Response to Comments of Referee #2

Thank you for the instructive and constructive comments for our paper. Those comments are very helpful for and serve as significant guidance for our research. We have studied the comments carefully and revised our manuscript accordingly. The changes in our manuscript are highlighted in **red**. The point-to-point responses to your questions/comments are listed as follows.

Comments to the Author:

This manuscript introduces the BRIGHT dataset, which is the first open building damage assessment dataset with global coverage, multi-hazard scenarios, multimodal imagery (Optical and SAR), and sub-meter resolution. The paper systematically describes data collection, annotation, and quality control methods, and validates the dataset with multiple deep learning models, including cross-disaster transfer (zero-shot and one-shot), semi-supervised, and unsupervised approaches. The dataset demonstrates clear novelty and practical value, and it is of significant importance for advancing research and applications in disaster emergency response, remote sensing, and artificial intelligence. Generally, the paper is well structured, logically clear, with detailed results and strong value in terms of data sharing. Although the manuscript is rich in content, there are still details that require improvement, and I recommend appropriate revisions.

Response: We really appreciate your spot-on summary of our manuscript and such a positive endorsement of our work. Our responses to your valuable comments and suggestions are itemized below.

Q1: The explanation of annotation consistency and reliability remains insufficient. Although the authors state that the data annotations were obtained from multiple institutions such as Copernicus EMS, UNOSAT, and FEMA and then refined manually, there may be inconsistencies in how different institutions define “damaged” and “destroyed”. This could affect the consistency of annotations across disaster scenarios. It is therefore necessary to further elaborate on the process of unifying annotations, provide more detail on the manual refinement procedures.

R1: We sincerely thank the reviewer for raising this crucial point. Ensuring annotation consistency across different data sources and disaster types is paramount for the reliability of BRIGHT as a benchmark dataset, and we appreciate the opportunity to elaborate on our rigorous unification and refinement process.

Our approach was a multi-stage process designed specifically to address the potential inconsistencies the reviewer has identified:

First, we recognized that the source agencies, while conceptually aligned, use slightly different grading scales. To address this, we **established a single, standardized three-tier classification scheme for all events** in BRIGHT: Intact (1), Damaged (2), and Destroyed (3), with clear definitions provided in Table 3 of our manuscript. This scheme served as the universal target for all incoming annotations.

Secondly, the reviewer correctly notes that the exact terminology and number of damage tiers can differ between agencies. However, **their underlying definitions for EO-based damage assessment are conceptually consistent**. All agencies grade damage based on visually verifiable structural failure. This conceptual alignment provided a solid foundation for our initial, rule-based mapping. The "Destroyed" category was the most consistent. Labels such as "Destroyed", "Collapsed", or "Completely Damaged" from all sources were directly mapped to our Destroyed (3) class. For partial damage, we aggregated multiple intermediate tiers. Labels like "Severe Damage", "Major Damage", "Highly Damaged", or "Moderately Damaged" were all mapped to our single Damaged (2) class. This conservative aggregation ensures that our "Damaged" category represents significant, visually verifiable structural harm.

Recognizing that subtle inconsistencies could persist even after the rule-based mapping, the most critical stage of our process was a comprehensive manual review and refinement. This final, expert-led stage served as the ultimate guarantor of consistency, ensuring that every annotation conforms to our unified standard. This procedure, conducted using tools like Google Earth Pro, involved:

- Correction of Inconsistencies: Our experts meticulously compared pre- and post-disaster VHR optical imagery for each annotation point to identify and correct discrepancies between the source label and the visual evidence.
- Harmonization of Ambiguous Labels: We paid special attention to **ambiguous source labels**, such as "Possibly Damaged". In these cases, if clear structural damage was not evident upon visual inspection, we adopted a conservative approach and re-classified the building as "Intact" to ensure a high confidence "Damaged" class.
- Disaggregation of Area-Based Labels: Crucially, we **identified and re-processed all area-based damage annotations** (i.e., where an entire block or neighborhood was assigned a single damage category). Our team manually disaggregated these coarse labels, assigning a precise, building-wise (point-level) damage label to each individual structure within the area. This step was vital for ensuring instance-level consistency and granularity across the entire dataset.

