

Visual Chronicles: Using Multimodal LLMs to Analyze Massive Collections of Images

Supplementary Material

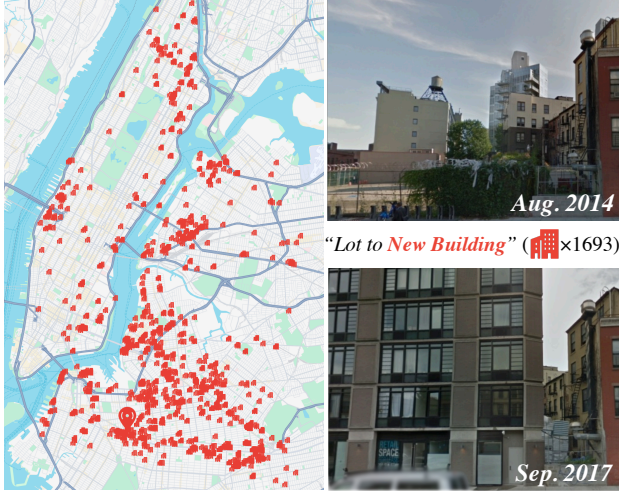


Figure 7. **Where are new buildings in NYC?** We plot all locations where we observe the trend, “A lot now has a multi-story building constructed on it”, in NYC. We find notable clustering in Brooklyn, Long Island City, and Hudson Yards.

In this supplementary, we provide more application results including an example of a non-temporal query and a spatial insight (Sec. 7), more details about our data, system, and evaluations (Sec. 8), more evaluations including ablations and complete results of MLLM baselines on the hyper-parameters of Visual Chronicles (Sec. 9), and potential societal impacts of Visual Chronicles (Sec. 11).

7. Extended Application Results

7.1. Where are new buildings in NYC?

Aside from the analysis for specific temporal period of interest (Sec. 5.2) or semantic subject of interest (Sec. 5.3), Visual Chronicles can also reveal insights about the spatial domain, *e.g.* by visualising the spatial distribution of a trend. In particular, in this study, we showcase the spatial distribution of one of the trends of the decade in NYC (Sec. 5.1), namely “A lot now has a multi-story building constructed on it”. As shown in Fig. 7, we find that more new buildings are observed in Brooklyn and Long Island City, compared to sparser, scattered points in Manhattan. Additionally, the well-known Hudson Yards development [55] appears as a rare spatial cluster in Manhattan.

7.2. Many Graffiti Observed Since 2020

As shown in Fig. 8, our system also identified the trend “New graffiti were added to the wall” as a significant trend since

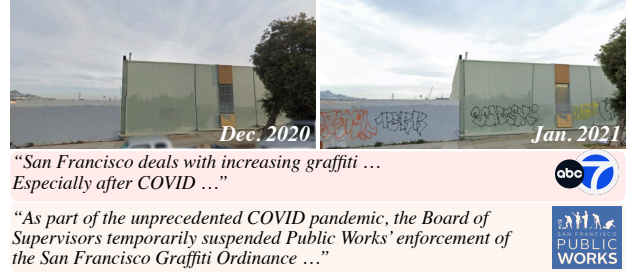


Figure 8. **Many Graffiti Observed, 2020–2022.** In SF since 2020, many new graffiti are observed. A quick search on the Internet reveals relevant reports in local news and a local policy change. The enforcement has since resumed.

2020. Such changes were spotted 3152 times post-2020 and 1150 times pre-2020. As reported by [ABC News](#), during COVID San Francisco paused enforcement of local regulations requiring building owners to remove graffiti promptly. Yet, this is an example of why we must be careful when drawing conclusions from this data, as at least two hypotheses seem compatible: 1) graffiti tagging became more popular, or 2) graffiti was removed less frequently, increasing its likelihood of being observed in our data.

7.3. Additional Trend Results

More Images of Shown Trends. In previous sections, we have shown example trends discovered in Visual Chronicles, each evidenced by a pair of images. In Fig. 9, we further provide an additional pair of images illustrated visual evidence for each of these previously shown trends, demonstrating the variation in the visual data.

More Trends in Images. In addition, we show in Fig. 10 more trends found using Visual Chronicles along with accompanying visual evidence. Such trends include a specific “canopied outdoor dining” trend in NYC, a “new zebra crosswalk” painted on the street, an added “under-bridge wooden plank”, an added green “bike lane”, and the opening of “new cafés”.

More Trends in Text. In previous figures, we have shown all trends with visual evidence. For visualization purpose, the shown text is shortened descriptions of the exact trend found in Visual Chronicles. In Fig. 11 and Tab. 4, we first show a list of complete text for all the trends we have shown. Moreover, we show an addition set of found trends also in complete text descriptions.

