a total of 30,000 images) from the STVchrono set of 71,900 images. We employed crowdsourcing platforms to gather human annotations in English. Each image pair received three to eight descriptive sentences detailing the dominant changes from one human annotator, while another annotator verified the effectiveness of these descriptions. An image pair annotation was approved only after the validation received from the second annotator.

Considering the possibility of numerous changes between two images, we focused on 10 dominant subjects commonly found in street view images, such as “weather” and “tree”, to be featured in the change captions. For each subject, the annotations specifically addressed the **distinction** in various aspects, including color, age, volume, or condition. The comprehensive annotation guidelines are presented in Table 2. Additionally, annotators were instructed to report dominant changes that go beyond the

| Subject | Attributes | Dataset example |
|---|---|---|
| Weather | Conditions, brightness, color | Image A is sunny, while image B is cloudy. **(IP, distinction)** |
| Tree | Growth pattern, color, volume, presence/absence | The tree on the right side becomes progressively thicker. **(IS, tendency)** |
| Building | Construction stages, age, cleanliness, heights, exterior alterations | Image 1 has the newest building on the left side. **(IS, superlative)** |
| Road | Age, cleanliness, width and volume, number and presence/absence of roads, cars, and traffic signs | In Images 1 and 2, a road is visible on the left; in Images 3 and 4, it disappears. **(IS, similarity)** |
| Lawn / Grassland | Color variations, volume, growth rates, transitions, presence/absence | The lawn on the right side looks greener in image B than in image A. **(IP, distinction)** |
| Soil / Land | Color variations, volume, transitions, presence/absence | The land on the right side of the sidewalk turned into a lawn from images 2 to 5. **(IS, tendency)** |
| River | Color variations, volume, transitions, presence/absence | The river is the cleanest in image 3. **(IS, superlative)** |
| Road fence | Age, color, cleanliness, height, presence/absence | The fence gate is not visible in image 1 but is present in images 2 and 3. **(IS, similarity)** |
| Human | Number, type and nature of activities, presence/absence | In image A, someone walks on the road; in image B, someone sits on a bench. **(IP, distinction)** |
| Animal | Number, type and nature of activities, presence/absence | There is a cat on the road in Image 3, but it is absent in the other images. **(IS, similarity)** |

Table 2. Annotation guidelines for the continual change captioning. The image pair task involves comparing two images, labeled A and B, to identify attribute distinctions. The image sequence task requires analyzing a series of 3-6 images to detail tendencies, superlatives, and similarities. The series start with the earliest image, designated as Image A and number 1 (IP: image pair; IS: image sequence).

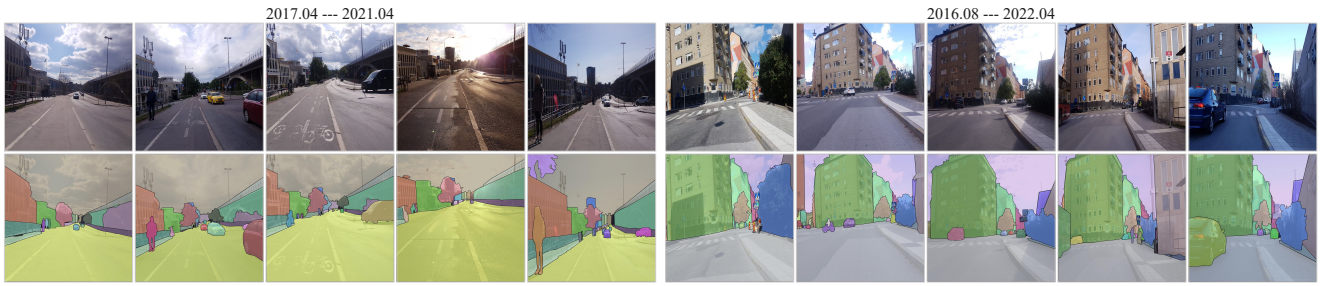


Figure 3. Two examples of image sequences (top) and their annotations (bottom) for the change-aware sequential instance segmentation task. Objects with consistent IDs share the same segmentation mask colors within each sequence.

guidelines, allowing for a more open-ended approach to change recognition.

