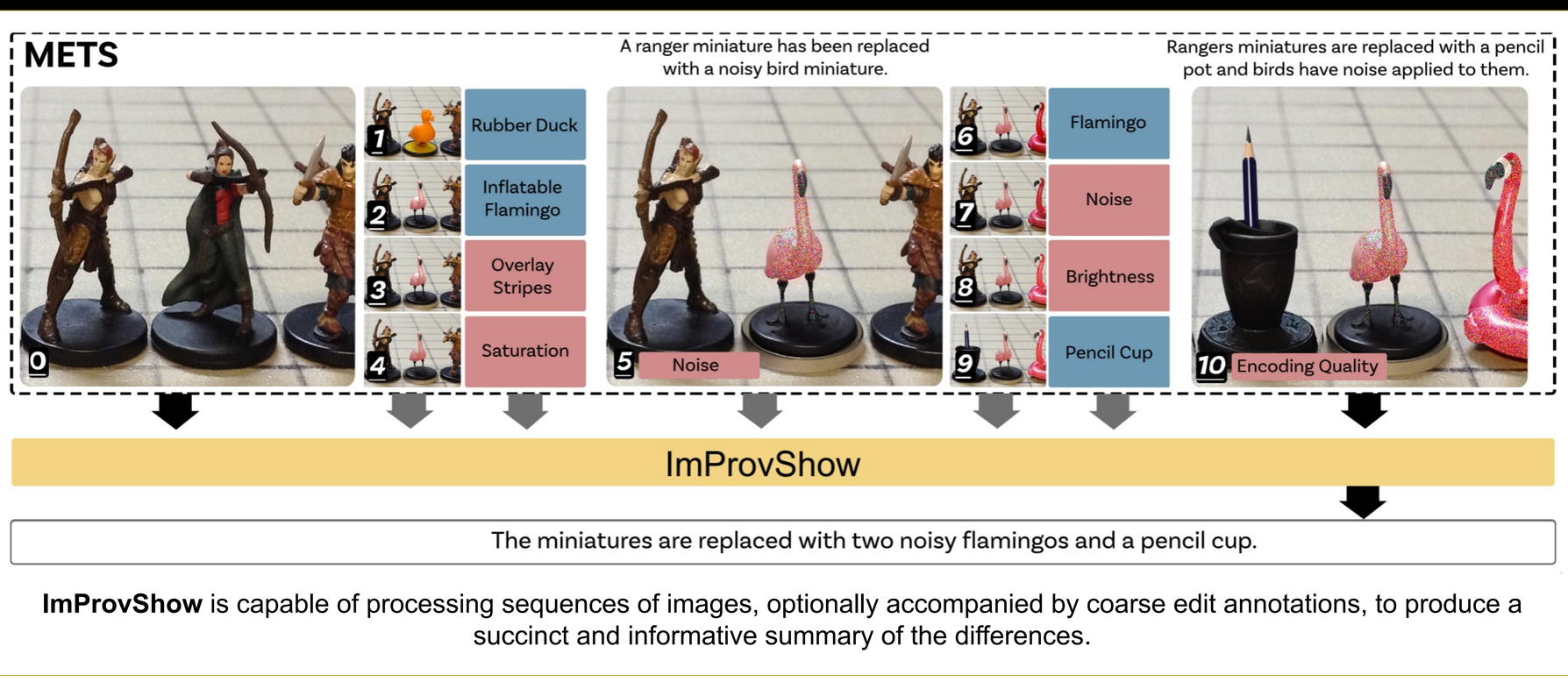
ImProvShow: Multimodal Fusion for Image Provenance Summarization

Alexander Black, Jing Shi, Yifei Fan, John Collomosse

alexander@black.com



ImProvShow is capable of processing sequences of images, optionally accompanied by coarse edit annotations, to produce a succinct and informative summary of the differences.