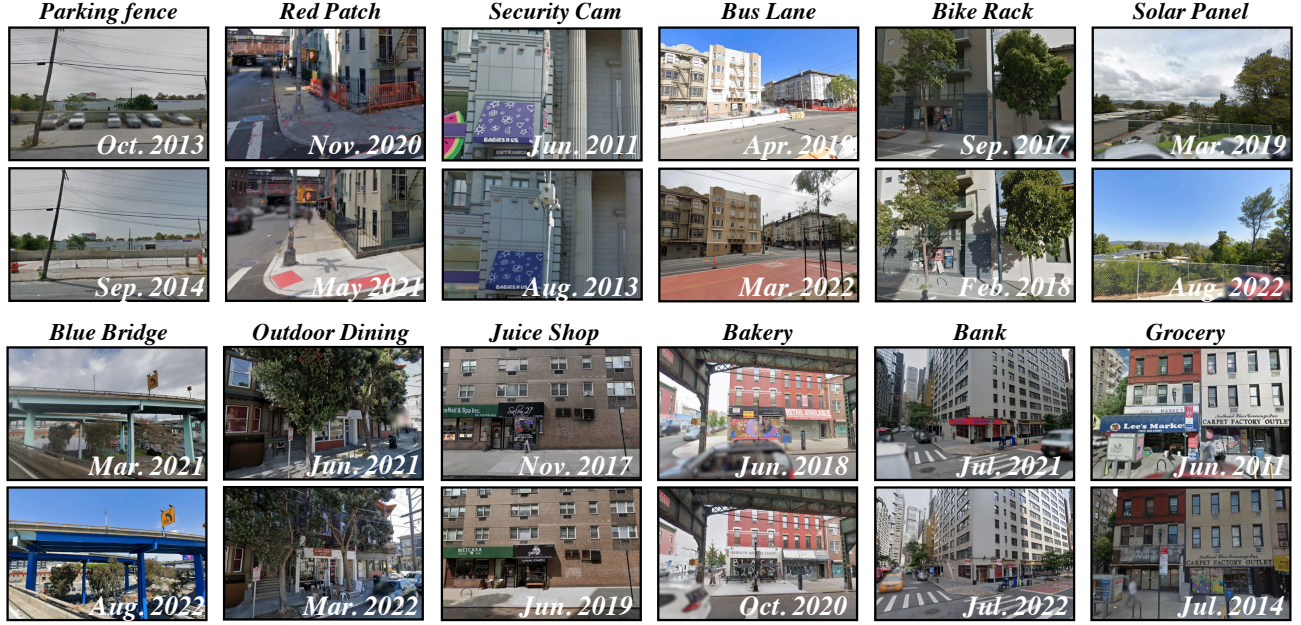


Figure 9. More Images of Shown Trends.

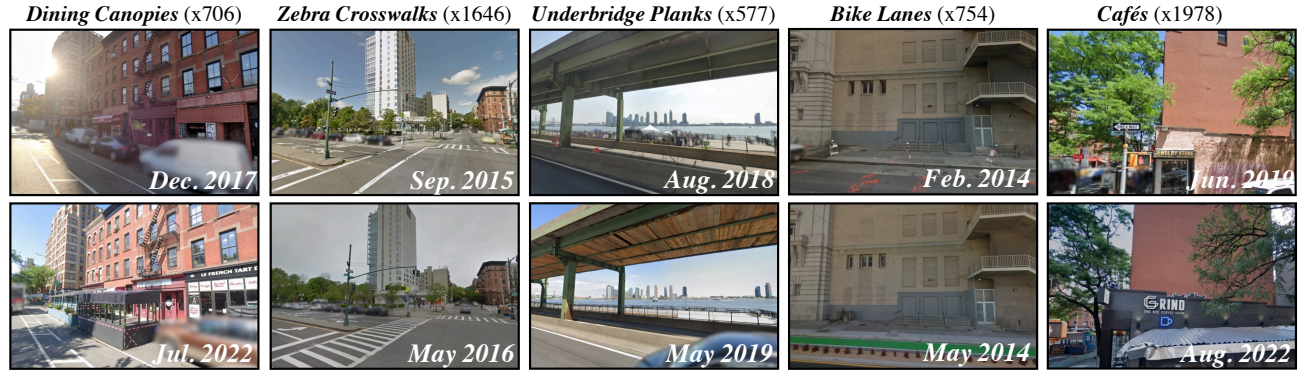


Figure 10. More Trends in Images.

8. Extended Details

8.1. Computation Costs.

Our change detection step takes < 24 hours for analysing 20M images from 1 city. As in L457, our aggregation step takes 4.6 hours for 200 trends. Both steps use 64 MLLMs running in parallel. The total cost of MLLM inference is $< 10K\$$, well within the budget of commercial city-scale analysis projects.

8.2. Data Preprocessing

To produce the sets of local image sequences for the local change detection step of our system, we first randomly sample a set of seed locations from the dataset. For each seed location, we then sample a view direction that’s likely to have buildings in view, based on the map. Finally, for each

panorama centered within 1.8 meters of the seed location, we reproject the panorama to the sampled view direction as a pinhole image and add the resulting image to the set. For a city-scale dataset with 20M images (e.g., NYC), this process results in a manageable set of 10 to 40 images in each set at $\sim 900K$ sampled locations.

8.3. Dataset Coverage

We collect our data from two major US cities, San Francisco and New York City. In particular, we sample locations from a selected region of each city, as shown in Fig. 12. Note that the regions are selected to represent busy, dense regions of the city for our study of trends. They do not cover the whole city by the definition of range of city administration. Yet, within the selected regions we already collect 20M images for our massive scale study.

New York City

1. (×1635) The street in front of the building had a dedicated red-painted lane, likely for buses or bicycles, added to it.
2. (×1422) Window air conditioning units were installed in windows.
3. (×1360) Benches appeared in front of a building where there were previously no benches.
4. (×1031) The Halloween decorations, including pumpkins or inflatable figures, were removed from a building.
5. (×935) The light pole had a box attached that wasn't there before.
6. (×865) A "No Parking Anytime" sign has been installed in front of the building.
7. (×780) The roadway changed from being relatively smooth and undamaged to showing signs of wear and tear, with visible cracks and patches.
8. (×656) The storefront added a banner advertising space available for lease.