### 3.3. Continual Change Captioning (Image Sequence)

The objective of the continual change captioning (image sequence) task is to narrate the progression of changes observed in a series of 3-6 images, captured at the same location over several years. From the 71,900 images in the STVchrono dataset, we utilized 19,800 images, divided into 4,400 sequences. These images were grouped into four categories, each containing 1,100 sequences with 3, 4, 5, and 6 images, respectively. We asked human annotators to focus on the same 10 change aspects identified in the continual change captioning (image pair) task. Each image sequence received annotations, which were then validated by two separate annotators. The annotations were directed to capture the **tendency, superlative, and similarity** in color, age, volume, or condition across various change aspects, as outlined in Table 2.

### 3.4. Change-Aware Sequential Instance Segmentation

The central goal of the consistent sequential instance segmentation task is to identify and track specific subject instances, within image sequences, captured at the same lo-

cation over different time intervals. We selected 520 sequences, representing a variety of cities and coordinates. Each sequence includes five images taken at different times (yielding a total of 2,600 images). Human annotators manually marked the instance regions and labels for each image. This task is particularly crucial for monitoring long-term trends such as the increase or decrease in vegetation, changes in river width, and the construction or demolition of buildings. A key challenge of this task is maintaining consistent instance labels for the same subjects despite their transformations over time. We provided labels for 12 subject categories, including vehicle (car_bus), building, tree, road, sky, lawn/grassland, soil/land, road fence, motorbike, bicycle, human, and animal. Two examples illustrating the task are shown in Figure 3.

### 3.5. Dataset Statistics

To ensure a comprehensive representation of street view changes, we selected 50 different cities from around the globe for our STVchrono dataset image collection. The distribution encompasses 14 cities in Asia, 13 in Europe, 8 in North America, 6 in South America, 6 in Oceania, and 4 in Africa. Istanbul was included in both the Asian and European tallies because of its transcontinental position. We split the dataset by cities into train and test sets, with ratios of 38/12 for image pair and sequence captioning, and 22/8 for segmentation task. For the two change caption tasks, the dataset boasts a vast range of vocabulary due to the fully human-annotated sentences. Specifically, the total vocabulary encompasses 1,223 unique words, with an average of 35.98 words per caption for the image pair task, and 50.65 words per caption for the image sequence task.

A comparative analysis of the STVchrono dataset with existing datasets is summarized in Table 1. The STVchrono dataset is the first of its kind to capture ongoing changes on a global scale (50 cities) and to consider the trends within sequences of images (2-6 images). It facilitates not only the detection of changes but also the recognition of change content through detailed human-labeled sentences. Additional details about the dataset, such as word and caption length distribution, time deltas distribution, a full list of the included cities, are available in the supplementary material.

## 4. Experiments

### 4.1. Baseline Methods

We evaluated the effectiveness of the existing state-of-the-art change captioning methods for both continual change captioning tasks (image pair and image sequence) using our STVchrono dataset. We conducted experiments using five change captioning methods: DUDA [7], MCCFormers-D, MCCFormers-S [8], CLIP4IDC [28], and VARD-Trans [29]. While recent studies highlight the effectiveness of LLMs in context reasoning, their incorporation into change recognition is still underexplored. Therefore, we decided to add two recent multimodal LLM-based methods for the comparison: OpenFlamingo [42] and BLIP2 [43] in conjunction with GPT4 [44].

As there are no existing methods for the change-aware sequential instance segmentation task, we selected two most closely related state-of-the-art video instance segmentation methods: Mask2Former [35, 38] and CTVIS [39]. We adapted these methods to track change instances (*e.g.* roads, trees, or buildings) in sequential images instead of tracking moving objects in videos. Experiments were conducted using various backbones: ResNet50 [45], ResNet101 [45], Swin Transformer Small and Large (SwinT-S, SwinT-L) [46] for the Mask2Former method, and both ResNet50 and SwinT-L for CTVIS.

### 4.2. Implementation Details

We used out-of-the-box implementations of DUDA [7], MCCFormers-D, MCCFormers-S [8], CLIP4IDC [28], and VARD-Trans [29] for the continual change captioning (image pair) task. For the continual change captioning (image sequence) task, we employed MCCFormers-S and CLIP4IDC, as both methods allow the sequencial input. We set the initial learning rate as $10^{-4}$ and adopted the Adam optimizer. All methods were trained for 80 epochs for captioning tasks, and 50 epochs for the segmentation task. For evaluation of OpenFlamingo [42] and BLIP2 [43] + GPT4 [44], we designed different prompts. The main paper shows the best results, while prompt design details for these multimodal LLMs are available in the supplementary material.