Through this systematic process of standardization, mapping, and exhaustive expert-led refinement, we have made every effort to harmonize the annotations and ensure that the final labels in BRIGHT are as consistent and reliable as possible. We have now added these details to the manuscript to make our process more transparent.

Here, we show the revised part below for your convenience.

The labels in BRIGHT consist of two components: building polygons and post-disaster building damage attributes. Expert annotators manually labeled the building polygons, then all labels underwent independent visual inspections of EO experts to ensure accuracy. Damage annotations were obtained from Copernicus Emergency Management Service[6], the United Nations Satellite Centre (UNOSAT) Emergency Mapping Products[7], and the Federal Emergency Management Agency (FEMA)[8]. These annotations were derived through visual interpretation of high-resolution optical imagery captured before and after the disasters by EO experts, supplemented by partial field visits. To harmonize these diverse annotations and ensure consistency across all 14 disaster events, we implemented a rigorous, multi-stage process. First, we established a single, standardized three-tier classification scheme, including Intact (with pixel value 1), Damaged (with pixel value 2), and Destroyed (with pixel value 3), with clear definitions provided in Table 3, drawing on the frameworks of FEMA's Damage Assessment Operations Manual, EMS-98, the BDD dataset (Adriano et al., 2021), and the xBD dataset (Gupta et al., 2019). While the source agencies' terminology can differ (e.g., "Severe Damage" vs. "Major Damage"), their underlying definitions for EO-based assessment are conceptually consistent. We leveraged this alignment for an initial rule-based mapping, where various intermediate damage tiers were conservatively aggregated into our single "Damaged" category. Second, our team of EO experts conducted a comprehensive manual verification and refinement of every annotation using multi-temporal VHR imagery on platforms like Google Earth Pro. This final stage served as the ultimate guarantor of consistency. We paid special attention to ambiguous source labels, such as "Possibly Damaged". Adopting a conservative approach, these were re-classified as "Intact" if clear structural damage was not evident, thereby ensuring a high-confidence "Damaged" class. We also manually disaggregated all area-based annotations (i.e., where an entire block was assigned a single category). We re-processed these to assign a precise, building-wise damage label to each individual structure, ensuring instance-level consistency and granularity across the entire dataset.

[6]https://emergency.copernicus.eu
[7]https://unosat.org/products
[8]https://www.fema.gov

Q2: The treatment of class imbalance is not sufficient. Figure 5(d) shows that intact buildings account for over 80%, while destroyed buildings account for less than 7%. This severe imbalance directly affects the accuracy of recognizing destroyed classes. Although the authors employed the Lovasz loss function to partially alleviate the issue, this is still not enough to solve the problem. Is this imbalance one of the reasons for the relatively low performance of the subsequent experimental results?

R2: We thank the reviewer for this insightful question. The reviewer has astutely identified one of the most significant and inherent challenges in the task of automated building damage assessment.

First, we would like to clarify that this severe class imbalance is not a unique artifact of our dataset but rather **a common characteristic of real-world post-disaster data**. Disasters, even when severe, typically damage or destroy a minority of the buildings in an affected area. For instance, the widely used **xBD dataset** exhibits a similar long-tail distribution, with the "intact" class constituting the vast majority of labeled buildings **(approximately 75%)**. The imbalance in BRIGHT, as shown in Figure 5(d), therefore realistically reflects the sparse nature of catastrophic damage.

Second, we agree with the reviewer that this imbalance is indeed one of the reasons for the lower performance observed in our benchmark results. Unlike general land-cover mapping tasks, where classes tend to be more balanced, the **rarity and variability of damage signatures** make it especially difficult for models to learn robust and generalizable representations from a limited number of disaster events.