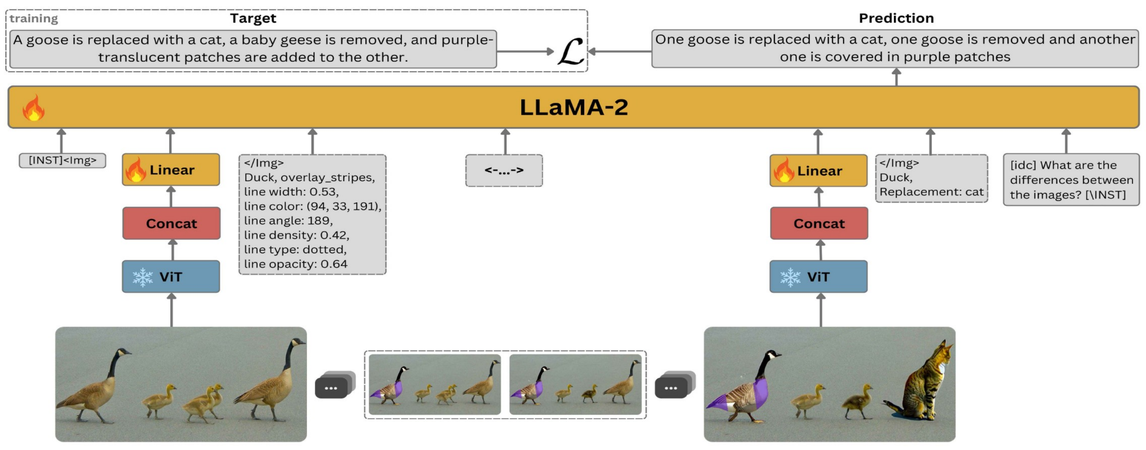
Introduction

Problem: Image manipulation becomes increasingly easier to perform and harder to notice. Often, multiple edits are applied in sequence by one or multiple editors, forming a provenance graph containing multiple versions of the same image at different stages of the editing process. To avoid the spread of misinformation, it is important to be able to communicate the history of these changes to the end user succinctly to enable them to make informed trust decisions.

We propose **ImProvShow** – a novel approach to summarizing the multi-stage edit history (or *provenance*) of an image. ImProvShow fuses visual and textual cues to succinctly summarize multiple manipulations applied to an image in a sequence; a novel extension of the classical image difference captioning (IDC) problem.

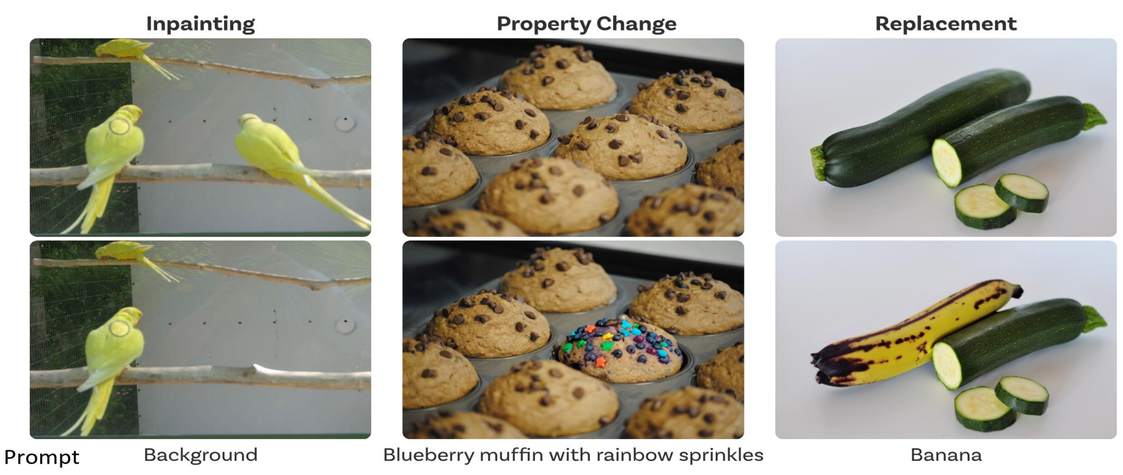
We train it with **METS** (Multiple Edits and Textual Summaries) – a novel dataset of long image editing sequences paired with machine annotations and human-written summaries at multiple steps.