### 4.3. Evaluation Metrics

For evaluation of the generated change captions, we employed standard captioning metrics: BLEU4 [47] and CIDEr [48], assessing the similarity between generated and reference captions. Additionally, we used GPT4 [44] evaluation to focus more on meaning similarity over sentence structures. The number of sentences in the STVchrono dataset's ground truth captions is limited to 3-8 reference captions per image sequence. As this number might not be enough to describe all the changes within the image sequence, we further implemented human ratings to assess the accuracy and coverage of the generated captions manually. Accuracy is calculated as the proportion of correct change descriptions relative to total changes, while coverage is the average number of correctly captured changes per image sequence. Human ratings were provided for the randomly sampled 100 sequences for each evaluated method. For the evaluation of the generated instance segmentation masks, we used the standard Average Precision (AP) metric.

| Methods | Continual change captioning (image pair) | | | | | Continual change captioning (image sequence) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | BLEU4↑ | CIDEr↑ | GPT4↑ | Human rating | | BLEU4↑ | CIDEr↑ | GPT4↑ | Human rating | |
| | | | | Accuracy↑ | Coverage↑ | | | | Accuracy↑ | Coverage↑ |
| DUDA [7] | 21.7 | 39.1 | 26.3 | 32.7 | 1.1 | - | - | - | - | - |
| MCCFormers-D [8] | 22.4 | 52.7 | 29.8 | 39.8 | 1.34 | - | - | - | - | - |
| MCCFormers-S [8] | 25.4 | 51.3 | 26.8 | 35.9 | 1.28 | 19.5 | **39.3** | 13.7 | 22.7 | 0.67 |
| CLIP4IDC [28] | **28.5** | **69.5** | 32.4 | **47.8** | 1.74 | **20.0** | 26.0 | 9.5 | 13.0 | 0.48 |
| VARD-Trans [29] | 16.4 | 19.4 | 21.9 | 28.3 | 1.0 | - | - | - | - | - |
| OpenFlamingo [42] | 7.8 | 37.3 | **37.9** | 43.8 | **1.85** | 11.2 | 23.4 | 20.9 | **34.4** | **1.10** |
| BLIP2 [43] + GPT4 [44] | 4.2 | 16.1 | 33.1 | 37.5 | 1.52 | 4.9 | 7.5 | **30.3** | 21.3 | 1.02 |
| Human | 21.2 | 50.8 | 40.0 | 94.4 | 3.58 | 24.3 | 39.7 | 40.2 | 89.8 | 4.62 |

Table 3. Change description evaluation on continual change captioning (image pair: left side; image sequence: right side).


Figure 4. Experimental results of the existing methods in continual change captioning tasks (left: image pair; right: image sequence with four images). Changes correctly retrieved are highlighted in blue.

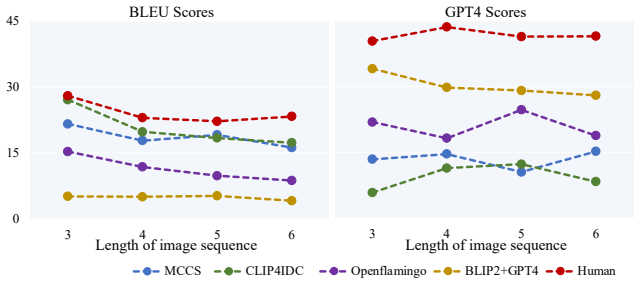


Figure 5. Experimental results on dataset examples with different sequence lengths (image numbers).

### 4.4. Continual Change Captioning (Image Pair)

The comparison of the selected baseline models and multimodal LLMs for this task is presented in Table 3 and Figure 4 (left side). Among all baselines (DUDA, MCCFormers -D and -S, CLIP4IDC, and VARD-Trans), CLIP4IDC achieves the highest BLEU4, CIDEr, and GPT4 scores, with 28.5, 69.5, and 32.4 points respectively. This performance is attributed to the large dataset size on which the model was pre-trained. OpenFlamingo and BLIP2+GPT4 show relatively low BLEU4 and CIDEr scores, while obtaining higher scores on GPT4 and human ratings. This is because these methods do not undergo a training process, tending to predict sentences with structures that differ from the ground truth sentences. In Figure 4, all methods capture only one to two changes. The highest human rating results come from CLIP4IDC and OpenFlamingo, but their best accuracy score of 47.8 and coverage score of 1.85 are extremely low, indicating that the models struggle to recognize changes within the images from the STVchrono dataset correctly.