San Francisco

1. (×1677) A small sapling has been planted in front of the building where there was no tree or sapling before.
2. (×1502) The house's front door had a Christmas wreath added to it.
3. (×1403) A traffic light was added to the intersection of two streets.
4. (×1357) The storefront of the building had its storefront windows covered.
5. (×1355) A building now has an American flag hanging from it.
6. (×760) The storefront added a "for lease" sign in the window.
7. (×678) The storefront had its window display covered by a metal security gate.
8. (×527) A basketball hoop was added around the building.

Figure 11. More Trends in Text.

	Trend Description
Fig. 1	(×1483) The storefront added tables and chairs outside. (×481) The support of the overpass was painted blue.
Fig. 3	(×745) Security cameras became visible on the light poles. (×509) The parking lot in front of the building now has a fence enclosing it. (×519) The crosswalk had its marking changed from white to red.
Fig. 4	(×1504) Solar panels were added to the roof of one of the buildings in the background. (×751) The street has been reconfigured from a standard two-way configuration to a configuration with a dedicated bus lane. (×1799) Bike racks were added in front of a building.
Fig. 5	(×318) A juice shop opened at the storefront. (×512) There is a bakery newly opened at the storefront. (×1614) The bank branch at the storefront closed. (×741) The grocery store on the street closed.
Fig. 7	(×1693) A lot now has a multi-story building constructed on it.
Fig. 8	(×3152) New graffiti were added to the wall.
Fig. 10	(×706) An outdoor seating area with a canopy was added in front of the buildings. (×1646) The crosswalk changed from having solid white painted lines to zebra-striped painted lines. (×577) Portions of the underside of the overpass structure are covered with wooden planks, where it was primarily metal beams and supports before. (×754) A green bike lane was added to the street in front of a building. (×1978) The storefront changed into a cafe.

Table 4. Full Texts for Trends in Various Figures.

8.4. Local Change Detection Details

Data Preparation. For the first step, local change detection, in Visual Chronicles, we first sample a set of locations from the city of interest. Since the street view imagery we use is sampled from capture vehicles, we naturally have all the captured locations scattered on the street instead of inside of buildings. We hence directly sample a set of locations from the capture locations. To avoid redundancy caused by sampling nearby points, we implement Non-Maximum Suppression (NMS). Specifically, we first calculate the number of nearby images (as defined in Sec. 3.3) for each location. Then we rank all locations by number of neighbors in decreasing order, and apply NMS to select non-redundant locations.

After selecting all locations, we collect local neighboring images following the process outlined in Sec. 3.3. Note that we filter out locations that have fewer than 10 local images at the end of this preparation step.

Detection Prompting. To detect local changes using MLLMs, we essentially feed a sequence of images, in chronological order, to the MLLM and ask it what changes it detects and from which images. We use Gemini-1.5 Pro (Gemini) as the MLLM here. We find that such prompts already work surprisingly well on Gemini. Gemini would describe significant temporal semantic changes it notices, including changes minute in visual scale such as the tiny security camera in Fig. 3. At the same time, it often voluntarily ignores semantically insignificant changes such as illumina-

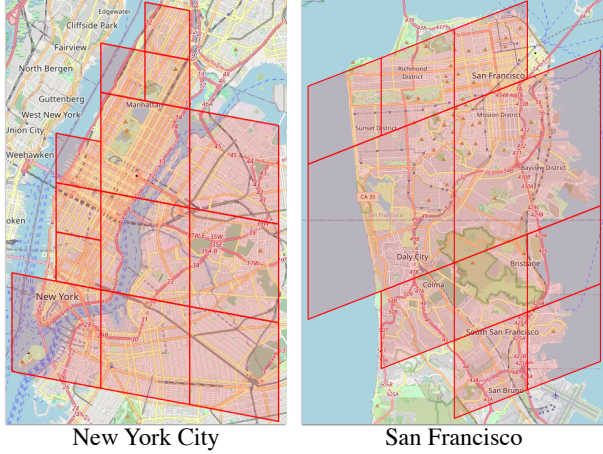


Figure 12. **Data Coverage.** We show the region we sample our locations from in red. Note that our data collection does not coverage the whole region of administration for both cities.

tion changes or camera displacement changes (as our camera poses are not perfectly aligned). Yet, through our empirical study, we find a few error patterns of such simple prompting. Among them, the notable ones include: 1) Gemini may hallucinate changes that can not be certainly seen in images. 2) The answers occasionally include changes that are caused by camera wiggles, normally in the periphery of the frame. 3) Sometimes Gemini would mistake color changes caused by illumination change as a permanent change, *e.g.* change from a dark color in a shadow to a very bright color under the sun. 4) Very rarely, Gemini may mistake the foreground object with the background, *e.g.* mistake the sign and text on the car in front of a building as text and sign on the building. To address these, we articulate a more extensive list of instructions in the prompt shown in Fig. 17. We include all the error patterns we empirically find in this prompt.

Self-Critic. Despite the careful prompting, we still find Gemini making mistakes such as hallucination in some of the answers. Inspired by OpenAI o1, we implement a simple Self-Critic strategy by feeding the visual evidence and the detected change back to Gemini and ask it to verify that the change is significant and that it can be found in the visual evidence. We find such a simple verification step greatly improves the precision of local change detection, as shown in Sec. 9.4.

