Measuring and Mitigating Hallucinations in Vision-Language Dataset Generation for Remote Sensing

Madeline Anderson[1], **Miriam Cha**[2], **William T. Freeman**[1], **J. Taylor Perron**[1], **Nathaniel Maidel**[3], **Kerri Cahoy**[1]

[1]MIT, [2]MIT Lincoln Laboratory, [3]DAF MIT AI Accelerator

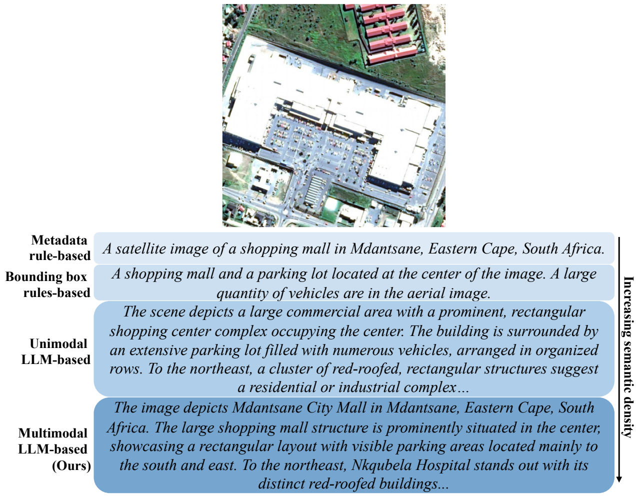
Abstract

Vision language models have achieved impressive results across various fields. However, adoption in remote sensing remains limited, largely due to the scarcity of paired image-text data. To bridge this gap, synthetic caption generation has gained interest, traditionally relying on rule-based methods that use metadata or bounding boxes. While these approaches provide some description, they often lack the depth needed to capture complex wide-area scenes. Large language models (LLMs) offer a promising alternative for generating more descriptive captions, yet they can produce generic outputs and are prone to hallucination. In this paper, we propose a new method to enhance vision-language datasets for remote sensing by integrating maps as external data sources, enabling the generation of detailed, context-rich captions. Additionally, we present methods to measure and mitigate hallucinations in LLM-generated text. We introduce fMoW-mm, a multimodal dataset incorporating satellite imagery, maps, metadata, and text annotations. We demonstrate its effectiveness for automatic target recognition in few-shot settings, achieving superior performance compared to other vision-language remote sensing datasets.

Introduction

In recent years, there have been significant advancements in vision-language models, leading to powerful applications across many fields (Zhang et al. 2024a; Long et al. 2022; Du et al. 2022). However, adoption within the remote sensing community has lagged, largely due to the limited availability of paired data for remote sensing imagery and text. Recently, researchers have started to address this gap by generating synthetic captions for remote sensing images (Khanna et al. 2024; Liu et al. 2024; Zhang et al. 2024b). Traditionally, rule-based methods leveraging metadata (Khanna et al. 2024) and bounding boxes (Liu et al. 2024) have been used, but these approaches fall short when it comes to fully describing the complexity of wide-area remote sensing scenes.

The adoption of large language models (LLMs) offers a promising alternative, as LLMs can potentially generate more descriptive and contextually rich captions (Zhang et al. 2024b). Yet, LLM-generated text for remote sensing data often remains generic, and importantly, is prone to hallucination. This issue of hallucination has yet to be thoroughly explored in the context of vision-language dataset for remote sensing, where accurate and detailed scene descriptions are important for data curation.



A satellite image of a shopping mall in Mdantsane, Eastern Cape, South Africa.

A shopping mall and a parking lot located at the center of the image. A large quantity of vehicles are in the aerial image.

The scene depicts a large commercial area with a prominent, rectangular shopping center complex occupying the center. The building is surrounded by an extensive parking lot filled with numerous vehicles, arranged in organized rows. To the northeast, a cluster of red-roofed, rectangular structures suggest a residential or industrial complex...