### 4.5. Continual Change Captioning (Image Sequence)

Experimental results for this task are in Table 3 (averaged for all sequences from 3 to 6 images) and Figure 4 (right side). Like for the image pair continual change captioning task, multimodal LLM-based methods exhibited lower scores in BLEU4 and CIDEr, but achieved better GPT4 score and human ratings. Specifically, BLIP2 + GPT4 scored the highest in GPT4, while OpenFlamingo averaged

| Methods | Backbone | AP | AP50 | AP75 |
|---|---|---|---|---|
| Mask2Former [35, 38] | ResNet50 [45] | 4.60 | 6.73 | 4.52 |
| | ResNet101 [45] | 4.64 | 6.29 | 4.70 |
| | SwinT-S [46] | 6.02 | 8.34 | 6.47 |
| | SwinT-L [46] | 6.46 | 9.52 | 6.32 |
| CTVIS [39] | ResNet50 | 5.86 | 7.82 | 6.37 |
| | SwinT-L | **7.08** | **10.42** | **7.00** |

Table 4. Evaluation on the change-aware sequential instance segmentation task (SwinT-S, -L: swintransformer small, large).

Figure 6. Examples of the change-aware sequential instance segmentation results (from top to bottom: input images; ground truth; results from Mask2Former and CTVIS). Objects with the consistent IDs share the same mask colors within each sequence.

nearly 1.10 changes detected with higher accuracy (34.4 points). Figure 4 presents OpenFlamingo and BLIP2 + GPT4 correctly identifying changes. Compared to change recognition from image pairs, all methods demonstrated reduced performance, when recognizing changes from image sequences. Figure 5 depicts BLEU4 and GPT4 scores for varying sequence lengths. BLEU4 scores drop with the length increase, attributed to lengthier ground truth captions and diminished model efficiency in grasping complex structures. GPT4 scores stabilize, indicating a consistent complexity level in recognizing the change trends across 3 to 6 images. The performance gap compared to human accuracy highlights a deficiency in identifying temporal transitions in sequences, even for the advanced multimodal LLMs.

### 4.6. Change-Aware Sequential Instance Segmentation

The comparison of the chosen baseline models for the change-aware sequential instance segmentation task is present in Table 4. Among the two baselines, the CTVIS method achieved the highest Average Precision (AP) score across all thresholds (7.08 AP, 10.42 AP50, 7.00 AP75), when used with the SwinT-L backbone. Notably, even with the adoption of more extensive backbones like SwinT-L, the scores were not significantly improved. Examples of the generated instance segmentation masks for the chosen baseline models are present in Figure 6. Both Mask2Former and CTVIS exhibited low accuracy in identifying buildings with changing viewpoints, and in segmenting small regions like cars and humans. This is attributed to the unique challenges the STVchrono dataset poses, which include significant appearance changes due to factors like: construction, traffic, weather, seasons, and varying camera angles. These factors distinguish STVchrono from the typical tasks such as video instance segmentation, highlighting its complexity. The results underscore the need for ongoing innovation and the development of new approaches to improve robustness in the change-aware sequential instance segmentation. Additional experimental results are available in the supplementary material.

## 5. Conclusion

Continuous long-term change is a prevalent and fundamental element in the real-world observations, finding applications in areas like urban and land analysis, agriculture, and cultural heritage sites. However, most existing research in change recognition primarily centers on short-term, discrete changes and often relies on the synthetic datasets limited to two-image observation pairs. To advance the research in the real-world change recognition, we introduce "STVchrono", a novel benchmark dataset, comprising street view images for 50 cities spanning 18 years. This dataset particularly emphasizes long-term continuous changes and facilitates evaluations based on paired images, serial image change descriptions, and consistent instance segmentation, across images from the identical locations. Our experiments with the STVchrono reveal a significant performance gap between the latest multimodal LLMs and human capabilities, highlighting current advanced models' limitations in recognizing dynamic changes.