8.5. Trend Discovery Details

Trend Proposal. In the second step of our system, we derive trend proposals from detected local changes. This step has 2 parts. First, we design a prompt that make MLLMs deduct potential proposals from a detected change, as shown in Fig. 18. Then, we implement a standard Canopy Clustering [39] using text embedding distance and set the tight distance threshold to 0.15 and the loss distance threshold

to 0.2. Because our verification needs at least k nearest neighbours in cluster, we filter clusters from the Canopy Clustering that has less than k elements. After the filtering, each cluster center becomes a trend proposal, and all its derived trends text will be candidates for trend verification.

Verification Prompting. We develop a detailed prompt promoting MLLM (in our system, Gemini) to verify if a change belongs to a trend proposal, shown in Fig. 19. In particular, we leverage the in-context learning ability of MLLMs, by including a small set of examples in the prompt. We empirically find that such examples improve the verification quality.

8.6. Evaluation Details

For our evaluation in Sec. 4.2 and Sec. 4.3, we sample the test set from our data collection from New York City. For the evaluation of change detection in Tab. 1, we randomly sample 200 locations in New York City and ask a human labeller to describe a list of changes for each location and from which 2 consecutive images the changes were found, as the ground truth changes. Given such labels, we can derive if there’s a change between any 2 consecutive images, which is the ground truth we use to calculate the Average Precision in Tab. 1. In our evaluation of trend discovery in Tab. 2, we aim to collect 50 clusters each with 40 candidate changes to verify. However, naively randomly sampling such a collection would lead to mostly negative labels, that is, the change does not belong to the trend. To better balance the test labels, we over-sample 75 clusters, with a selection heuristic of enough nearby neighbours, *i.e.* more than N changes in the pool with text embedding distance smaller than 0.25. Thereafter, to sample balanced change candidates, we over-label change changes until we collect at least 20 positive labels for the cluster, *i.e.* 50% positive rate. We also apply early-stop, *i.e.* setting the maximum number of changes to label for each cluster, to avoid excessive labelling for one cluster. After labelling all 75 clusters, we pick 50 clusters with the most balanced test labels as our final test set. For Tab. 3, since NV-Emb is predicting continuous distance values, we grid-search 1000 thresholds to binarize its predictions and pick the highest accuracy to report.

9. Extended Evaluations

9.1. Validity of Evaluation Data

Our analysis is conducted on massive-scale datasets. It’s impractical to annotate even 1% of the 20M images in the datasets. We instead designed an evaluation set large enough to provide comparisons of different methods with low variance, with conclusions generalizable to larger-scale applications. To verify this, we computed the standard deviation of metrics using 1K random 75%-subsets of the evaluation set, and found them to be $\sim 1\%$ in AP for both subtasks.

MLLM	Change Detection	Trend Discovery
LLAVA [34]	64.31%	81.11%
Gemini [45]	76.56%	86.63%

Table 5. **Ablation on Different MLLMs** in AP. The smaller MLLM, LLAVA, still performs decently on both sub-problems, while being slightly worse than more powerful Gemini.

	Without Self-Critic	With Self-Critic
Precision	57.83%	81.74%
Recall	91.50%	89.87%

Table 6. **Ablation on Self-Critic in Local Change Detection.** The Self-Critic strategy (Sec. 8.4) significantly improves the precision of local change detection.

Method	$k = 2N$	$k = 3N$	$k = 4N$	$k = 5N$
Acc.@50	88.6%	93.9%	95.9%	97.2%
Acc.@100	91.4%	94.6%	96.1%	97.1%
Acc.@200	95.7%	98.3%	99.2%	99.5%

Table 7. **Ablation of Hyperparameters of Our Hybrid Solution in Trend Discovery.** k is the number of nearest neighbours our hybrid solution verify using MLLMs. N is the frequency threshold of a trend. We choose $k = 3N$ for all our experiments for its balance between accuracy and efficiency.

9.2. Ablation Studies

We conducted extensive ablation studies evaluating design choices of our system, including MLLMs for change detection (Sec. 4.2 & Tab. 1), prompting strategy in change detection (Sec. 9.4 & Tab. 6), MLLMs for trend discovery (Sec. 4.3 & Tab. 2), hybrid solutions for scaling (Sec. 4.3 & Tab. 3), its hyperparameters (Sec. 9.5 & Tab. 7), and alternative MLLMs (Sec. 9.3 & Tab. 5).

9.3. Alternative MLLMs

We investigate the impact of using different MLLMs on the effectiveness of our system. To this end, we replace Gemini with a smaller open-sourced MLLM, LLAVA [34], and evaluate its performance on both sub-problems in our system. We find that it still produces reasonably accurate analysis while being slightly worse than Gemini.

9.4. Self-Critic Prompting in Change Detection

As mentioned in Sec. 8.4, we find that applying a round of self-critic to local change detection greatly improves the precision of the detection. In Tab. 6, we quantitatively evaluate the impact of self-critic. We find that it boosts the precision by a significant margin of more than 33% while not degrading much on the recall.

9.5. Hyperparameters in Trend Discovery

Our hybrid solution to trend discovery requires setting a hyper-parameter k , *i.e.* the number of nearest changes for which we use MLLMs to verify regarding a trend proposal. In Tab. 7, we show an ablation study of varying the value of k , as multiples of the trend threshold N . We find that as k increases, the accuracy goes up as expected, which comes at a cost of more MLLM inferences and slower verification. However, $k = 3N$ can already produce satisfying accuracy ($> 90\%$ across the board). Hence, we choose it as a good balance between accuracy and speed.

9.6. MLLMs Without Images

Complete Answers from MLLMs. As mentioned in Sec. 4.1, we find MLLMs’ answers without looking at any images unsatisfactory. Such answers are often abstract and lack intriguing details. In Fig. 13, we illustrate the complete answers from MLLMs to the prompting questions, “What are the trends of changes happened in [NYC / SF] that we can see visually in 2011-2023?”. Here we use Gemini-1.5 Pro for queries for both cities.