The image depicts Mdantsane City Mall in Mdantsane, Eastern Cape, South Africa. The large shopping mall structure is prominently situated in the center, showcasing a rectangular layout with visible parking areas located mainly to the south and east. To the northeast, Nkqubela Hospital stands out with its distinct red-roofed buildings...

Figure 1: **Comparison of captioning methods:** Rule-based captions are limited in detail. Unimodal LLM captions are fluid but often generic. Wide-area scenes covering diverse structures and objects require semantically rich descriptions. We leverage the semantic density of maps to generate comprehensive and detailed captions.

In this paper, we propose a new approach for curating vision-language datasets for remote sensing by integrating external sources of information, such as maps and metadata. Maps offer a rich source of contextual information, including labels and segmentation maps. Using these external sources, we introduce a method to generate more comprehensive and detailed captions for remote sensing images than existing methods allow (Figure 1).

To address the issue of hallucinations, we present methods to measure and mitigate hallucination in LLM-generated captions. Additionally, we introduce **fMoW-mm**, a new multimodal dataset (built upon fMoW (Christie et al. 2018)), which includes satellite imagery, maps, metadata, and text annotations. Finally, we demonstrate the effectiveness of

Dataset Name	Captioning Method	Caption Quality	External Info
RSVQA (Lobry et al. 2020)	Map data rules-based	Focuses on reasoning	Maps (OSM)
Skyscript (Wang et al. 2023)	Map data rules-based	Rigid and limited to OSM tags	Maps (OSM)
DiffusionSAT (Khanna et al. 2024)	Metadata rules-based	Rigid with details limited to metadata	Metadata
RemoteCLIP (Liu et al. 2024)	Bounding box rules-based	Rigid with details limited to bounding box objects	None
ChatEarthNet (Yuan et al. 2024)	LLM-based	Fluid sounding with details limited to landcover	Landcover (WorldCover)
RS5M (Zhang et al. 2024b)	Web-filtered, LLM-based	Coarse detail with focus on central objects	None
fMoW-mm (Ours)	**Multimodal LLM-based**	**Fluid sounding with comprehensive, specific details**	**Metadata + maps (OSM)**

Table 1: Overview of various vision-language remote sensing datasets.

fMoW-mm in automatic target recognition under few-shot conditions, showcasing the potential of this dataset for enhancing remote sensing applications with limited labeled data. Our contributions are as follows:

- We introduce a novel dataset curation method that leverages external data sources, specifically maps, for enhanced language descriptions of remote sensing images.
- We present fMoW-mm, a comprehensive multimodal dataset cross-referenced with fMoW, consisting of satellite imagery, map, metadata, and text annotations.
- We explore methods to measure and mitigate hallucinations in LLM-generated captions for remote sensing.
- We demonstrate the utility of fMoW-mm for automatic target recognition in limited-label scenarios.

Related Work

Vision-Language Datasets for Remote Sensing

Although large vision-language datasets are less common in the remote sensing domain, several have been developed in recent years. We review six existing datasets in Table 1. Datasets that rely on rules-based captions, such as RSVQA (Lobry et al. 2020), Skyscript (Wang et al. 2023), DiffusionSAT (Khanna et al. 2024), and RemoteCLIP (Liu et al. 2024), often produce rigid captions with limited detail. The content is constrained by the external information fed into the rules-based frameworks, such as OpenStreetMap (OSM) data for RSVQA and Skyscript, metadata for DiffusionSAT, and bounding boxes for RemoteCLIP. ChatEarthNet (Yuan et al. 2024) and RS5M (Zhang et al. 2024b) leverage LLMs, resulting in more fluid sounding captions. However, ChatEarthNet captions primarily describe landcover, while RS5M captions, generated using BLIP-2 (Li et al. 2023), contain coarse details. RS5M also includes internet-scraped image-text data, often centered on a single object, which may not represent typical remote sensing images. We address these shortcomings by leveraging a multimodal LLM (GPT-4o) with multiple sources of external information (maps and metadata) to generate comprehensive, detailed, and fluid captions for complex remote sensing scenes.