Then, we wish to clarify the primary scope of our work. As a dataset and benchmark paper, **our central contribution is to capture and present these real-world challenges** in multimodal building damage mapping, including the severe class imbalance, to **provide a realistic and challenging testbed for the community**. Our objective is to establish baselines by evaluating existing models on this data, thereby transparently highlighting this problem and providing a reference point for future studies.

We concur with the reviewer that simply using Lovasz loss is only a partial mitigation, not a complete solution. However, we want to point out that **fully addressing this deep-rooted imbalance is a significant research challenge in its own right**, likely requiring multiple dedicated methodology papers focusing on novel algorithms (e.g., specialized loss functions, data resampling strategies, or generative augmentation). **Such an endeavor, while crucial, extends beyond the scope of a single dataset-focused paper**. Our work aims to provide a foundational dataset to enable and inspire that future research. This is precisely why we highlight this issue in our manuscript: to serve not only as a caution to users but also as a clear focus for future methodological advancements.

Thank you again for providing us with the opportunity to clarify the context and scope of our contribution.

Q3: The discussion on cross-disaster generalization needs to be strengthened. Table 6 shows that in different disaster types, certain events perform particularly poorly. In particular, the mIoU values are the lowest for explosion and chemical accident events such as Bata-EP-2021 and Kyaukpyu-CC-2023, while earthquake events such as Morocco-EQ-2023 and Noto-EQ-2024 also remain highly challenging. This indicates that the models face significant difficulties in handling highly destructive, structurally complex, and spatially heterogeneous disaster scenarios. The authors should analyze these challenges in more depth, such as the heterogeneity and extreme local variations in explosion damage, the diversity of collapse patterns in earthquake events, and the limitations of SAR data in capturing fine-grained details. It is also recommended to provide typical error cases and compare model errors across disaster types to better illustrate the shortcomings in generalization.

R3: We sincerely thank the reviewer for these detailed and highly constructive suggestions. The points raised are crucial for understanding the complexities of cross-disaster generalization, and we appreciate the opportunity to clarify and strengthen our discussion.

Our response to this valuable comment is structured in three parts: 1) we will gently clarify the context of the tables to ensure a common understanding of the results; 2) we will guide the reviewer to the sections in our manuscript where we already performed the in-depth analysis you suggested; 3) we will describe the new content we have added based on your excellent recommendation.

First, we would like to gently clarify this point to avoid any potential misunderstanding regarding the tables. **The results in Table 6 are part of the standard machine learning evaluation**, providing an event-wise breakdown of the overall results shown in Table 5. The purpose of Table 6 is to prevent the evaluation from being dominated by events with a large number of samples, thus offering a more granular view of model performance on each disaster event. The experiments specifically designed to evaluate cross-event transfer generalization are presented and discussed in Section 5.4 (Section 4.6 in the revised manuscript), with the quantitative results shown in Table 8 (Table 10 in the revised manuscript). We apologize if this structure was not sufficiently clear.

Secondly, we are pleased that the reviewer highlighted these critical areas for analysis, as we also believe they are central to understanding the challenges. We would like to respectfully guide the reviewer to the following sections of our manuscript where **some of these points were discussed** in detail:

- Regarding the heterogeneity of damage, we analyzed this extensively in [Section 5.4.3: Why is cross-event transfer challenging] (Section 4.6.3 in the revised manuscript). Specifically, **Figure 9 provides violin plots** that visualize the significant shifts in SAR backscatter distributions for damaged and destroyed buildings across different events, including those of the same disaster type. This directly addresses the “heterogeneity and extreme local variations” and “diversity of collapse patterns” that the reviewer mentioned.
- Regarding model performance across disaster types and SAR limitations: This was analyzed in [Section 5.2 - What have the models learned and what can they learn] (Section 4.2 in the revised manuscript). The **bar chart in Figure 8** directly compares the models’ average IoU across seven major disaster types. The accompanying text discusses the varying performance, explicitly noting the models’ accuracy/errors across disaster types.

