Supplementary material for ARTeFACT: Benchmarking Segmentation Models on Diverse Analogue Media Damage

Daniela Ivanova[1], Marco Aversa[2], Paul Henderson[1], and John Williamson[1]

[1] University of Glasgow, UK
[2] Dotphoton, Switzerland

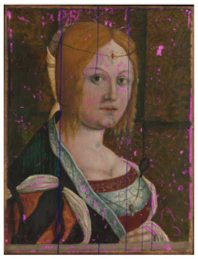
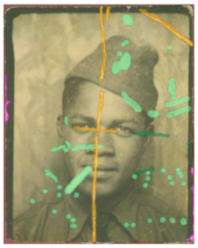
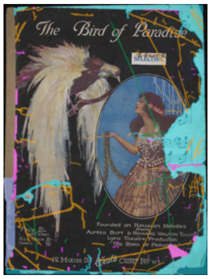
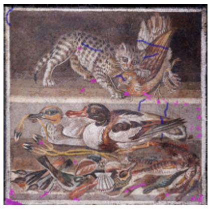
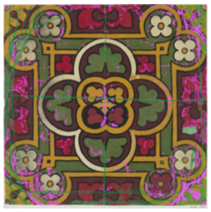
1 Data collection & curation

We sourced most of our dataset from the digitised collections of the following cultural heritage institutions:

- The Metropolitan Museum of Art
- The US National Archives
- The Library of Congress
- The Philadelphia Museum of Art
- The J. Paul Getty Museum
- The Louvre
- The Rijksmuseum
- The Victoria & Albert Museum
- The British Museum
- The British Library
- The National Gallery of Scotland
- The Burrell Collection
- The British National Portrait Gallery
- The National Library of Ireland
- The Fitzwilliam Museum
- The Bavarian National Museum
- The Albert-Kahn Museum

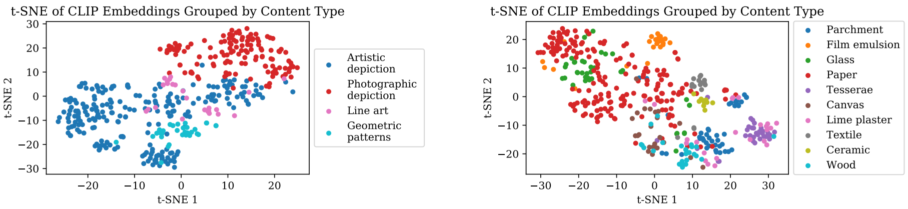
The remainder of the images were collected from Wikimedia, and from Flickr groups dedicated to vintage photographs, such as Wallet Worn and Vintage Scans. All images were made available in the Public Domain by their authors, and collected under CC0, CC-BY and CC-BY-NC licenses. For each image we acquired the highest resolution available, with the average image height and width in our dataset being 2496 and 2306 pixels respectively, and the maximum height and width reaching 11233 and 12797 pixels, allowing for flexibility in preprocessing depending on the task. All text annotations (Content and Materual classifications, Type classification and descriptions) are provided in a tabular format along with source attribution. Ground-truth segmentation masks are provided in a standard format as single-channel PNG files, along colour-coded RGB segmentations for visualisation purposes. Examples shown in Table 1.

Table 1: Example segmentation, class and text annotations from our dataset.

	**Type**: Painting **Material**: Wood **Content**: Artistic depiction **LLaVA description**: A woman wearing a dress and a crown is depicted in a painting on wood. **Our description**: Portrait painting of woman wearing a dress and a crown. **Damage description**: The painting has cracks, dirt spots, scratches and peels.
	**Type**: Photo **Material**: Paper **Content**: Photographic depiction **LLaVA description**: A man wearing a uniform and a hat is smoking a cigarette. **Our description**: Photo portrait of a man wearing a uniform. **Damage description**: The photo has staining, folds, scratches and peels.
	**Type**: Book cover **Material**: Paper **Content**: Artistic depiction **LLaVA description**: A book cover featuring a woman and a bird, with the woman wearing a flowing dress and holding a bird in her hand. **Our description**: Book cover illustration of a woman in a dress in a mirror facing a bird with long feathers sitting on a branch. **Damage description**: The book cover has stickers, writings, folds, scratches, peels and tears.
	**Type**: Mosaic **Material**: Tesserae **Content**: Artistic depiction **LLaVA description**: A mosaic on tesserae depicting a cat and a duck, with the cat standing over the duck, and a bird in the foreground. **Our description**: Mosaic of a cat catching a duck above other ducks eating from a pile of fish and clams in the foreground. **Damage description**: The mosaic has cracks and peels.
	**Type**: Tile **Material**: Ceramic **Content**: Geometric pattern **LLaVA description**: A tile on ceramic depicting a green and red design, with a flower pattern and a cross in the center. **Our description**: Ceramic tile depicting a green and red design, with a flower pattern and a cross in the center. **Damage description**: The tile has discolourations, scratches and peels.

2 Image-level caterogorisation and textual descriptions

2.1 Content and Material categories.

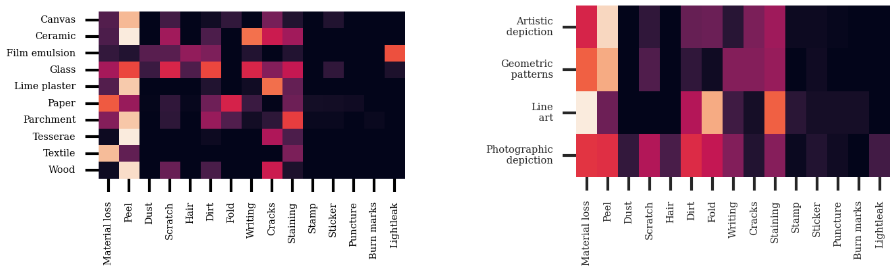


(a) Image embeddings coloured by Content type.

(b) Image embeddings coloured by Material type.

Fig. 1: CLIP embeddings of the images in the dataset, dimensionality reduced via t-SNE. Each colour corresponds to a material (a) or content (b) category as provided in our dataset.

Figure 1a shows that our Content categories assigned by our experts are well separeted in the CLIP image feature space, since CLIP focuses on semantic image content. Material labels form “fuzzier" regions in Figure 1b, demonstrating the diversity of objects depicted in the images across Material classes, while also indicating that certain subject matters are more commonly depicted on certain materials, e.g. Textile will most commonly have stylised Geometric pattern depictions.



(a) Damage type frequency by Material.

(b) Damage type frequency by Content.

Fig. 2: Distributions of kinds of damage across the two proposed splits, Material and by Content.

It can also be observed that certain types of damage occur more frequently in images of certain content, as shown in Figure 2b, however, this is less pronounced than for different material types, as shown in Figure 2a.

Table 2: Cosine distances between the CLIP embeddings of image descriptions generated by LLaVA and our descriptions, averaged across Material and Content classes.