Hallucination Metrics and Mitigation Strategies

Measuring and mitigating hallucinations in LLM-generated captions is critical. Existing hallucination metrics include statistical, model-based, and vision-language measures (Ji et al. 2023). Statistical metrics like ROUGE (Lin 2004), BLEU (Papineni et al. 2002), and PARENT (Dhingra et al. 2019) assess hallucinations based on n-gram overlaps. Model-based metrics include Information Extraction (Singh 2018), QA-based methods (Deutsch, Bedrax-Weiss, and Roth 2021), Natural Language Inference (Dušek and Kasner 2020), and Faithfulness Classification (Liu et al. 2022), and LM-based approaches (Filippova 2020). However, many rely on task-specific datasets or LLM access, which may not be available. Metrics specific to vision-language hallucinations are very scarce (Rohrbach et al. 2019). To address the lack of suitable metrics, we propose a statistical metric inspired by BLEU precision that uses OSM tags as source text to measure hallucination rates.

LLM hallucination mitigation strategies include data-based, modeling-based, and post-processing methods (Ji et al. 2023). Data-based strategies include caption ranking or filtering and information augmentation with synthetic or external data. Modeling techniques include planning and sketching (Wang et al. 2021), reinforcement learning (Uc-Cetina et al. 2022), multi-task learning (Weng et al. 2020), and controllable generation (Rashkin et al. 2021; Wu et al. 2021). Post-processing focuses on correcting hallucinations after captions are generated. Without direct LLM access required by many modeling methods, we mitigate hallucinations using data-based strategies (external data augmentation) and post-processing techniques (prompt ensembling).

Multimodal Dataset Curation

Figure 2 outlines the fMoW-mm curation process: 1) Gather satellite images and metadata from fMoW-rgb, 2) Use bounding box metadata to perform an OSM Mapbox query and retrieve map tiles, 3) Input satellite images, maps, and metadata into GPT-4o to generate captions, 4) Combine these elements to create fMoW-mm. Each step is detailed in the following subsections.

Functional Map of the World (fMoW-rgb)

The fMoW-rgb dataset consists of 83,412 remote sensing images that feature objects in 63 categories (Christie et al. 2018). Each image comes with corresponding metadata such as category label, latitude, longitude, timestamp, ground sampling distance (GSD), and bounding box.

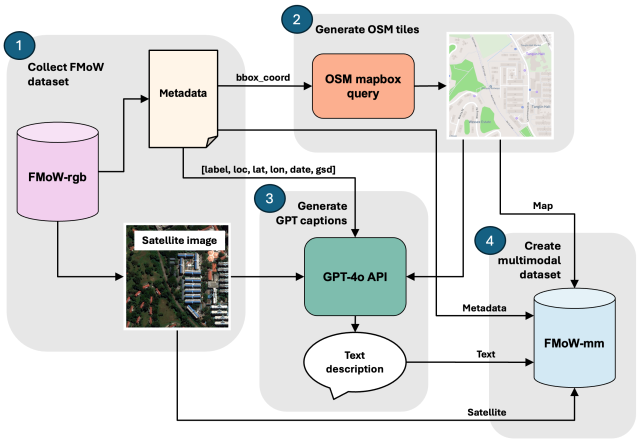


Figure 2: fMoW-mm data curation pipeline

OpenStreetMap (OSM) Tile Retrieval

We use the bounding box coordinates from the fMoW-rgb metadata to query the corresponding OSM Static Image tiles through the Mapbox API. Map styles are customized using the online Mapbox studio.

Caption Generation with GPT-4o

To generate captions, we use the GPT-4o API from OpenAI, which accepts visual and text inputs. For each sample, we input the fMoW-rgb satellite image, metadata and OSM tile. The input metadata includes the category label, location (city, state/region, country), latitude, longitude, and GSD. We prompt GPT-4o to describe the remote sensing scene and to include landmarks, relative positions, sizes, colors, and quantities, while leveraging the metadata and map for context. Other LLMs, including open-source options, can be substituted for GPT-4o, as long as they accept visual inputs.