Finally, we **completely agree with the reviewer that providing typical error cases is an excellent way** to visually illustrate the model’s generalization shortcomings. While our original analysis was primarily quantitative, visual examples provide a more intuitive understanding of the specific failure modes. Therefore, we have now added **a new figure (Figure 9 in the revised manuscript) and corresponding analysis to Section 4.2** in the revised manuscript. This new content shows concrete examples of model misclassifications. We believe this addition, prompted by the reviewer's valuable suggestion, significantly strengthens our analysis by bridging the quantitative results with qualitative, real-world examples of model errors.

Here, we show the revised part below for your convenience.

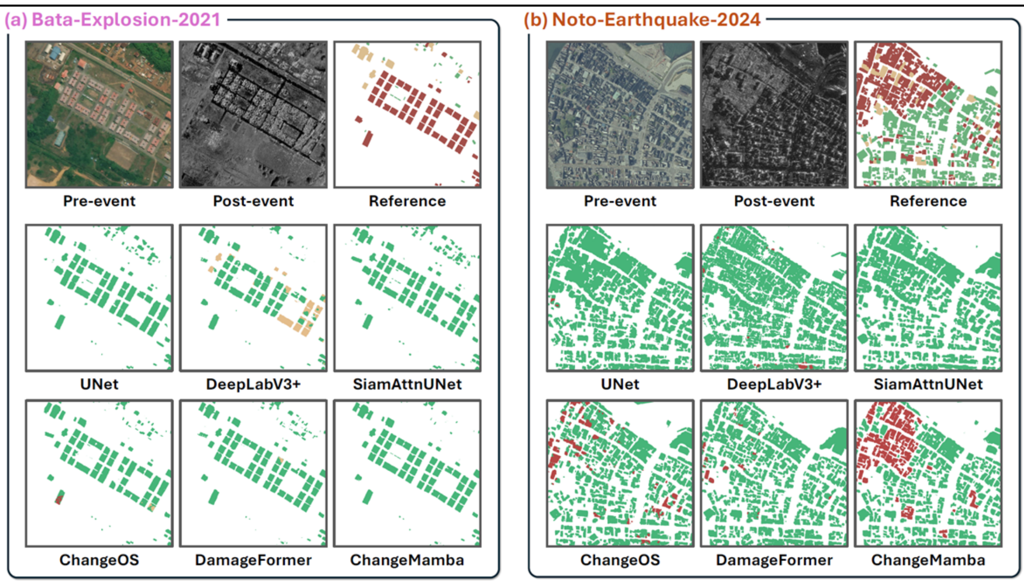

Figure 9. Typical failure cases of different models on Bata-Explosion-2021 and Noto-Earthquake-2024 in BRIGHT.

heterogeneous patterns of structural collapse typical of seismic events, where damage is often subtle, partial, and highly variable. These conditions pose significant challenges for SAR-based assessment. Interestingly, the model achieves relatively high IoU scores in the flood and hurricane events for the "Damaged" category with approximately 50% and 60%, respectively. This indicates that SAR effectively captures contextual environmental changes, such as water inundation or terrain disruption, which indirectly aid in assessing building damage. In the case of the conflict event, the model's performance on the "Destroyed" class is surprisingly low. This might be attributed to the limited number of destroyed samples in the dataset for this category, which leads to insufficient learning and poor generalization.

These quantitative limitations are vividly illustrated by the typical failure cases shown in Figure 9. In the Bata-Explosion-2021 event, models misclassify severely destroyed buildings as intact, reflecting the difficulty of interpreting heterogeneous debris patterns. Similarly, in the Noto-Earthquake-2024 event, large-scale collapses are largely missed, highlighting the challenge of diverse and subtle seismic damage. These examples visually confirm that while models can leverage broad contextual cues, they still struggle to distinguish partial from complete collapse where SAR backscatter changes are weak or inconsistent.