LLaVA vs	**Material**										**Content**			
	Canvas	Ceramic	Film emulsion	Glass	Lime plaster	Paper	Parchment	Tesserae	Textile	Wood	Artistic depiction	Geometric patterns	Line art	Photographic depiction
Ours	0.801	0.861	0.731	0.772	0.744	0.73	0.700	0.792	0.752	0.747	0.742	0.789	0.682	0.751
LLaVA + Damage description	0.917	0.924	0.818	0.895	0.928	0.909	0.884	0.971	0.951	0.935	0.920	0.935	0.923	0.883
Ours + "Damaged"	0.740	0.812	0.649	0.696	0.713	0.659	0.669	0.756	0.722	0.705	0.699	0.746	0.646	0.656
Ours + Damage description	0.761	0.820	0.613	0.716	0.743	0.689	0.665	0.787	0.736	0.725	0.715	0.766	0.660	0.685

2.2 Textual descriptions.

We provide textual descriptions for each image, detailing both content and damage types. The expert descriptions are derived by correcting draft captions produced by LLaVA [1], by prompting it with the following: *This is a <type> on <material>. Describe what this image depicts concisely, in a single sentence, but be as detailed as possible, also noting the initial information about the type of image and the media it is depicted on.* The resulting descriptions were then manually verified and corrected by our annotation experts. We derived separate damage descriptions from our pixel-level annotations. To illustrate how much human experts improve on LLM-generated captions, we computed cosine similarity between LLaVA and expert descriptions using CLIP text embeddings (Table 2), quantifying LLaVA's "incorrectness" regarding semantic image content. We also report distances between original LLaVA descriptions and: extended LLaVA descriptions, extended expert descriptions, and expert descriptions prefixed with "Damaged." Specifying damage presence decreases cosine similarity, showing LLaVA's failure to recognize damage, often ignoring or misinterpreting it. For instance, in the second example in Table 1, LLaVA's caption is `A man wearing a uniform and a hat is smoking a cigarette`, entirely omitting the presence of damage, and misinterpreting a peeled area pf the photograph as a cigarette. In contrast, our expert annotation is `Photo portrait of a man wearing a uniform`, while the damage description reads `The photo has staining, folds, scratches and peels`.

3 Qualitative assessment

3.1 Supervised segmentation

We qualitatively compared the segmentation masks produced by various models across distinct materials. We show the outcomes for both binary and multi-class segmentation tasks. The selected images cover a wide range of damage type where state-of-the-art models encounter difficulties in achieving optimal performance.

Binary segmentation Binary segmentation is the easier setting of the task, where the models learn to differentiate between damaged and non-damaged areas of the images, without needing to classify the type of damage. The qualitative results shown in Figure 3 for Material categories and in Figure 4 for Content categories support our quantitative evaluation results, showing that SegFormer marginally outperforms the other models, but overall model performance is poor. All models are prone to making false positive predictions, and struggle with both small-scale artefacts such as Dust and Hairs in photos. The DinoV2 + MLP model, which is the worst performing according to the quantitative evaluation, fails at predicting large missing areas, and instead only focuses on the edges of such artefacts (e.g in Glass, Tesserae, Artistic depiction categories).

Table 3: Weighted average metrics per damage types across Material categories for each model at the task of multiclass semantic segmentation of damage.

Damage Type	**Segformer**			**UPerNet + Swin**			**UPerNet + ConNeXt**			**DinoV2 + MLP**			**SAM ViT-H + MLP**		
	Acc	**F1**	**IoU**	**Acc**	**F1**	**IoU**	**Acc**	**F1**	**IoU**	**Acc**	**F1**	**IoU**	**Acc**	**F1**	**IoU**
Material loss	0.458	0.543	0.38	0.767	**0.71**	0.572	0.713	0.685	0.548	0.517	0.427	0.277	0.294	0.095	0.051
Peel	0.236	0.294	0.176	0.31	0.371	0.233	0.327	**0.392**	0.249	0.395	0.359	0.226	0.101	0.108	0.059
Dust	0	0	0	0	0	0	0	0	0	0	0	0	0.005	0.004	0.002
Scratch	0.008	0.007	0.004	0.047	0.049	0.026	0.087	**0.101**	0.056	0	0	0	0.017	0.009	0.004
Hair	0	0	0	0	0	0	0	0	0	0	0	0	0	0.001	0
Dirt	0.04	0.047	0.025	0.141	0.132	0.071	0.149	**0.14**	0.076	0.007	0.003	0.002	0.008	0.007	0.004
Fold	0.021	0.018	0.01	0.493	0.489	0.33	0.583	**0.517**	0.356	0.012	0.009	0.005	0.009	0.013	0.007
Writing	0.292	0.245	0.167	0.474	**0.405**	0.263	0.448	0.36	0.224	0.592	0.302	0.197	0.016	0.024	0.013
Cracks	0.051	0.062	0.033	0.099	0.131	0.073	0.124	0.14	0.078	0.245	**0.216**	0.126	0.02	0.029	0.015
Staining	0.038	0.036	0.019	0.245	0.221	0.141	0.287	**0.257**	0.168	0.023	0.024	0.013	0.002	0.003	0.001
Stamp	0	0	0	0	0	0	0	0	0	0	0	0	0.003	0	0
Sticker	0	0	0	0	0	0	0	0	0	0	0	0	0.017	0.002	0.001
Puncture	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Burn marks	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Lightleak	0	0	0	0	0	0	0	0	0	0.008	**0.007**	0.004	0.003	0.003	0.001

Table 4: Weighted average metrics per damage types across Content categories for each model at the task of multiclass semantic segmentation of damage.

Damage Type	**Segformer**			**UPerNet + Swin**			**UPerNet + ConNeXt**			**DinoV2 + MLP**			**SAM ViT-H + MLP**		
	Acc	**F1**	**IoU**	**Acc**	**F1**	**IoU**	**Acc**	**F1**	**IoU**	**Acc**	**F1**	**IoU**	**Acc**	**F1**	**IoU**
Material loss	0.572	**0.522**	0.357	0.576	0.517	0.356	0.621	0.517	0.353	0.539	0.409	0.262	0.306	0.115	0.061
Peel	0.181	0.231	0.133	0.2	0.218	0.124	0.207	0.24	0.138	0.345	**0.333**	0.206	0.069	0.078	0.042
Dust	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Scratch	0.001	0.001	0.001	0.012	**0.02**	0.01	0.011	0.012	0.006	0	0	0	0.04	0.008	0.004
Hair	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Dirt	0.02	0.012	0.007	0.046	0.017	0.009	0.06	**0.075**	0.039	0	0	0	0.002	0.002	0.001
Fold	0.197	0.155	0.086	0.166	0.128	0.07	0.188	**0.186**	0.104	0.194	0.131	0.07	0.021	0.021	0.011
Writing	0.106	0.021	0.012	0.172	0.095	0.063	0.216	0.112	0.068	0.433	**0.14**	0.079	0.033	0.023	0.012
Cracks	0.021	0.006	0.003	0.026	0.011	0.006	0.026	**0.012**	0.006	0.016	0.009	0.005	0.003	0.004	0.002
Staining	0.04	0.034	0.018	0.131	0.077	0.041	0.083	**0.079**	0.041	0.004	0.005	0.002	0.019	0.014	0.007
Stamp	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Sticker	0	0	0	0	0	0	0	0	0	0	0	0	0.057	0.001	0
Puncture	0	0	0	0	0	0	0.051	**0.053**	0.029	0	0	0	0	0	0
Burn marks	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Lightleak	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Multi-class segmentation In the multi-class segmentation setting, models also need to differentiate between the different types of damage. We provide qualitative results in Figure 5 across Material test categories, and in Figure 6 across