Multimodal Functional Map of the World (fMoW-mm)

We combine the fMoW-rgb satellite image and metadata, the OSM tile, and the GPT-4o generated caption to create 83,412 tuples of {satellite, metadata, map, text}. Figure 3 shows a sample from the fMoW-mm dataset. The full dataset is available at `https://bit.ly/fMoW-mm`.

Hallucination Metric

Hallucinations often occur when the LLM infers incorrect landmarks during caption generation. To quantify these hallucinations, we compute the false discovery rate (FDR), inspired by BLEU precision, which measures the proportion of false positives in the generated text. Unlike BLEU, which evaluates n-gram overlaps, we calculate precision over variable-length proper nouns and define FDR as $1 - precision$:

$$FDR = 1 - \frac{\sum_{c \in C} \mathbb{1}_R(c)}{K} \tag{1}$$

where the candidate list $C = [c_1, c_2, ..., c_K]$ is an array of K proper nouns, and the reference list $R = [r_1, r_2, ..., r_M]$ is an array of M proper nouns. The indicator function $\mathbb{1}_R(c) = 1$ if $c \in R$ and 0 otherwise. FDR reflects the proportion of false positives among all predicted positives, quantifying the rate of hallucinations in the generated (candidate) captions.

Experiments

We perform ablations to evaluate how components of our curation pipeline affect hallucination rates (FDR) and measure the percentage of uncertain words as a proxy for LLM uncertainty. We then demonstrate fMoW-mm's effectiveness in enhancing few-shot object detection performance.

Ablations

- **Map Resolution:** We vary the resolution of the OSM input to GPT-4o, considering {256, 512, 1024}.
- **Map Types:** We explore four map variations:
  - All Labels: Includes all available labels on the map.
  - Landmarks-Only: Includes only landmark labels, excluding street names.
  - Streets-Only: Includes only street names, excluding landmark labels.
  - No Labels: Displays the segmentation map without any text labels.
- **Prompt Ensembling:** We generate multiple prompts for the same question and aggregate the responses to analyze convergence. We experiment with {1, 3, 5} prompts.

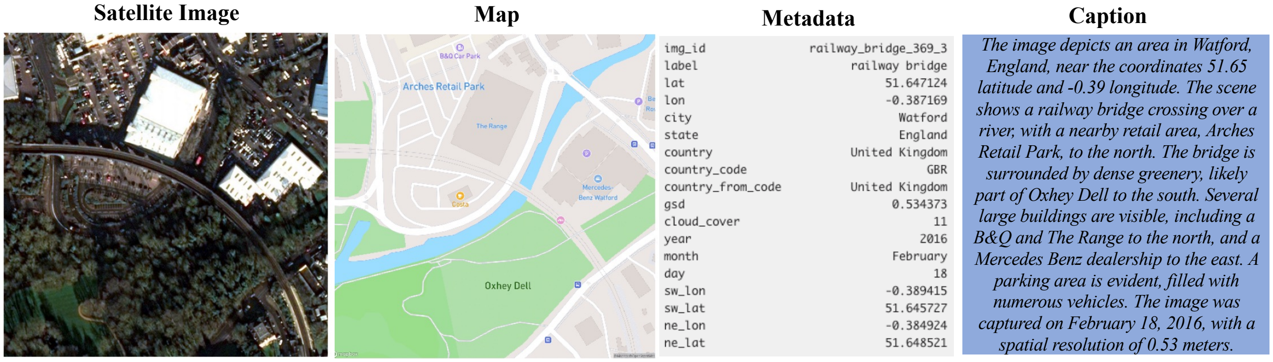
Figure 4 shows that increasing map resolution reduces hallucination rates and uncertain word percentages, highlighting the importance of map legibility. We use a 1024 × 1024 resolution for fMoW-mm. While further increases in resolution may offer additional benefits, we leave this exploration for future work due to computational constraints.