In summary, these findings confirm both the promise and limitations of optical-SAR modality for all-weather, global-scale disaster response. Although this combination performs well in events characterized by large-scale surface disruption (e.g., wildfires, volcanoes), it struggles with subtle or localized damage patterns. Incorporating richer data sources, such as fully polarimetric SAR and LiDAR data, can further enhance the accuracy and reliability of future all-weather building damage assessments.

Q4: The manuscript currently lacks a comparison with optical-only baselines, which is crucial to highlight the value of multimodal methods. Readers may question whether the inclusion of SAR brings significant benefits and whether the additional cost of multimodality is justified. To avoid such doubts, I suggest adding experiments with optical-only inputs and comparing them with Optical+SAR results. This would further emphasize the unique value of the BRIGHT dataset and provide stronger evidence for the necessity of multimodal fusion.

R4: We sincerely thank the reviewer for this excellent suggestion. The question of how multimodal methods compare against an optical-only baseline is fundamental to justifying their value, and we appreciate the opportunity to provide a detailed clarification and new experimental evidence.

First, we would like to respectfully clarify the specific scenario that the BRIGHT dataset is designed to model. The core premise of BRIGHT is to facilitate **all-weather, rapid disaster response**. It is constructed around the common and challenging real-world situation where a disaster event (e.g., a hurricane, flood, or wildfire) is followed by adverse atmospheric conditions (e.g., cloud cover, smoke) that prevent the timely acquisition of usable post-event optical imagery. Therefore, **the dataset's composition is intentionally <u>pre-event optical + post-event SAR</u>**. This represents a pragmatic and operationally vital workflow. Consequently, a direct **<u>"optical-only" baseline</u>** you suggested (i.e., pre-event optical + post-event optical) **<u>is not feasible on the main BRIGHT dataset</u>** by design, as high-quality post-event optical imagery is not a component of its primary structure.

However, we completely agree with the reviewer that a direct comparison is crucial for understanding the relative strengths of each modality when both happen to be available under ideal conditions. To address this valuable point, we have **conducted a new set of experiments on a specific subset of our data**, including Bata-Explosion-2020, Beirut-Explosion-2021, Hawaii-Wildfire-2023, Libya-Flood-2023 and Noto-Earthquake-2024, for which high-quality, cloud-free and preprocessed post-event optical imagery was also available to us.

We benchmarked three different setups on this subset:

- Optical-Only: Pre-event optical + Post-event optical
- SAR-Only (**BRIGHT's standard**): Pre-event optical + Post-event SAR
- Optical+SAR Fusion: Pre-event optical + Post-event optical + Post-event SAR

As the results show (**<u>Table 8</u>** in the revised manuscript), when high-quality, cloud-free post-event optical imagery is available, the post-event optical approach outperforms the post-event SAR approach across all tested models. For instance, using the state-of-the-art **DamageFormer** model, the **post-event optical setup achieves a final mIoU of 69.76%**, higher than the **65.56% from the post-event SAR setup**. This is expected, as optical imagery provides rich and intuitive visual information for damage assessment. Crucially, the results also demonstrate that fusing both modalities consistently provides the best results, outperforming even the strong post-event optical methods in every case. For example, DamageFormer's mIoU increases from 69.76% (post-event optical) to **70.79% with the addition of SAR data**, suggesting that SAR provides complementary information that can enhance the results even when high-quality optical data is present.

While this experiment provides a valuable benchmark, its results ultimately reinforce our motivation for BRIGHT. The high performance of the **optical-only model is entirely contingent on the availability of ideal, cloud-free post-event imagery**, a condition frequently not met in the critical window after many disasters. Therefore, **<u>the small performance trade-off of the post-event SAR-based approach is justified by its invaluable all-weather, day-and-night operational capability</u>**. BRIGHT is designed precisely to advance the development of models for these realistic, often non-ideal, but operationally critical scenarios. We have added this new experiment and discussion to the new Section in the revised manuscript to further emphasize the unique value of our dataset and the necessity of multimodal fusion.

Here, we show the revised part below for your convenience.