Alternative Prompts and MLLMs. We also vary the wording of the prompts in search for potentially better alternatives, such as replacing “changes ... that we can see visually” with “visual changes” or simply “changes”. Yet, we empirically find the results not varying much in quality, except we see more abstract trends, *e.g.* “Increased focus on sustainability”, when using simply “changes”. Additionally, we show in Fig. 14 the results from an alternative MLLM, namely GPT-4o. We find that GPT-4o focuses more on unique changes, such as “construction of numerous super-tall buildings ... along 57th Street ... known as Billionaire’s Row.”, but misses more frequent trending changes such as bike lane expansions, which both Gemini and Visual Chronicles identify as a significant trend. We also notice that to answer the query, GPT-4o would search online and retrieve relevant information, both image and text, in the answering process, and even output a few images in the output, as shown in Fig. 14. Yet these images are not strictly all visual evidence for the discovered trends, unlike in Visual Chronicles.

9.7. MLLMs Processing All Images

In Sec. 4.1, we discussed the context size as MLLMs’ major limitation. Due to such constraint, it is infeasible to feed all our images to MLLMs. Instead, in Fig. 15, we sub-sample the maximum subset of our images that can be fed into MLLMs, ranging from 7K to 8K images because we feed complete image sequences to MLLMs and such sequences vary in lengths. Specifically, we sample 4 such subsets and take the union of the answers as the final answer shown in Fig. 15. Note that we only sample 4 subsets for increasing the number of subsets mainly creates duplicate trends.

9.8. MLLMs Processing All Changes

As mentioned in Sec. 4.3, trend discovery is not trivial to MLLMs for simply feeding all change descriptions to MLLMs is not adequate. In Fig. 16, we show the complete results of feeding as many changes as possible to MLLMs and ask for trends among them. Here we sample 50K changes for each batch to feed to MLLMs, from our pool of 3M changes. We sample 4 batches in total and take the union of answers from each batch as the final answer shown in Fig. 16. Note that we only sample 4 batches because in our observations, more batches mostly create redundancy in found trends.

9.9. Hybrid Trend Discovery Method and RAG

Our hybrid solution to trend discovery bears some connections to Retrieval-Augmented-Generation (RAG [30]). Particularly, both use embeddings to filter relevant examples from a database. However, ours is fundamental different from RAG in a few aspects: 1) Our goal is different from RAG. We aim to find salient patterns of examples in the database, whereas RAG retrieves examples to assist question answering. 2) The scale of relevant examples is different. In our task, we mark thousands of examples in the database as relevant, while RAG normally select several most relevant examples. 3) We use relevant examples differently from RAG. RAG often combines all relevant examples together as context to LLM or MLLMs. In contrast, we process each relevant example independently to verify visual trend proposals.

10. Limitations and Failure Cases

As described in the Conclusion section of the main paper, our approach has limitations that lead to failure cases and topics for future work. The following provides a few examples.

The proposed approach is natural when the input image collection can be decomposed into small sets for MLLM analysis. We utilized image poses to create sets of images at nearby viewpoints from which local changes can be detected. However, this method would not work well for detecting all types of trends, including ones that involve global changes or ones that cannot be observed in images.

We rely upon simple prompts to guide the MLLM to produce descriptions of local changes that we expect people would find interesting. However, other types of changes might be of interest in different applications (e.g., to find new dirt on the sidewalk). That would require prompt engineering to get the MLLM to detect certain types of changes – further study is required to discover how readily good prompts can be found.

Our formulation produces sets of visual changes that happened multiple times at different locations, but does not reason about how often the reverse change happened, and so it sometimes reports changes that are the result of tran-

sient activities (e.g., the addition of Halloween decorations), which are not in fact long-term trends. Similarly, our system does not measure the fraction of times a change happened for the same starting state, which could be valuable for studying certain types of trends. Adding these capabilities is an interesting topic for future studies.

11. Potential Societal Impacts

To our best knowledge, Visual Chronicles marks the first proof-of-concept of using MLLMs to analyze millions of images. In the context of searching for trends of visual changes in the urban environment, our studies show that a system like ours could potentially facilitate studies in urban planning, civil engineering, or public policy, by automatically providing insights from massive scale dataset. Nevertheless, as we mentioned in our discussion on our results, it’s vital to keep aware of the potential bias in our data, hence being cautious in drawing conclusions that demand statistical soundness. Particularly, careful uses of systems of this kind are critical in battling misinformation. Our focus on providing verifiable visual evidence for any trend proposed by our system can be seen as part of such efforts.

Prompt: What are the trends of changes happened in **New York City** that we can see visually in 2011-2023?

MLLM (Gemini 1.5) Answers: Visually noticeable changes in New York City between 2011 and 2023 include:

Infrastructure & Development:

- **Rise of Supertalls:** The proliferation of extremely tall, slender skyscrapers, particularly in Midtown Manhattan (e.g., 432 Park Avenue, 111 West 57th Street) and Downtown Brooklyn, dramatically altered the skyline.
- **Hudson Yards Development:** The massive Hudson Yards project on Manhattan's west side, with its distinctive Vessel structure and surrounding high-rises, became a completely new neighborhood, visually impacting the cityscape.
- **Second Avenue Subway:** The long-awaited Second Avenue Subway extension finally opened its first phase, visibly changing the streetscape along Second Avenue on the Upper East Side with new station entrances.
- **Brooklyn Waterfront Transformation:** Continued development along the Brooklyn waterfront, from Williamsburg to DUMBO to Greenpoint, resulted in new parks, residential buildings, and commercial spaces, reshaping the view from Manhattan.
- **Bike Lane Expansion:** A significant increase in dedicated bike lanes throughout the city is a visible change, altering street markings and traffic patterns.