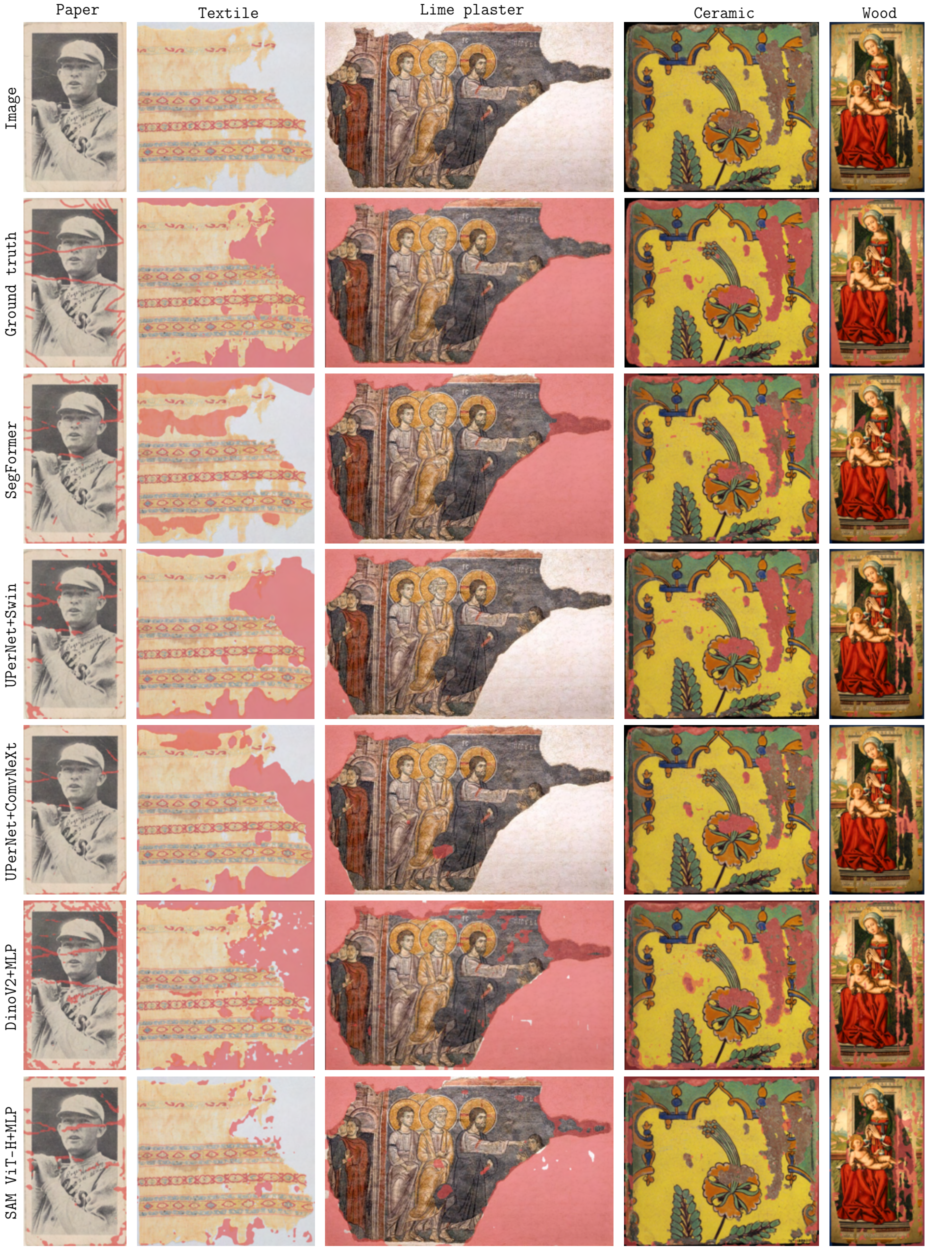


Fig. 3: Qualitative comparison for binary damage segmentation across Material categories.

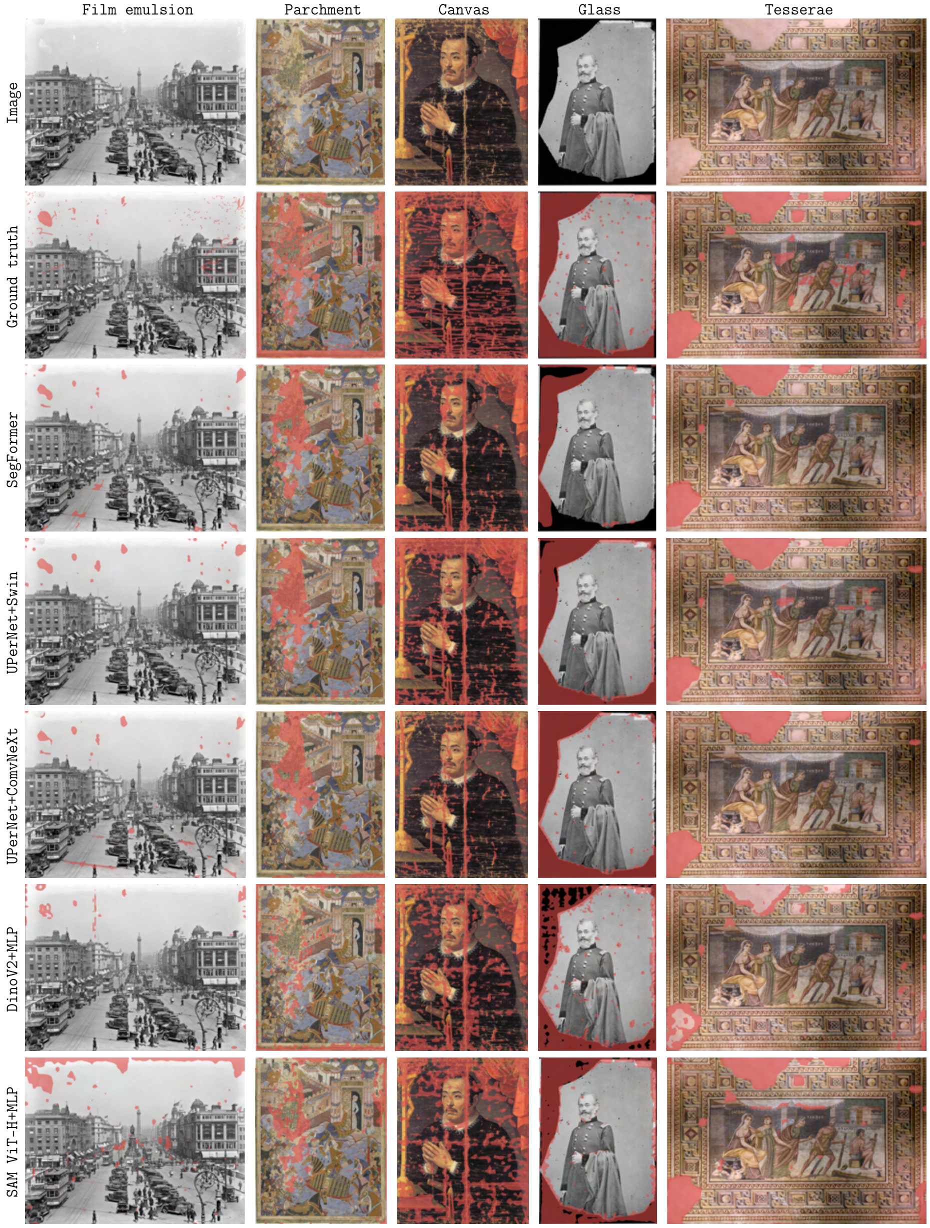


Fig. 3: Qualitative comparison for binary damage segmentation across Material categories.