Table 8. Performance comparison of different post-event modalities on a subset of BRIGHT. Results are reported for UNet, DeepLabV3+, and DamageFormer on five disaster events where high-quality post-event optical imagery is available: Bata-Explosion-2020, Beirut-Explosion-2021, Hawaii-Wildfire-2023, Libya-Flood-2023, and Noto-Earthquake-2024.

Method	Post-event modality	F_1^{loc} (%)	F_1^{clf} (%)	Final mIoU (%)	IoU per class (%)			
					Background	Intact	Damaged	Destroyed
UNet	SAR	85.05	71.43	62.94	94.39	65.60	42.34	49.43
	Optical	86.46	75.64	65.96	94.56	68.62	45.27	55.36
	Optical+SAR	86.70	74.46	66.29	94.72	69.16	41.73	59.57
DeepLabV3+	SAR	83.55	67.52	60.57	93.86	65.62	35.64	47.16
	Optical	85.79	74.39	64.87	94.33	67.90	44.31	52.94
	Optical+SAR	85.90	74.87	65.84	94.48	69.60	44.68	54.60
DamageFormer	SAR	88.41	73.43	65.56	95.30	70.62	41.31	55.00
	Optical	88.32	78.04	69.76	95.37	72.72	47.26	63.68
	Optical+SAR	88.86	79.27	70.79	95.56	73.64	48.44	65.51

4.4 Impact of post-event modality on building damage assessment performance

Although the primary design of BRIGHT is to facilitate all-weather disaster response through the use of pre-event optical and post-event SAR imagery, it is also important to understand how these modalities compare when high-quality post-event optical imagery is available. To this end, we conducted supplementary experiments on a subset of events, including Bata-Explosion-2020, Beirut-Explosion-2021, Hawaii-Wildfire-2023, Libya-Flood-2023, and Noto-Earthquake-2024, for which pre-processed post-event optical data were accessible. We evaluated three experimental setups: (i) optical-only (pre-event optical + post-event optical), (ii) SAR-only (pre-event optical + post-event SAR, i.e., the standard BRIGHT setting), and (iii) optical+SAR fusion (pre-event optical + post-event optical + post-event SAR).

Table 8 presents the experimental results. As expected, when ideal post-event optical imagery is available, the optical-only setup achieves higher performance than the SAR-only setup. For example, with DamageFormer, the optical-only configuration reaches a final mIoU of 69.76%, compared to 65.56% for SAR-only. Importantly, the performance gap between optical and SAR is not substantial, demonstrating that SAR alone provides a strong alternative in the absence of usable optical imagery. Moreover, the fusion of optical and SAR consistently yields the best results across all tested models. For instance, DamageFormer's mIoU further increases to 70.79% with Optical+SAR fusion, indicating that SAR contributes complementary information that strengthens performance even under optimal optical conditions.

These findings underscore two important insights. First, multimodal fusion is beneficial even when high-quality optical data are available, as SAR provides unique structural information that enriches the optical signal. Second, the performance of the SAR-only approach, being reasonably close to the optical-only results, highlights the practical value of SAR in real-world disaster scenarios where post-event optical imagery is often unavailable. BRIGHT is therefore designed to advance the development of models for these realistic, often non-ideal, but operationally critical all-weather disaster response settings.

Q5: The discussion of limitations and future directions is insufficient. At present, the conclusion mainly emphasizes the dataset's contributions, but it does not address its shortcomings in detail. It is suggested to include a separate subsection summarizing the limitations, such as the use of single-polarization SAR, the lack of time-series data, and the fact that most disaster events are concentrated after 2020.

R5: We sincerely appreciate this constructive suggestion to improve the discussion on the limitations of our dataset. We fully agree that a thorough and transparent account of the dataset's shortcomings is essential for the community. Just for clarity, our original manuscript actually already included relevant content in Section 6 – Discussion:

- **Section 6.1 - Limitation of BRIGHT**: to address what we identified as the primary limitations, including potential registration errors, label quality, and sample/regional imbalance.
- **Section 6.2 - Significance of BRIGHT**: by suggesting that "Incorporating richer data sources, such as fully polarimetric SAR and LiDAR data, can further enhance the accuracy and reliability of future all-weather building damage assessments".