Adding text labels, such as landmarks and street names, predictably increases hallucination rates. The inclusion of street names (e.g., streets-only, all-labels) results in a more pronounced increase, likely because non-horizontally aligned street names introduce ambiguity that leads to hallucinations. Landmark names, which are consistently horizontal, cause fewer issues. Captions generated without labels (i.e., no-label) achieve the lowest hallucination rates but are often overly generic, with a high rate of uncertain word usage. For the fMoW-mm dataset, we selected the landmarks-only configuration as it strikes a good balance, minimizing hallucinations while maintaining reasonable specificity.

Prompt ensembling did not result in noticeable improvements. We suspect that repeated hallucinations across responses may increase overlap, propagating errors into the final captions. For fMoW-mm, we aggregate responses from three prompts, yielding the lowest FDR.

Few-Shot Object Detection with CLIP

We continually pretrain the CLIP (Radford et al. 2021) ViT-L/14 model using the fMoW-mm dataset and evaluate the learned visual representation on few-shot object detection (Bou et al. 2024). The model was continually trained for



img_id	railway_bridge_369_3
label	railway bridge
lat	51.647124
lon	-0.387169
city	Watford
state	England
country	United Kingdom
country_code	GBR
country_from_code	United Kingdom
gsd	0.534373
cloud_cover	11
year	2016
month	February
day	18
sw_lon	-0.389415
sw_lat	51.645727
ne_lon	-0.384924
ne_lat	51.648521

The image depicts an area in Watford, England, near the coordinates 51.65 latitude and -0.39 longitude. The scene shows a railway bridge crossing over a river, with a nearby retail area, Arches Retail Park, to the north. The bridge is surrounded by dense greenery, likely part of Oxhey Dell to the south. Several large buildings are visible, including a B&Q and The Range to the north, and a Mercedes Benz dealership to the east. A parking area is evident, filled with numerous vehicles. The image was captured on February 18, 2016, with a spatial resolution of 0.53 meters.

Figure 3: A sample from the fMoW-mm dataset. The generated caption accurately incorporates information from the satellite image, map, and metadata.

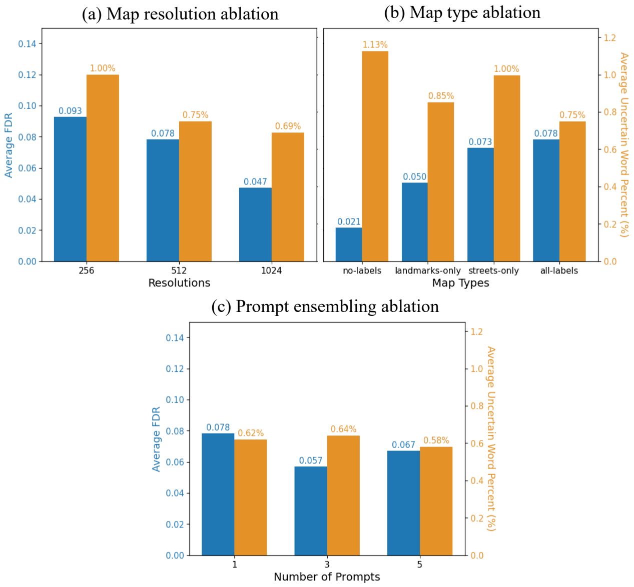


Figure 4: **Ablations. (a) Map Resolution:** Higher resolution reduces hallucination rates and uncertainty in generated captions. **(b) Map Types:** Using landmarks-only gives the best balance, reducing hallucinations while limiting uncertainty. **(c) Prompt Ensembling:** Combining captions from multiple prompts did not significantly impact the metrics, however increasing from 3 to 5 prompts may result in repeated hallucinations that propagate into the final caption.