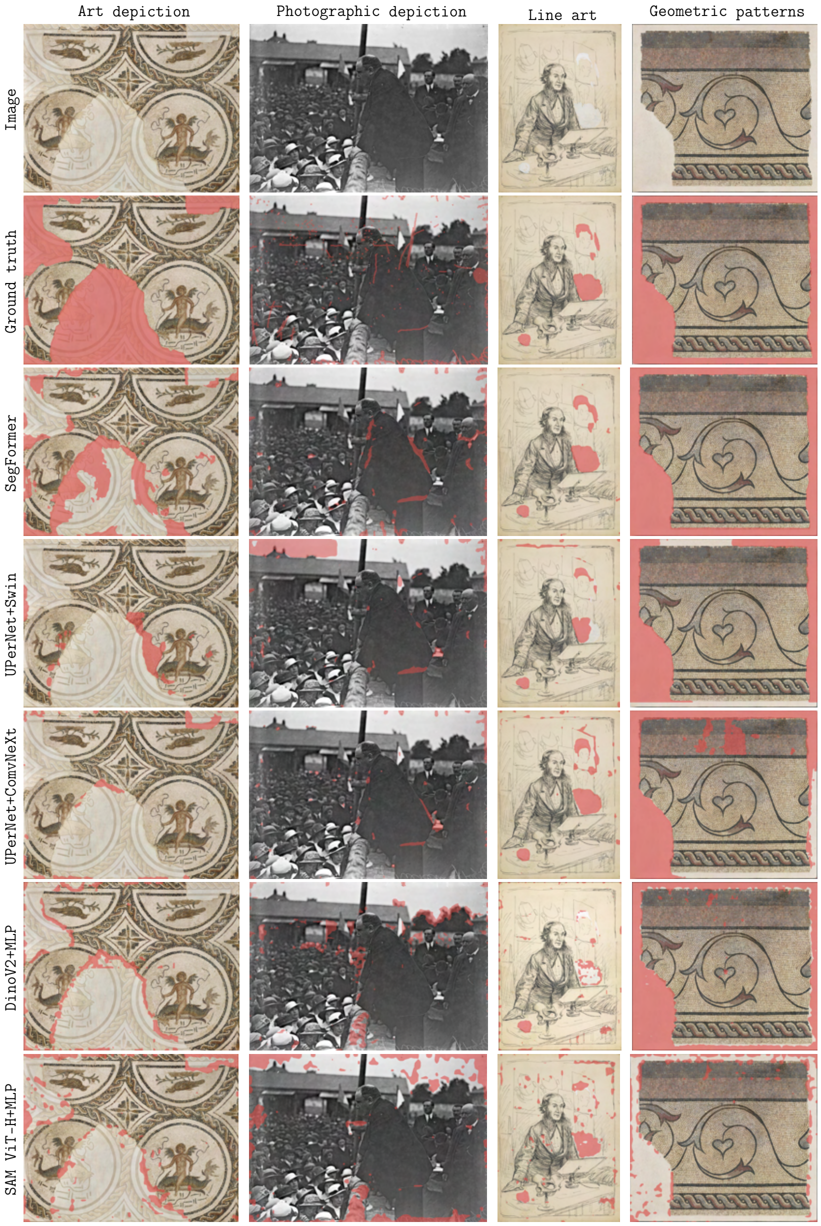


Fig. 4: Qualitative comparison for binary damage segmentation across Content categories.

Content test categories. The predictions are similarly poor as in the binary setting across all four models. We can observe that even when the damaged area is correctly detected, models often assign it an incorrect class. We further examine this behavior by computing metrics for each specific damage type, as summarised in Table 3 and Table 4 for Material and Content categories respectively. The values in the tables are averaged across all LOOCV splits and weighted based on the frequency of Damage type per each Material or Content category. In some cases (LOOCV Content splits), all instances of a certain type of damage belong in the same test class - e.g., all Lightleaks would be in the Photographic depiction category, hence they are either all in the training or the test data.

Qualitatively, performance is exhibited across all damage types in the Content splits, including ones such as Material loss and Peel, which are common in multiple categories. We see that the models cannot successfully segment large-scale damage; they also fail at detecting very small-scale damage types, such as Dust, Scratches, and Hairs, contributing to its poor performance in the Photographic depiction class.

When Lightleaks are not unique to a specific category, as in the Material split, Table 3 shows that the linear probes of DinoV2 and SAM are the only models which score above 0 in all three metrics for this damage type, but still achieve an extremely low score. Dust and Scratches are again where the other three models fail as well; furthermore, all models struggle with Stamps, Stickers, Burn marks.

Where models do make somewhat successful predictions, they show similar performance to the binary setting, highlighting edges and failing at more "blobby" damage instances. Still, false positives are incredibly common, with models detecting arbitrary image features as damage.

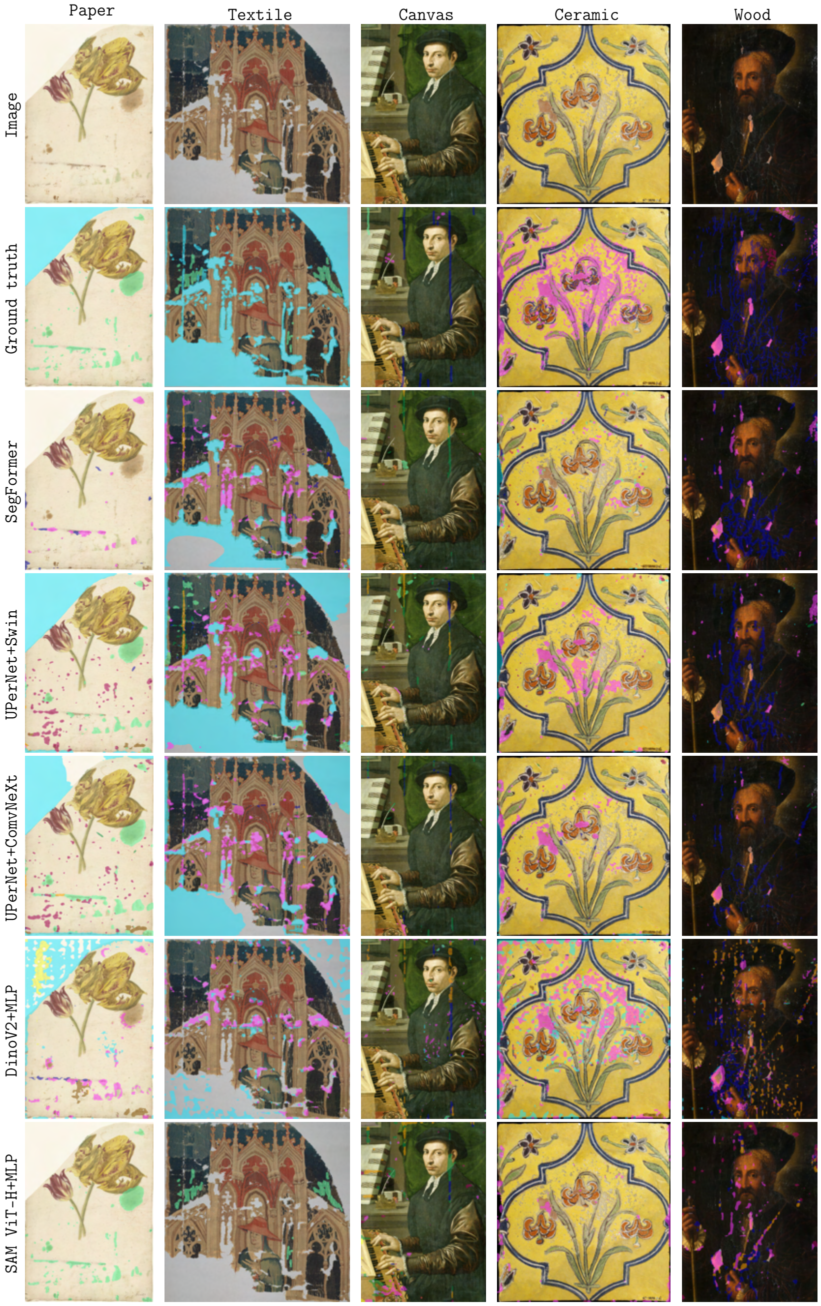


Fig. 5: Qualitative comparison for multiclass damage segmentation across Material categories.