That said, we acknowledge that our initial discussion did not explicitly address two important points raised by the reviewer: the absence of time-series data and the temporal concentration of disaster events after 2020. We greatly appreciate this observation. In response, we have **revised Section 6.1 (Section 5.1 in the revised manuscript) to incorporate these limitations**, thereby providing a more comprehensive discussion. We believe these additions, prompted by the reviewer's insightful comment, have significantly strengthened the manuscript. We thank the reviewer again for helping us improve the quality of our work.

Here, we show the revised part below for your convenience.

5 Discussion

5.1 Limitation of BRIGHT

We begin this subsection by acknowledging that the composition of the BRIGHT dataset is fundamentally shaped by practical constraints in data availability. While BRIGHT represents a significant step forward in assembling a large-scale, multimodal, and globally distributed dataset for disaster response, it is important to recognize several inherent limitations. These limitations arise not only from the scarcity of open-access VHR SAR imagery, especially over disaster-affected regions, but also from the challenges of manual annotation and the uneven distribution of events. To provide a clearer picture for potential users, we summarize these constraints in four aspects below.

4. **Modality and temporal scope.** The dataset's scope is defined by two key characteristics of the available data. First, it exclusively utilizes single-polarization SAR imagery. The current version lacks the more informative multi-polarization or dense time-series SAR data, which, if available, could enable more nuanced damage characterization and long-term recovery monitoring, respectively. Second, the dataset's temporal coverage is concentrated on events from 2020 onwards. This is a direct consequence of its reliance on modern commercial VHR SAR providers (Capella Space and Umbra), whose open-data initiatives largely commenced around that time.

Q6: The description of study areas and disaster events is somewhat redundant.

R6: Thank you so much for this valuable suggestion. In preparing the manuscript, we followed the style of other disaster-related papers published in ESSD, which typically provide detailed descriptions of study areas and events. However, we agree with you that presenting too many event-specific details in the main text can be overwhelming and may distract readers from the core contributions of the work.

In response, we have **revised the structure of the manuscript to streamline this section**. Specifically, we have **moved the detailed descriptions of individual disaster events to the Appendix**, while retaining a concise overview in the main text. Specifically, the general description originally included in Section 2 has been merged into Section 3, now serving as its opening subsection. We believe this restructuring improves the readability of the manuscript by highlighting the key information while still making the detailed event descriptions accessible to interested readers.

Here, we show the revised part below for your convenience.

> **2 Dataset Description**
>
> **2.1 Study areas and disaster events**
>
> We selected 14 disaster events across the globe for BRIGHT, as illustrated in Figure 2 and Table 2. Since both Capella Space and Umbra satellites were launched in 2020, we focused on study areas where disasters have occurred since then. The selected regions are primarily in developing countries, where public administration and disaster response capacities tend to be weaker compared to those in developed nations, making international assistance more critical. The dataset covers five major types of natural disasters: earthquakes, storms (including hurricanes and cyclones), wildfires, floods, and volcanic eruptions. Additionally, it includes human-made disasters, such as accidental explosions and armed conflicts. Detailed descriptions of the 14 disaster events are provided in Appendix A.

Q7: The terminology for "one-shot" may not be accurate. The authors describe it as using "a small number of labeled samples," which may be better defined as "few-shot." Since these concepts are borrowed from previous work, it is suggested to cite the corresponding references.

R7: We sincerely thank you for this precise and helpful comment. We agree that a clear and accurate definition of terms like "one-shot" and "few-shot" is crucial for methodological rigor.

The reviewer is correct that the phrase "a small number of labeled samples" generally corresponds to a "few-shot" learning scenario. Our intention in **using this more general phrase** in the original manuscript was to illustrate the practical context of disaster response, where it might be feasible for experts to quickly label a handful of examples from a new event. We recognize that this phrasing created an ambiguity between the real-world analogy and our specific experimental setup. Therefore, we **have modified the corresponding contents** in the revised manuscript.