Social & Cultural:

- **Increased Outdoor Dining:** The COVID-19 pandemic led to a boom in outdoor dining structures, from simple sheds to elaborate setups, significantly changing the look and feel of many streets. While some have been removed, many remain and have become a more permanent fixture.
- **Public Art Installations:** Larger and more frequent public art installations have become more common in parks and public spaces, adding visual interest to the city.
- **Retail Landscape Changes:** The rise of online shopping and the pandemic impacted brick-and-mortar stores, resulting in visible vacancies and changes in storefront usage in some areas.
- **Greening Initiatives:** Efforts to increase green space, including park expansions and rooftop gardens, are slowly but surely adding more visible greenery to the city.

Less Visible but Noteworthy:

- **Demographic Shifts:** While not immediately visually obvious, changes in demographics (gentrification, population growth/decline in certain neighborhoods) influence the types of businesses, public services, and overall aesthetic of different areas over time.
- **Security Measures:** Increased security measures after 9/11 and other events, though not always visually prominent, have subtly altered public spaces and infrastructure.

To truly appreciate these changes, comparing photos of specific locations from 2011 and 2023 can be particularly revealing. Google Street View's historical imagery feature is a useful tool for this purpose. You'll see the transformation of specific streetscapes and neighborhoods unfold before your eyes.

Prompt: What are the trends of changes happened in **San Francisco** that we can see visually in 2011-2023?

MLLM (Gemini 1.5) Answers: Visually noticeable changes in San Francisco between 2011 and 2023 include:

Increased Construction and Development:

- **High-rise buildings:** The skyline has grown denser, particularly in areas like South of Market (SoMa), Mission Bay, and Transbay. Many new residential and commercial high-rises have been constructed.
- **Crane activity:** While fluctuating, crane activity has been generally high throughout this period, a visible sign of ongoing construction.
- **Infilling:** Vacant lots and older, smaller buildings have been replaced by larger, denser structures in many neighborhoods.

Transportation Changes:

- **Bike lanes and infrastructure:** A significant increase in protected and dedicated bike lanes is visible throughout the city, reflecting the push for alternative transportation.
- **Transit changes:** While the Central Subway finally opened in this period, other visual changes are less noticeable from street level. However, increasing numbers of electric buses might be seen.
- **Scooter proliferation (and subsequent regulation):** From around 2017-2019, dockless scooters flooded the city, dramatically changing the streetscape, before regulations brought more order.

Neighborhood Transformations:

- **Mid-Market revitalization (and struggles):** The area around Twitter's headquarters has seen noticeable changes, with new businesses and offices, but also persistent challenges with homelessness and public safety.
- **Mission Bay development:** This formerly industrial area has transformed into a bustling neighborhood with new housing, UCSF facilities, and the Chase Center.
- **Gentrification and displacement:** While not always directly visible, the effects of gentrification can manifest in changes to storefronts, the types of businesses present, and the overall aesthetic of a neighborhood.

Other Visible Changes:

- **Increased homelessness:** Sadly, a significant rise in visible homelessness has been a major change in this period, with tent encampments and individuals on the streets becoming more prevalent.
- **Retail vacancies:** Particularly after the pandemic, an increase in empty storefronts, especially in downtown areas, has been a noticeable change.
- **Parklets and outdoor dining:** Accelerated by the pandemic, parklets and outdoor dining spaces became more common, altering the look and feel of many commercial streets.
- **Art installations and murals:** San Francisco's vibrant street art scene continues to evolve, with new murals and installations regularly appearing and disappearing.

To actually see these changes, you would need to compare imagery from the two periods:

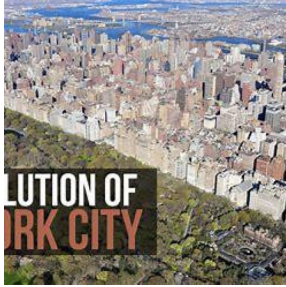
- **Google Street View:** Offers a historical record of street-level views, allowing for direct comparisons over time.
- **Satellite imagery:** Services like Google Earth can show broader changes in building density and land use.
- **Photo archives:** Local historical societies, libraries, and online photo collections can provide visual documentation of past cityscapes.
- **Documentaries and news footage:** Videos can offer a dynamic perspective on how the city has changed over time.

By comparing imagery and exploring these different sources, you can gain a clearer visual understanding of San Francisco's transformation between 2011 and 2023.

Figure 13. Full Gemini 1.5 Predictions Without Images (Accessed time: Nov. 2024).

Prompt: What are the trends of changes happened in **New York City** that we can see visually in 2011-2023?

MLLM (GPT-4o) Answers:



Between 2011 and 2023, New York City underwent significant visual transformations across its skyline, infrastructure, and public spaces. Key trends include:

1. Skyline Evolution:

- *Rise of Super-Tall Skyscrapers:* The city saw the construction of numerous super-tall buildings, particularly along 57th Street, known as “Billionaire’s Row.” Notable additions include 432 Park Avenue and Central Park Tower, which have redefined the city’s silhouette.