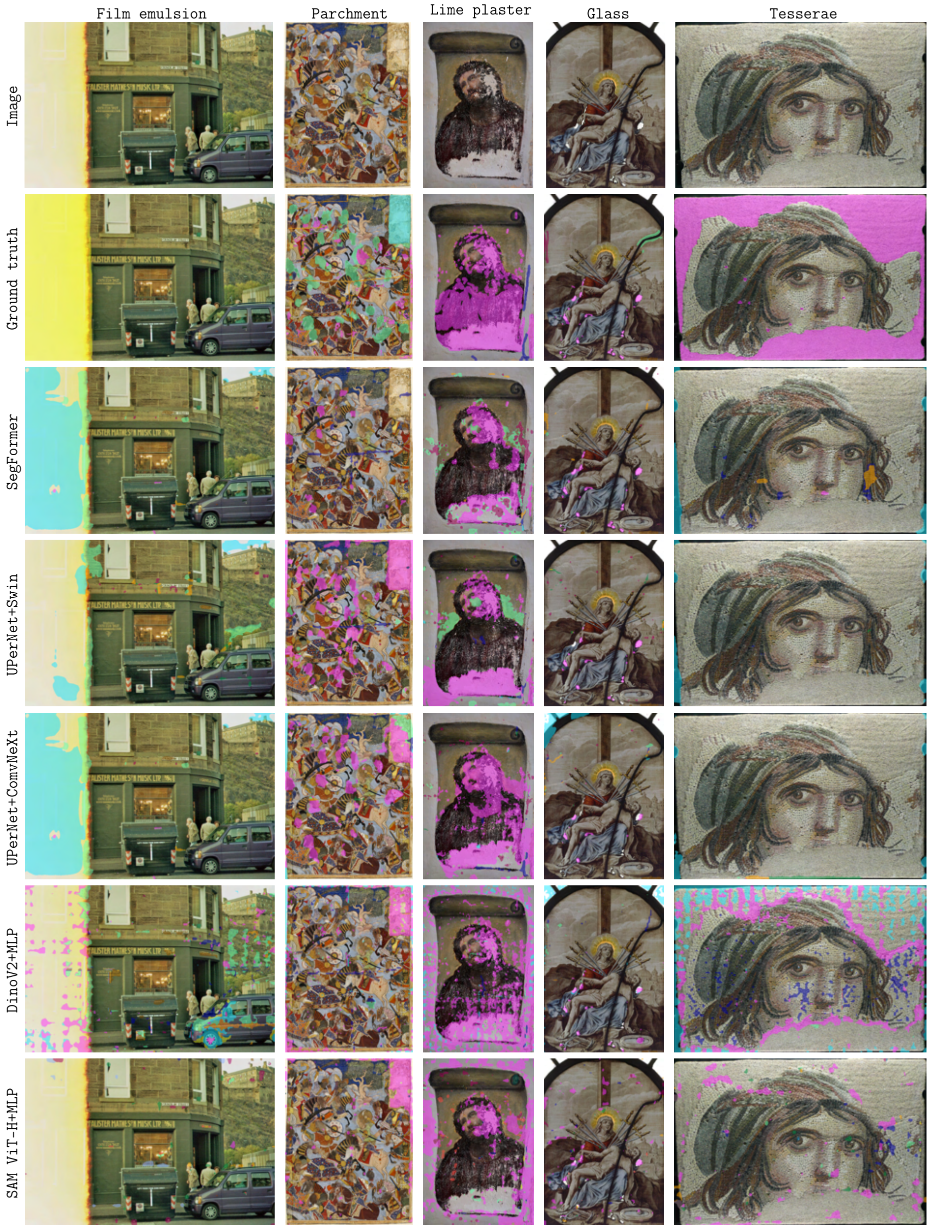


Fig. 5: Qualitative comparison for multiclass damage segmentation across Material categories.

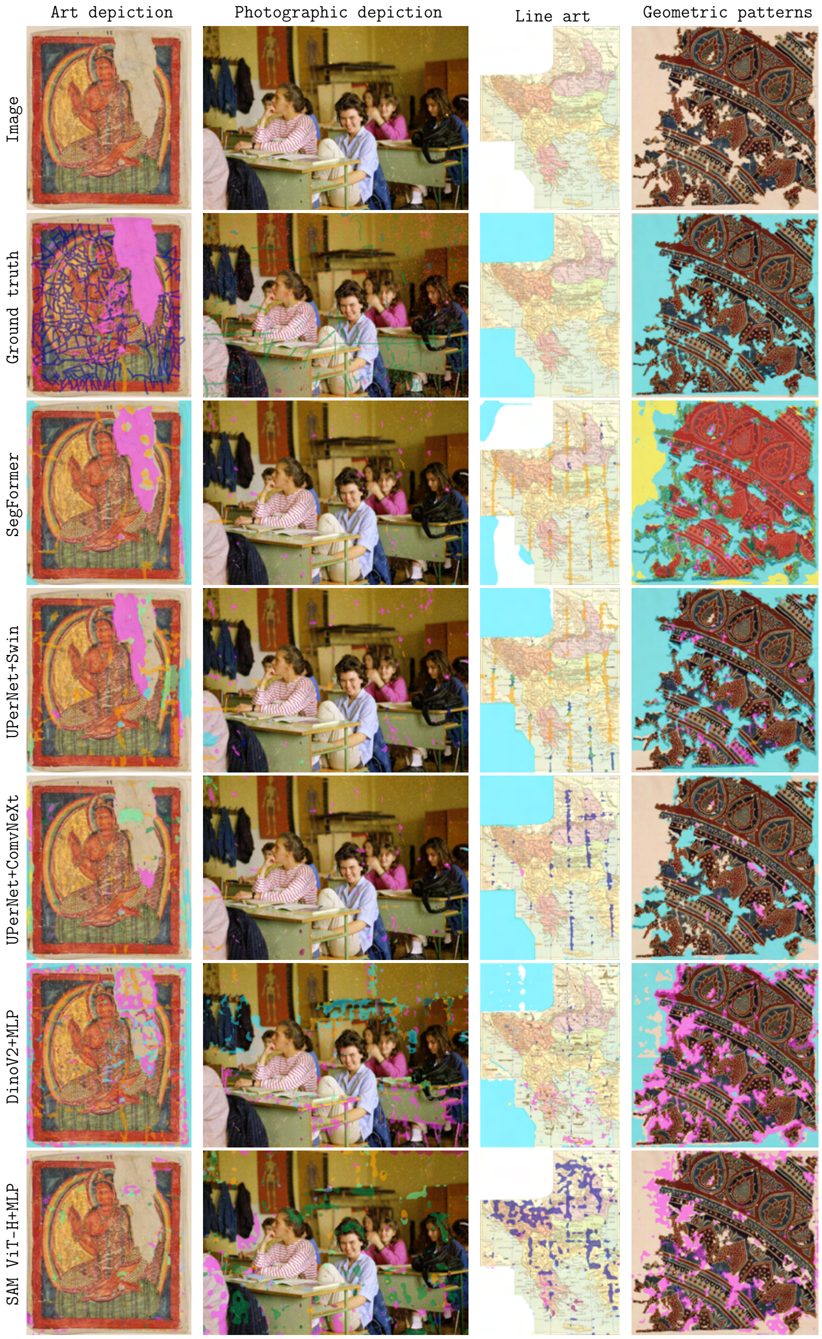


Fig. 6: Qualitative comparison for multiclass damage segmentation across Content categories.