Furthermore, we would like to take this opportunity to clarify a key distinction in our application of the "one-shot" paradigm. In many computer vision tasks, one-shot learning typically refers to learning to recognize new semantic classes from a single example. In our work, however, the classes (e.g., intact, damaged, destroyed) remain consistent across all disaster events. Our challenge is not learning new classes but rather adapting the model to a new data domain. That is a new, unseen disaster. Therefore, we use the one-shot sample to facilitate cross-event adaptation, helping the model adjust to the unique visual characteristics, sensor properties, and damage signatures of the target event.

To ensure our manuscript accurately reflects this, we have revised the description to be more precise about both the terminology and the methodological context. Here, we show the revised part below for your convenience.

> • One-shot setup: Recognizing the difficulty of the zero-shot setup, we introduce a one-shot setup. This setting simulates a realistic scenario where a single, representative sample from the new disaster can be quickly labeled to guide model adaptation. In this setting, a limited subset of labeled data (one pair for training and one pair for validation) from the target disaster event is incorporated into the training process. At the same time, the majority of the test set remains unseen. This setup evaluates the model's ability to leverage a minimal amount of manually labeled data to improve disaster-specific adaptation.
>
> It is worth noting that our cross-event transfer setup differs from classic few-shot learning tasks in the computer vision field (Amirreza Shaban and Boots, 2017; Wang et al., 2020). Our goal is not to recognize new classes, but to adapt the model's knowledge of existing classes to a new domain, *i.e.*, an unseen disaster event.

Q8: At line 420, the authors state that SAR is not sensitive to fine structural changes. Would this limitation reduce the value of multimodality in certain scenarios?

R8: Thank you for your thoughtful comment. We thank the reviewer for raising this insightful question. It is important to clarify that **every remote sensing modality captures only certain aspects of the Earth's surface, and each has its own strengths and limitations**. For example, optical imagery records reflect light in the visible and near-infrared spectrum, while SAR measures backscattered microwave signals. Consequently, each modality is inherently more or less sensitive to particular types of features.

In the case of single-polarization SAR, it is true that very fine structural changes/damages may not be well captured. However, **this limitation is not unique to SAR**. Indeed, even optical VHR satellite imagery can sometimes struggle to detect subtle or small-scale damage, as noted in [1]. Therefore, the challenge of capturing fine-grained structural changes/damages is a broader limitation of satellite data rather than a drawback that renders multimodality less valuable.

The core value of multimodal integration lies in ensuring operational continuity for disaster response, which is the primary motivation for our dataset. As discussed in the Introduction, optical EO data, while semantically rich, is frequently rendered unusable by cloud cover in the critical hours and days following a disaster. SAR data's all-weather capability is not just an advantage; it is often the only viable option for timely data acquisition. **The combination of pre-event optical data and post-event SAR data is therefore a pragmatic and powerful solution for rapid assessment,** even if neither modality alone is perfect.

Of course, in an ideal scenario, incorporating additional modalities such as fully polarimetric SAR, LiDAR, or multi-temporal imagery, would provide more complete coverage of structural damages. As noted in Section 6.2 (Section 5.2 in the revised manuscript), we explicitly highlight this as an important direction for future dataset development.

In summary, while we acknowledge the limitations of single-polarization SAR for detecting fine-scale damage, this represents **a pragmatic trade-off rather than a fundamental flaw**. It does not diminish the value of multimodality but instead highlights the importance of combining complementary data sources to build robust, timely, and practical disaster response systems.

[1] T. Manzini, P. Perali, J. Tripathi and R. R. Murphy, "Now you see it, Now you don't: Damage Label Agreement in Drone & Satellite Post-Disaster Imagery," *Proc. 2025 ACM Conf. Fairness, Accountability, and Transparency (FAccT '25)*, New York, NY, USA, pp. 1998–2008, 2025.