50 epochs with a batch size of 125. We compare performance with vision-language baselines: CLIP, OpenCLIP, GeoRSCLIP, and RemoteCLIP.

Table 2 shows the mAP50 scores for 5, 10 and 30-shot detection on the DIOR dataset (Li et al. 2020), averaged over 5 splits. Our model demonstrates improved performance across all n-shots, showing its viability for data-scarce scenarios. Although the fMoW-mm dataset is much smaller than the datasets used for GeoRSCLIP (RS5M, ~5M) and RemoteCLIP (~150k), it achieves superior performance, highlighting the benefits of increased semantic density in the generated captions. To isolate the impact of the dataset, comparisons are limited to CLIP models.

Backbone	5-shot	10-shot	30-shot
CLIP (Radford et al. 2021)	0.1447	0.1872	0.1810
OpenCLIP (Cherti et al. 2023)	0.1477	0.1863	0.1804
GeoRSCLIP (Zhang et al. 2024b)	0.1401	0.1791	0.1815
RemoteCLIP (Liu et al. 2024)	0.1571	0.1893	0.1903
Ours	**0.1574**	**0.1902**	**0.1972**

Table 2: mAP50 scores for 5, 10, and 30-shot object detection on the DIOR dataset using various visual backbones with ViT-L/14, averaged across 5 splits.

Conclusion

In this work, we explored methods to measure and mitigate hallucinations in captions describing remote sensing imagery. Previous approaches to caption generation have often resulted in rigid and generic descriptions. Our approach enhances vision-language datasets in remote sensing by integrating maps as external data sources, enabling the creation of more detailed and contextually rich captions. Through the introduction of fMoW-mm—a multimodal dataset extending the fMoW dataset with satellite imagery, maps, metadata, and text annotations—we demonstrate a reduced rate of hallucinations and improved performance in automatic target recognition under few-shot conditions.

Acknowledgments

Research was sponsored by the Department of the Air Force Artificial Intelligence Accelerator and was accomplished under Cooperative Agreement Number FA8750-19-2-1000. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Department of the Air Force or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

Bou, X.; Facciolo, G.; von Gioi, R. G.; Morel, J.-M.; and Ehret, T. 2024. Exploring Robust Features for Few-Shot Object Detection in Satellite Imagery. arXiv:2403.05381.

Cherti, M.; Beaumont, R.; Wightman, R.; Wortsman, M.; Ilharco, G.; Gordon, C.; Schuhmann, C.; Schmidt, L.; and Jitsev, J. 2023. Reproducible Scaling Laws for Contrastive Language-Image Learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2818–2829. IEEE.

Christie, G.; Fendley, N.; Wilson, J.; and Mukherjee, R. 2018. Functional Map of the World. arXiv:1711.07846.

Deutsch, D.; Bedrax-Weiss, T.; and Roth, D. 2021. Towards Question-Answering as an Automatic Metric for Evaluating the Content Quality of a Summary. arXiv:2010.00490.

Dhingra, B.; Faruqui, M.; Parikh, A.; Chang, M.-W.; Das, D.; and Cohen, W. W. 2019. Handling Divergent Reference Texts when Evaluating Table-to-Text Generation. arXiv:1906.01081.

Du, Y.; Liu, Z.; Li, J.; and Zhao, W. X. 2022. A Survey of Vision-Language Pre-Trained Models. arXiv:2202.10936.

Dušek, O.; and Kasner, Z. 2020. Evaluating Semantic Accuracy of Data-to-Text Generation with Natural Language Inference. In Davis, B.; Graham, Y.; Kelleher, J.; and Sripada, Y., eds., *Proceedings of the 13th International Conference on Natural Language Generation*.

Filippova, K. 2020. Controlled Hallucinations: Learning to Generate Faithfully from Noisy Data. In Cohn, T.; He, Y.; and Liu, Y., eds., *Findings of the Association for Computational Linguistics: EMNLP 2020*.

Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y. J.; Madotto, A.; and Fung, P. 2023. Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55(12): 1–38.

Khanna, S.; Liu, P.; Zhou, L.; Meng, C.; Rombach, R.; Burke, M.; Lobell, D.; and Ermon, S. 2024. DiffusionSat: A Generative Foundation Model for Satellite Imagery. arXiv:2312.03606.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. arXiv:2301.12597.

Li, K.; Wan, G.; Cheng, G.; Meng, L.; and Han, J. 2020. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS Journal of Photogrammetry and Remote Sensing*, 159.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*.

Liu, F.; Chen, D.; Guan, Z.; Zhou, X.; Zhu, J.; Ye, Q.; Fu, L.; and Zhou, J. 2024. RemoteCLIP: A Vision Language Foundation Model for Remote Sensing. arXiv:2306.11029.

Liu, T.; Zhang, Y.; Brockett, C.; Mao, Y.; Sui, Z.; Chen, W.; and Dolan, B. 2022. A Token-level Reference-free Hallucination Detection Benchmark for Free-form Text Generation. arXiv:2104.08704.

Lobry, S.; Marcos, D.; Murray, J.; and Tuia, D. 2020. RSVQA: Visual Question Answering for Remote Sensing Data. *IEEE Transactions on Geoscience and Remote Sensing*.

Long, S.; Cao, F.; Han, S. C.; and Yang, H. 2022. Vision-and-Language Pretrained Models: A Survey. arXiv:2204.07356.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020.

Rashkin, H.; Reitter, D.; Tomar, G. S.; and Das, D. 2021. Increasing Faithfulness in Knowledge-Grounded Dialogue with Controllable Features. arXiv:2107.06963.

Rohrbach, A.; Hendricks, L. A.; Burns, K.; Darrell, T.; and Saenko, K. 2019. Object Hallucination in Image Captioning. arXiv:1809.02156.

Singh, S. 2018. Natural Language Processing for Information Extraction. arXiv:1807.02383.

Uc-Cetina, V.; Navarro-Guerrero, N.; Martin-Gonzalez, A.; Weber, C.; and Wermter, S. 2022. Survey on reinforcement learning for language processing. *Artificial Intelligence Review*, 56(2): 1543–1575.

Wang, P.; Lin, J.; Yang, A.; Zhou, C.; Zhang, Y.; Zhou, J.; and Yang, H. 2021. Sketch and Refine: Towards Faithful and Informative Table-to-Text Generation. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*.

Wang, Z.; Prabha, R.; Huang, T.; Wu, J.; and Rajagopal, R. 2023. SkyScript: A Large and Semantically Diverse Vision-Language Dataset for Remote Sensing. arXiv:2312.12856.

Weng, R.; Yu, H.; Wei, X.; and Luo, W. 2020. Towards Enhancing Faithfulness for Neural Machine Translation. In Webber, B.; Cohn, T.; He, Y.; and Liu, Y., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2675–2684. Online: Association for Computational Linguistics.

Wu, Z.; Galley, M.; Brockett, C.; Zhang, Y.; Gao, X.; Quirk, C.; Koncel-Kedziorski, R.; Gao, J.; Hajishirzi, H.; Ostendorf, M.; and Dolan, B. 2021. A Controllable Model of Grounded Response Generation. arXiv:2005.00613.

Yuan, Z.; Xiong, Z.; Mou, L.; and Zhu, X. X. 2024. ChatEarthNet: A Global-Scale Image-Text Dataset Empowering Vision-Language Geo-Foundation Models. arXiv:2402.11325.

Zhang, J.; Huang, J.; Jin, S.; and Lu, S. 2024a. Vision-Language Models for Vision Tasks: A Survey. arXiv:2304.00685.

Zhang, Z.; Zhao, T.; Guo, Y.; and Yin, J. 2024b. RS5M and GeoRSCLIP: A Large-Scale Vision- Language Dataset and a Large Vision-Language Model for Remote Sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–23.