Table 5: Results across both Material and Content splits for DiffEdit binary mask generation in both center cropped and full resolution settings.

Test class	DiffEdit (Crop)			DiffEdit (Full Res)		
	Acc	**F1**	**mIoU**	**Acc**	**F1**	**mIoU**
Wood	0.77	0.13	0.07	0.74	0.15	0.08
Ceramic	0.70	0.13	0.07	0.67	0.19	0.10
Textile	0.79	0.19	0.11	0.74	0.18	0.10
Lime Plaster	0.72	0.18	0.10	0.70	0.18	0.10
Canvas	0.79	0.21	0.12	0.78	0.26	0.15
Tesserae	0.76	0.11	0.06	0.71	0.90	0.05
Paper	0.81	0.16	0.09	0.78	0.19	0.11
Glass	0.78	0.19	0.11	0.76	0.21	0.12
Film emulsion	0.78	0.08	0.04	0.73	0.11	0.06
Parchment	0.83	0.13	0.07	0.78	0.11	0.06
Geom Patterns	0.80	0.20	0.12	0.76	0.19	0.11
Line Art	0.84	0.14	0.08	0.79	0.14	0.08
Photo Depiction	0.79	0.18	0.10	0.77	0.22	0.13
Art Depiction	0.79	0.13	0.07	0.75	0.14	0.08

3.2 Zero-shot generative segmentation

In addition to quantitatively comparing DiffSeg and DiffEdit in the main paper, we visualise qualitative results in Figure 7. Additionally, we show a qualitative comparison between DiffEdit used on center-cropped images and in full resolution following the patch-process methodology described in the main paper, Figure 8. Quantitative results are shown in Table 5. Results are presented over all categories of Material and Content, and cover various types, scales, and degree of damage.

3.3 DiffEdit

We can observe that while the approach may sometimes predict damaged areas correctly, it tends to also highlight generally salient features, discriminative features such as faces, as well as high frequency features, none of which are related to the prompt. The detection of high-frequency features is particularly common when working in full resolution. Since the denoising process is performed over the later portion of steps in the denoising chain, this behaviour can be explained via the property of diffusion models to generate more fine-detailed features at later steps [2]. We find that the approach is extremely sensitive to scale, making inconsistent predictions over center-cropped areas compared to over the entire images. The results support our quantitative evaluation, demonstrating further that language guidance is too imprecise to describe damage. Furthermore, we show that the averaged noise prediction approach used by DiffEdit does not provide disentanglement between predicted noise variance from language conditioning and predicted noise variance over high-frequency features due to the choice of denoising step range. These results illustrate that while modern multi-modal generative approaches may show promising results in general purpose image segmentation via text conditioning, they fail at the task of damage detection - this highlights the complexity of our task.

3.4 DiffSeg

Although DiffSeg achieves better quantitative scores (partly due to the use of an oracle), we can see in Figire 7 that it struggles with higher frequency damage. This is due to the fact that with this approach the segmentation process is performed at a very low-resolutions at which the self-attentions is extracted, the highest being 64×64; the even lower latent resolutions of 16×16 and 32×32 might be helpful when forming general semanic clusters, but here they also fail to capture finer artefacts.

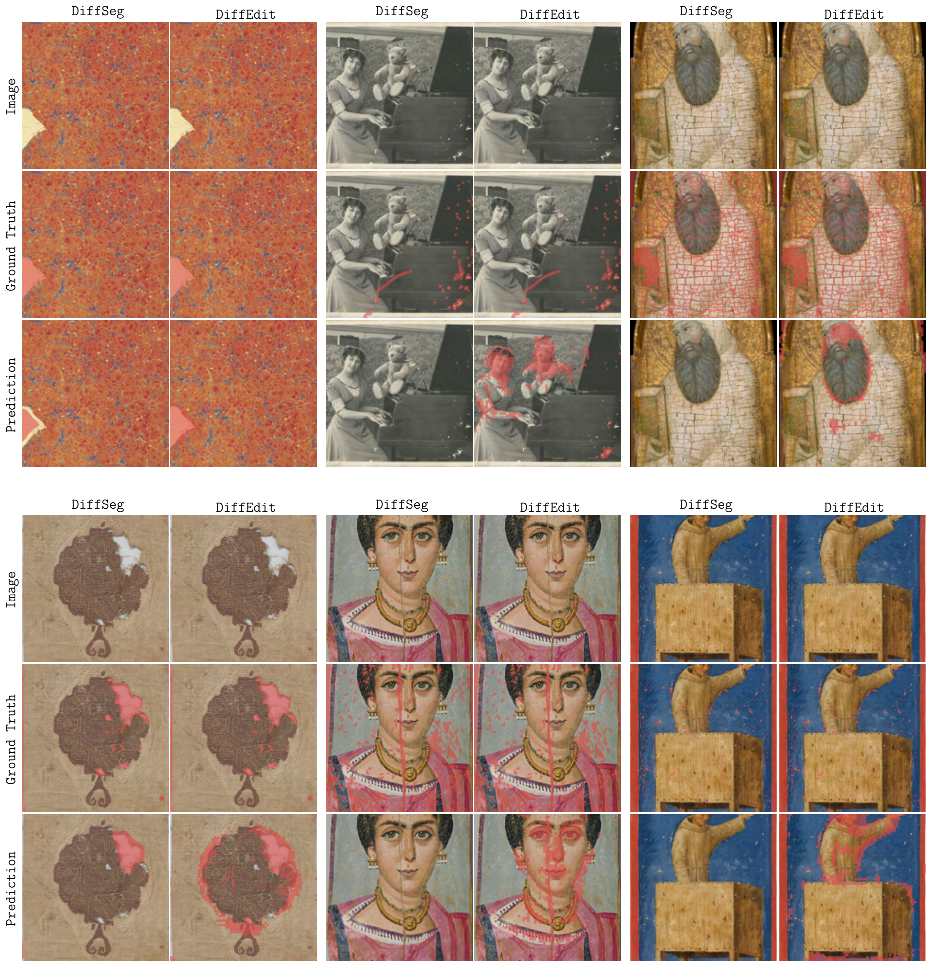


Fig. 7: Qualitative comparison for diffusion-based damage segmentation between DiffSeg and DiffEdit.