2. Infrastructure and Public Space Enhancements:

- *Pedestrianization Initiatives:* Broadway underwent significant changes, with sections transformed into pedestrian plazas to enhance walkability and reduce vehicular traffic. Notably, areas in Times Square and Herald Square were closed to cars, creating more space for pedestrians. *Reopening of Public Spaces:* After 15 years, a 15,000-square-foot greenspace under the Brooklyn Bridge, part of “The Arches” area, reopened in November 2024, providing much-needed greenery and recreational space for the community.

3. Adaptive Reuse of Buildings:

- *Office-to-Residential Conversions:* In response to changing work habits and increased office vacancies, developers have been converting obsolete office buildings into residential apartments, particularly in areas like the Financial District. This trend aims to address housing shortages and revitalize underused spaces.

4. Preservation and Restoration Efforts:

- *Neon Signage Conservation:* Efforts to preserve iconic neon signs have intensified, with restorations of landmarks like Radio City Music Hall and the Essex House hotel, reflecting a desire to maintain the city’s historical visual elements amid modernization.

5. Sustainable Urban Design:

- *Green Infrastructure Projects:* The city has embraced sustainable development, with projects like the transformation of Fifth Avenue into a more pedestrian-friendly boulevard, featuring expanded sidewalks, seating, and greenery, aiming to enhance urban vitality and environmental sustainability.

These developments collectively highlight New York City’s dynamic evolution, balancing modernization with preservation and sustainability.

Figure 14. GPT-4o Predictions Without Images. (Accessed time: Nov. 2024)

Prompt: I have some image sequences, each captured from one location over time. They are sorted in chronological order.

[... Image Sequences Containing 7K-8K Images...]

Now you, as a brilliant analyst, need to tell me what are the trends of temporal changes in these sequences. Describe each trend in a line in the format of “[Before] → [After] (Evidence ID: [IDX],[IDX])”.

Replace [Before] with descriptions for before the trend, replace [After] with descriptions for after the trend, and replace [IDX] with an evidence image sequence index for the trend. For each trend, you need to tell me at least 500 evidence.

MLLM (Gemini 1.5) Answers:

- Construction site with a partially built structure → Completed building.
- One business → Different business
- One business → Vacancy/Lease available
- Older signage/awning → Newer signage/awning No bike lane markings → Bike lane markings added
- One advertisement → Different advertisement
- One business storefront → Different business storefront
- Older storefronts/signage → New storefronts/signage Bare trees/winter scenery → Leafy trees/summer scenery
- Building under scaffolding → Building with no scaffolding
- Brownstone building at a street corner → Brownstone building with an awning and signage
- One storefront/business signage → Different storefront/business signage/closed down
- Smaller plants/Bare trees → Larger plants/Leafy trees/Bare Trees
- Older graffiti → New graffiti/graffiti removal
- Building exterior or area without scaffolding/construction equipment → Building exterior or area with scaffolding/construction equipment

Figure 15. MLLMs Processing All Images.

Prompt: Here are 50K text descriptions, each describing a change happened in the city.

[... Contents of 50K Changes ...]

Now you, as a brilliant analyst, need to tell me what are the trends among these changes. Describe each trend in a line in the format of “[Before] → [After] (Evidence ID: [IDX],[IDX])”.

Replace [Before] with descriptions for before the trend, replace [After] with descriptions for after the trend, and replace [IDX] with an evidence change index for the trend. For each trend, you need to tell me at least 500 evidence.

MLLM (Gemini 1.5) Answers:

- There is no scaffolding on the building → There is scaffolding on the building.
- The scaffolding is present. → The scaffolding is gone.
- The crosswalk markings are solid white lines → The crosswalk markings are thick white stripes.
- The building is covered in scaffolding → The scaffolding is removed.
- The building has a [Business_name] sign → The [Business_name] sign is replaced.
- The storefront is occupied by a business → The storefront is vacant.
- The storefront is occupied by one business → The storefront is occupied by another business.
- The trees on the street is thin → The trees grow taller and thicker.
- Scaffolding is erected in front of the building → The scaffolding is removed, revealing the building’s facade.
- The billboard has an advertisement → The advertisement changed.
- A business with [Business_name] sign → The [Business_name] sign is gone.
- There were no bike lanes → There is a new bike lane.
- Entrance to the building has a [Color] door → The entrance to the building has a [Color] door.
- Green scaffolding covers the building → The scaffolding is gone.

Figure 16. **MLLMs Processing All Changes.**

Prompt: Below are pictures of the same location taken over time.

This is image No. 2, taken in June 2021.



This is image No. 3, taken in May 2022.



You, as a brilliant role, need to describe what have changed over time in these images in details and also tell me from what images you found the changes. Please format the answer strictly in the following format, one change a line:

Start: [Start_DESCRIPTION] → End: [End_DESCRIPTION] (happened after image No.[X]).

Format Instructions:

Replace [Start_DESCRIPTION] with detailed descriptions of the start of the change;

Replace [End_DESCRIPTION] with detailed descriptions of the end of the change;

Replace [X] with the index of the image where the change happened after. That said, this number should come from the "image No.[X]" descriptions above.