The STVchrono dataset, while groundbreaking, has its limitations, including uneven city data distribution and a restricted range of changes associated with the weather and time of day. We aim to continually refine and expand STVchrono, incorporating a broader variety of visual changes and detailed linguistic change descriptions. Currently, change recognition methods involve separate change description and region detection. The development of a comprehensive change recognition methodology, that seamlessly integrates change description and adaptive detection, presents a promising avenue for the future research.

## References


[1] Evan Herbst, Peter Henry, Xiaofeng Ren, and Dieter Fox. Toward object discovery and modeling via 3-d scene comparison. In *ICRA*, 2011. 2

[2] V Coletta, V Marsocci, and R Ravanelli. 3dcd: A new dataset for 2d and 3d change detection using deep learning techniques. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2022. 2, 3

[3] Tao Ku, Sam Galanakis, Bas Boom, Remco C Veltkamp, Darshan Bangera, Shankar Gangisetty, Nikolaos Stagakis, Gerasimos Arvanitis, and Konstantinos Moustakas. Shrec 2021: 3d point cloud change detection for street scenes. *Computers & Graphics*, 2021. 2, 3

[4] Ken Sakurada and Takayuki Okatani. Change detection from a street image pair using cnn features and superpixel segmentation. In *BMVC*, 2015. 2

[5] Kento Doi, Ryuhei Hamaguchi, Yusuke Iwasawa, Masaki Onishi, Yutaka Matsuo, and Ken Sakurada. Detecting object-level scene changes in images with viewpoint differences using graph matching. *Remote Sensing*, 14(17):4225, 2022. 2, 3

[6] Ragav Sachdeva and Andrew Zisserman. The change you want to see. In *WACV*, 2023. 2, 3

[7] Dong Huk Park, Trevor Darrell, and Anna Rohrbach. Robust change captioning. In *ICCV*, 2019. 2, 3, 6, 7

[8] Yue Qiu, Shintaro Yamamoto, Kodai Nakashima, Ryota Suzuki, Kenji Iwata, Hirokatsu Kataoka, and Yutaka Satoh. Describing and Localizing Multiple Changes with Transformers. In *ICCV*, 2021. 2, 3, 6, 7

[9] Xin Hong, Yanyan Lan, Liang Pang, Jiafeng Guo, and Xueqi Cheng. Transformation driven visual reasoning. In *CVPR*, 2021. 2

[10] Yue Qiu, Yanjun Sun, Fumiya Matsuzawa, Kenji Iwata, and Hirokatsu Kataoka. Graph representation for order-aware visual transformation. In *CVPR*, 2023. 2

[11] Harsh Jhamtani and Taylor Berg-Kirkpatrick. Learning to describe differences between pairs of similar images. In *EMNLP*, 2018. 2, 3

[12] Rareş Ambruş, Nils Bore, John Folkesson, and Patric Jensfelt. Meta-rooms: Building and maintaining long term spatial models in a dynamic world. In *IROS*, 2014. 2, 3

[13] Shimaa Holail, Tamer Saleh, Xiongwu Xiao, and Deren Li. Afde-net: Building change detection using attention-based feature differential enhancement for satellite imagery. *IEEE Geoscience and Remote Sensing Letters*, 2023. 2, 3

[14] Deyi Ji, Siqi Gao, Mingyuan Tao, Hongtao Lu, and Feng Zhao. Changenet: Multi-temporal asymmetric change detection dataset. *arXiv preprint arXiv:2312.17428*, 2023. 2, 3

[15] Chenyang Liu, Rui Zhao, Hao Chen, Zhengxia Zou, and Zhenwei Shi. Remote sensing image change captioning with dual-branch transformers: A new method and a large scale dataset. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–20, 2022. 2, 3

[16] Yue Qiu, Yutaka Satoh, Ryota Suzuki, Kenji Iwata, and Hirokatsu Kataoka. 3d-aware scene change captioning from multiview images. *IROS*, 2020. 2

[17] Yue Qiu, Yutaka Satoh, Ryota Suzuki, Kenji Iwata, and Hirokatsu Kataoka. Indoor scene change captioning based on multimodality data. *Sensors*, 20(17):4761, 2020. 2

[18] Yue Qiu, Shintaro Yamamoto, Ryosuke Yamada, Ryota Suzuki, Hirokatsu Kataoka, Kenji Iwata, and Yutaka Satoh. 3d change localization and captioning from dynamic scans of indoor scenes. In *WACV*, 2023. 2