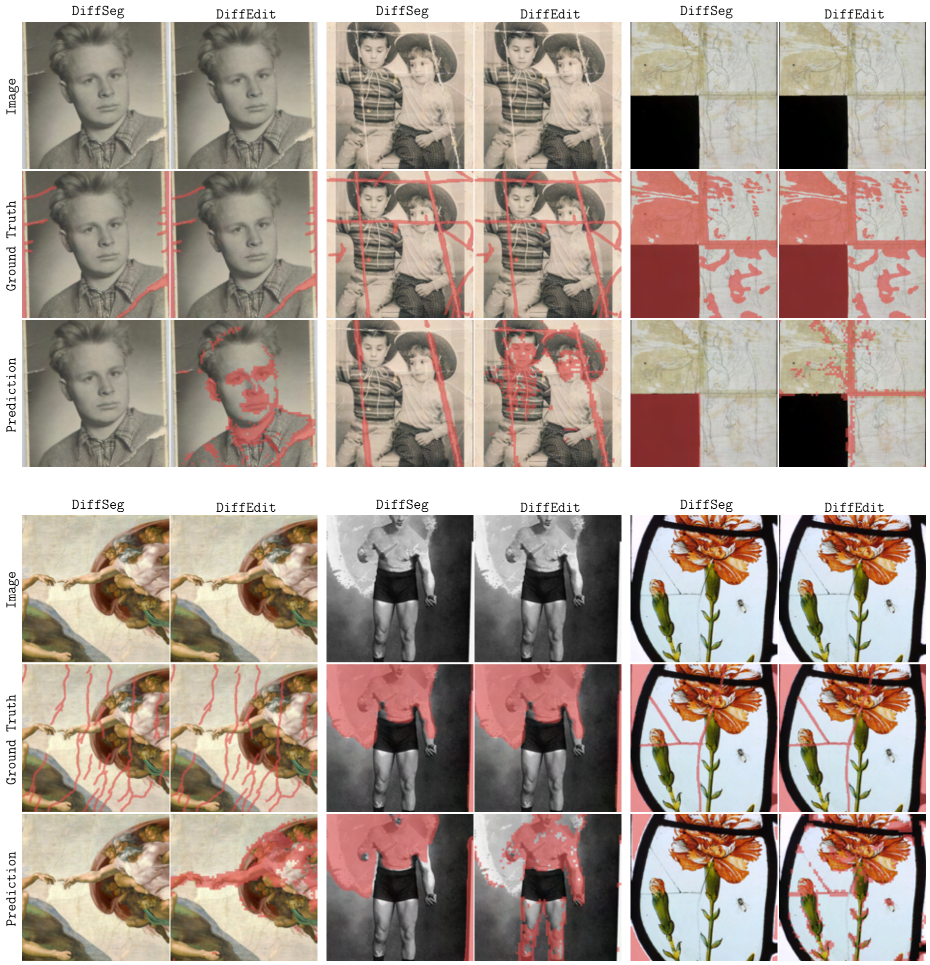


Fig. 7: Qualitative comparison for diffusion-based damage segmentation between DiffSeg and DiffEdit.

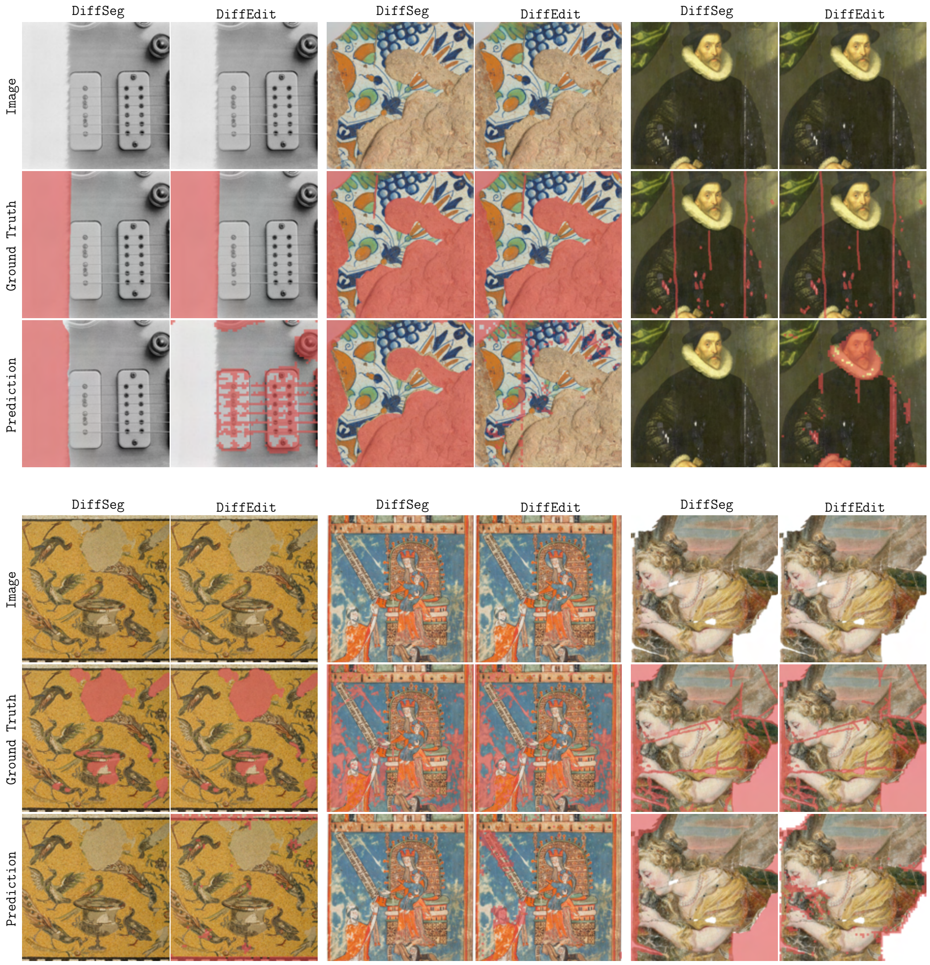


Fig. 7: Qualitative comparison for diffusion-based damage segmentation between DiffSeg and DiffEdit.

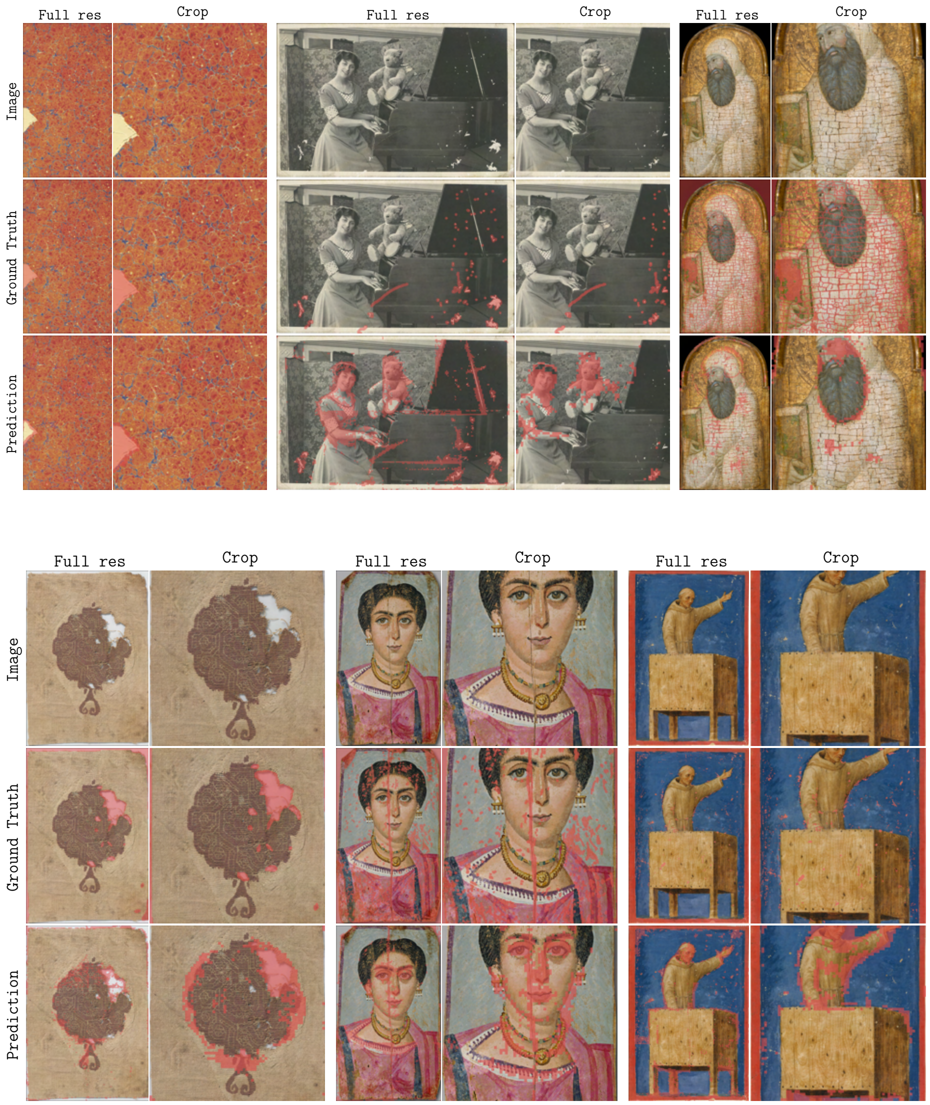


Fig. 8: Qualitative comparison for text-guided damage segmentation using DiffEdit over center-cropped images and in full resolution.