There are a few guidelines that you should follow when answering the question:

- DO NOT HALLUCINATE. You should stick to the changes that can be clearly and certainly seen in the images.
- IGNORE CHANGES IN THE PERIPHERAL. Because the viewpoint may wiggle, ignore the changes that happened very close the the peripheral of images because they usually can not be confirmed.
- IGNORE TINY UNRECOGNIZABLE TEXT. If certain text in the image is too small to be clearly recognized, such text content should not be part of a change because it can not be confirmed.
- IGNORE CHANGES CAUSED BY OCCCLUSIONS. There will be occlusions, like trucks or cars parked in front of some subjects in images. If a subject appear or disappear due to such occlusions, it is not a change.
- TRACK THE SAME SUBJECT. When detect changes, please be sure to describe the change happen to the same subject.
- DO NOT MISTAKE FOREGROUND AND BACKGROUND. Do not mistake the foreground subject, such a street sign, with attachments to the background subject, like signs on the building behind the street sign.
- IGNORE CHANGES CAUSED BY ILLUMINATION CHANGES. Ignore subtle color changes caused by illumination changes, like colors become darker because of shadows, you can get the clue of illumination from the time of day, i.e. the time after "taken at " in the prompt.
- ONE EVENT A CHANGE. For each detected change, it should focus on one event only, not multiple events. For example, the change "Start: The storefront of the building has a black awning and a red truck parked in front of it. → End: The storefront of the building has a red awning now and the truck is gone" should be 2 changes one on the awning and the other on the truck.
- CHANGES SHOULD BE SIGNIFICANT. This means the description after "Start:" should describe a clearly different status of THE SAME SUBJECT from the description after "End:". For example, the change "Start: The building is free of scaffolding. → End: There is no scaffolding on the building." is not a significant change because both Start and End are the same status of the same subject.

Now, please tell me changes that you can find in the images. Make sure you tell me ALL THE CORRECT CHANGES, following the above guidelines.

MLLM (Gemini 1.5) Answers:

...

Start: The support beams of the overpass were in pale green color. → End: The support beams were painted into a bright blue color. (Happened after image No. 2)

...

Figure 17. Local Change Detection Prompt.

Prompt:

I have a text description of a change that happened in a city. For example: Before: The Ginger Man's Bar did not have any "Bar for Lease" sign. After: The Ginger Man's Bar put up a "Bar for Lease" sign on the storefront window. As you can see, in the description, there will be the status before the change and the status after the change. Your job, as a brilliant analyst, is to derive more abstract versions of the same change by removing details.

Formulate the answer strictly in the following list format:

There are [N] levels of details on where the change happened:

p[X]. [Description of one level]

Meanwhile, there are [M] levels of details on the change itself:

c[X]. [Description of one level]

[More descriptions of the reasoning].

Answer: (reason) [A derived abstract version].

Here are a few examples:

... [Examples] ...

Here is the change you need to derive abstract versions for:

Before: The single door to the right of the main entrance has a sign above it that reads "151".

After: The sign above the single door is blank.

Now, please tell me the abstract versions of this change, following the above instructions.

MLLM (Gemini 1.5) Answers:

Derivation:

There are 3 levels of details on where the change happened:

- p1. The single door to the right of the main entrance. (reason: the original very specific place).
- p2. The single door. (reason: a more general location description without mentioning it's to the right of the main entrance).
- p3. A door. (reason: the most general location description). Meanwhile there are 2 levels of details on the change itself:
- c1. The sign above it changing from reading "151" to being blank. (reason: the original very specific change.)
- c2. The sign above it changing from reading "151" to something else. (reason: more general without specifying the sign being blank.)
- c3. The sign above it changing. (reason: the most general change, without even specifying how the sign changed).

Answer:

- (p1 + c1) The sign above the single door to the right of the main entrance changed from reading "151" to being blank.
- (p1 + c2) The sign above the single door to the right of the main entrance changed from reading "151" to something else.
- (p1 + c3) The sign above the single door to the right of the main entrance changed.
- (p2 + c1) The sign above the single door changed from reading "151" to being blank.
- (p2 + c2) The sign above the single door changed from reading "151" to something else.
- (p2 + c3) The sign above the single door changed.
- (p3 + c1) The sign above the door changed from reading "151" to being blank.
- (p3 + c2) The sign above the door changed from reading "151" to something else.
- (p3 + c3) The sign above the door changed.

Figure 18. Trend Proposal Prompt.

Prompt:

I have a text description of a change that happened in a city. For example: Before: The Ginger Man's Bar did not have any "Bar for Lease" sign.

After : The Ginger Man's Bar put up a "Bar for Lease" sign on the storefront window.

As you can see, in the description, there will be the status before the change and the status after the change.

I also have a test description of a group of similar changes that happened. For example:

A storefront put up a "For Lease" sign on its window.

Your job, as a brilliant analyst, is to decide if the given change belongs to the given group, "Y" for yes and "N" for no, and why.

Formulate the answer strictly in the following format:

Answer: [Y/N].

Reason: [Description of the reasoning].

For the example above, the answer would be:

Answer: Y.

Reason: The given change describes a bar putting up a "Bar for Lease" sign. We can remove the details of the location being a specific "Ginger Man's Bar" but only keep it as a storefront. Meanwhile, we can keep the event of putting up the sign the same mostly intact, only removing the detail of a bar, i.e. "Bar for Lease" to "For Lease", to align with a storefront. Then, we can arrive at the group description. Therefore, the given change belongs to the given group.

Here are a few more examples:

... [Examples] ...

Here is the change for you:

Before: The storefront on the corner is a GNC store.

After: The storefront is an OPT X store.

Here is the group for you:

A juice shop opened at the storefront.

MLLM (Gemini 1.5) Answers:

Answer: N.

Reason: The group description specifies a change to a "juice shop". The provided change describes a change from a "GNC store" to an "OPT X store". While both describe storefront changes, the specific type of store in the group description ("juice shop") is not reflected in the change description. We cannot generalize "OPT X store" to "juice shop" without adding information.

Figure 19. Trend Verification Prompt.