[19] Luca Weihs, Matt Deitke, Aniruddha Kembhavi, and Roozbeh Mottaghi. Visual room rearrangement. In *CVPR*, 2021. 2

[20] Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. A corpus of natural language for visual reasoning. In *ACL*, 2017. 3

[21] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. In *ACL*, 2019. 3

[22] Sagar Vaze, Nicolas Carion, and Ishan Misra. Genecis: A benchmark for general conditional image similarity. In *CVPR*, 2023. 3

[23] Benjamin Ramtoula, Matthew Gadd, Paul Newman, and Daniele De Martini. Visual dna: Representing and comparing images using distributions of neuron activations. In *CVPR*, 2023. 3

[24] Frederik Warburg, Soren Hauberg, Manuel Lopez-Antequera, Pau Gargallo, Yubin Kuang, and Javier Civera. Mapillary street-level sequences: A dataset for lifelong place recognition. In *CVPR*, 2020. 3

[25] Utkarsh Mall, Bharath Hariharan, and Kavita Bala. Change-aware sampling and contrastive learning for satellite images. In *CVPR*, 2023. 3

[26] Favyen Bastani, Piper Wolters, Ritwik Gupta, Joe Ferdinando, and Aniruddha Kembhavi. Satlaspretrain: A large-scale dataset for remote sensing image understanding. In *ICCV*, 2023. 3

[27] Xiangxi Shi, Xu Yang, Jiuxiang Gu, Shafiq Joty, and Jianfei Cai. Finding it at another side: A viewpoint-adapted matching encoder for change captioning. In *ECCV*, 2020. 3

[28] Zixin Guo, Tzu-Jui Wang, and Jorma Laaksonen. CLIP4IDC: CLIP for image difference captioning. In *ACL*, 2022. 3, 6, 7

[29] Yunbin Tu, Liang Li, Li Su, Junping Du, Ke Lu, and Qingming Huang. Adaptive representation disentanglement network for change captioning. *TIP*, 2023. 3, 6, 7

[30] Yunbin Tu, Liang Li, Li Su, Zheng-Jun Zha, Chenggang Yan, and Qingming Huang. Self-supervised cross-view representation reconstruction for change captioning. In *ICCV*, 2023. 3

[31] Ragav Sachdeva and Andrew Zisserman. The change you want to see (now in 3d). In *ICCVW*, 2023. 3

[32] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 3

[33] Ross Girshick. Fast r-cnn. In *ICCV*, 2015. 3

[34] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. 2021. 3

[35] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2022. 3, 6, 8

[36] 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. In *CVPR*, 2023. 3

[37] Bin Yan, Yi Jiang, Jiannan Wu, Dong Wang, Ping Luo, Zehuan Yuan, and Huchuan Lu. Universal instance perception as object discovery and retrieval. In *CVPR*, 2023. 3

[38] Bowen Cheng, Anwesa Choudhuri, Ishan Misra, Alexander Kirillov, Rohit Girdhar, and Alexander G Schwing. Mask2former for video instance segmentation. *arXiv preprint arXiv:2112.10764*, 2021. 3, 6, 8

[39] Kaining Ying, Qing Zhong, Weian Mao, Zhenhua Wang, Hao Chen, Lin Yuanbo Wu, Yifan Liu, Chengxiang Fan, Yunzhi Zhuge, and Chunhua Shen. Ctvis: Consistent training for online video instance segmentation. In *ICCV*, 2023. 3, 6, 8

[40] Junfeng Wu, Yi Jiang, Song Bai, Wenqing Zhang, and Xiang Bai. Seqformer: Sequential transformer for video instance segmentation. In *ECCV*, 2022. 3

[41] Tao Zhang, Xingye Tian, Yu Wu, Shunping Ji, Xuebo Wang, Yuan Zhang, and Pengfei Wan. Dvis: Decoupled video instance segmentation framework. In *ICCV*, 2023. 3

[42] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023. 6, 7

[43] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 6, 7

[44] R OpenAI. Gpt-4 technical report. *arXiv*, pages 2303–08774, 2023. 6, 7

[45] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6, 8

[46] 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 *ICCV*, 2021. 6, 8

[47] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002. 6

[48] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, 2015. 6