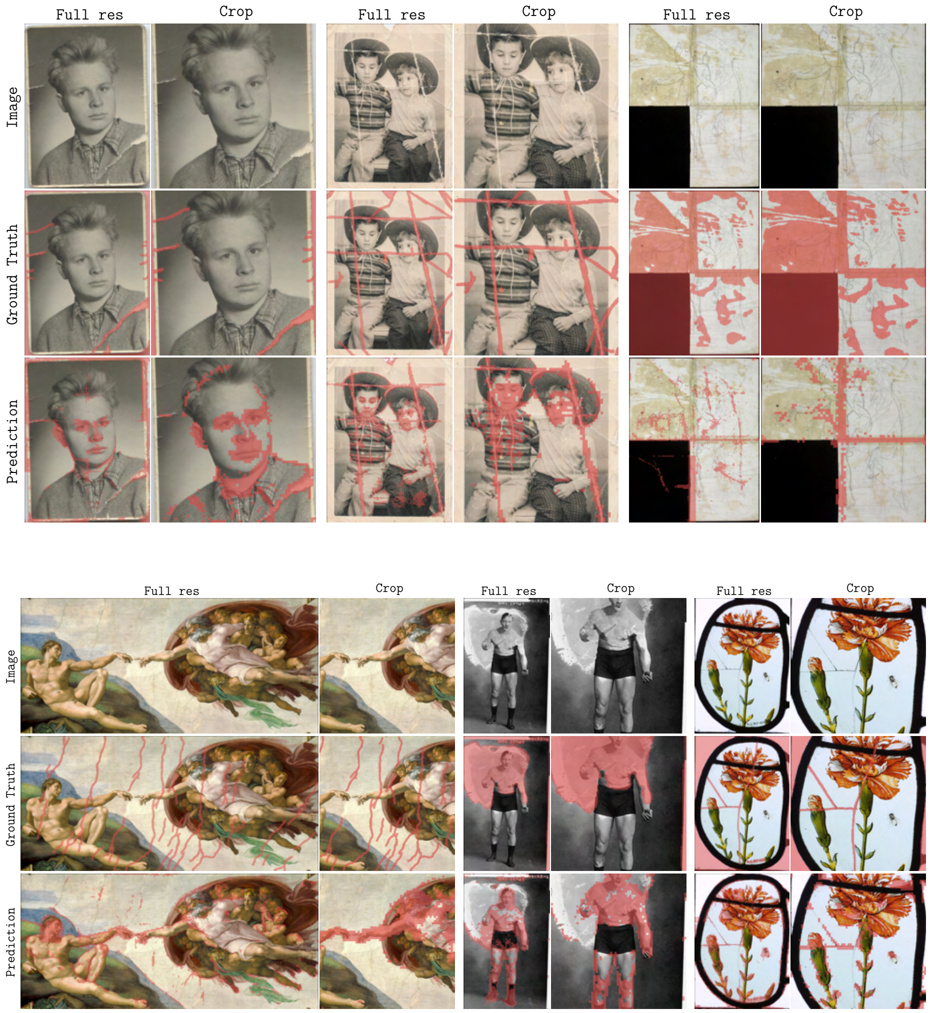


Fig. 8: Qualitative comparison for text-guided damage segmentation using DiffEdit over center-cropped images and in full resolution.

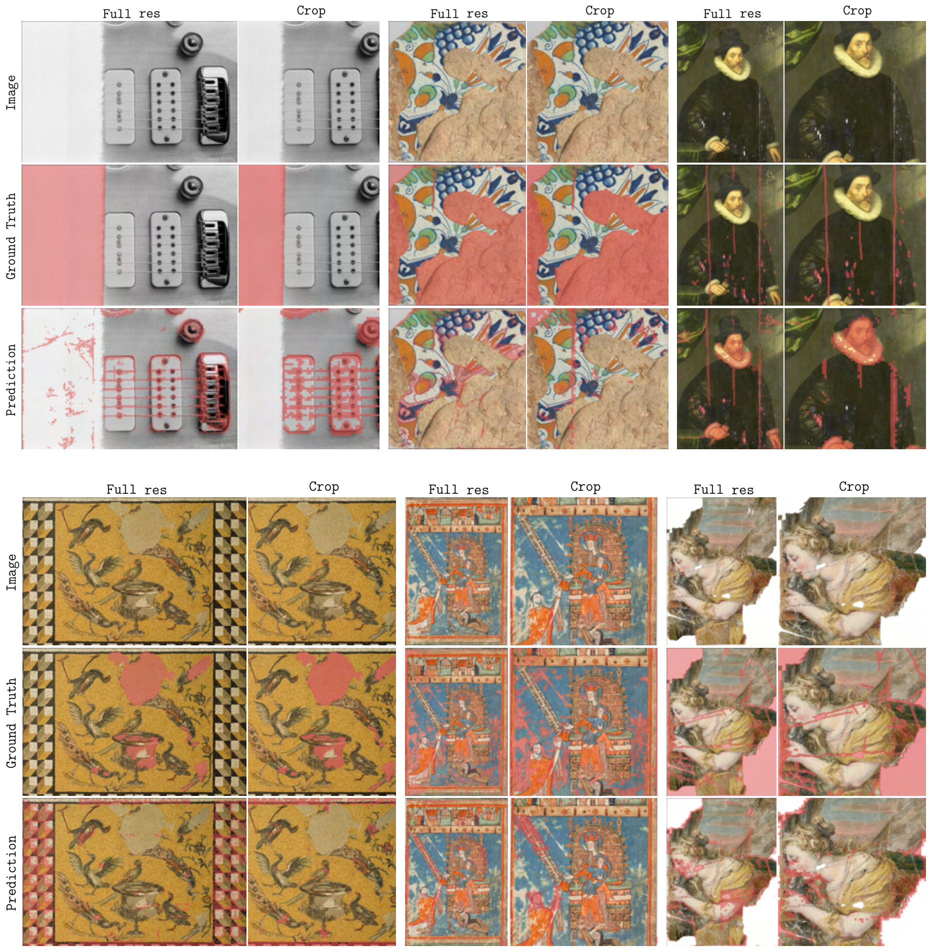


Fig. 8: Qualitative comparison for text-guided damage segmentation using DiffEdit over center-cropped images and in full resolution.

References

1. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeurIPS (2023) 4
2. Luo, G., Dunlap, L., Park, D.H., Holynski, A., Darrell, T.: Diffusion hyperfeatures: Searching through time and space for semantic correspondence. Advances in Neural Information Processing Systems **36** (2024) 13