Architecture



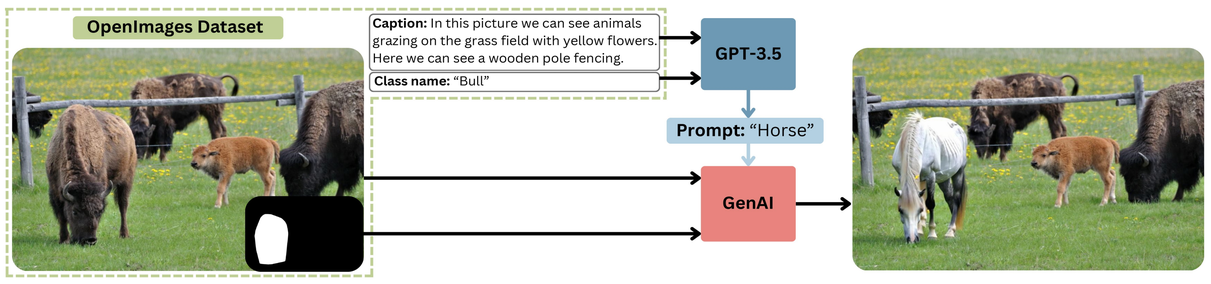
- The language model is conditioned using the multi-modal instruction template, which includes at least two image features and optional auxiliary textual information.
- Auxilary content withing dashed boxes is optional and provides extra context.
- The image features extracted from the ViT image encoder are concatenated in groups of 4 and projected to the LLM embedding space.

METS Dataset – Edit Types



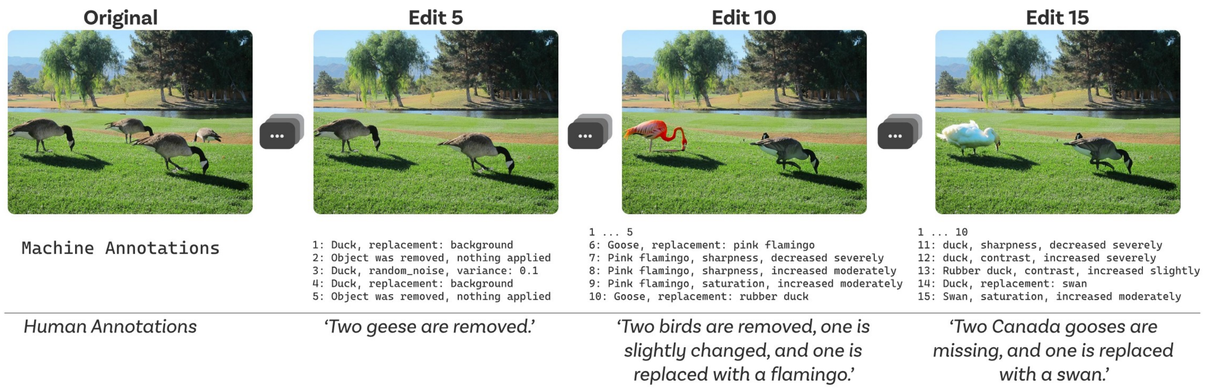
- **Property change:** likely change in color, material, texture or other applicable property of the object.
- **Replacement:** a replacement object of close match to the original shape and size (based on bounding box data) but different semantically.

METS Dataset – Sequence of Edits Generation



- Images, captions and segmentation masks from OpenImages Dataset.
- Plausible manipulation prompt generatedby LLM (GPT-3.5).
- Image + mask + prompt into diffusion generative model (Firefly Generative Fill).

METS Dataset – Annotations



- **Machine Annotations:** metadata of object, manipulation type and manipulation specific parameters.
- **Human Annotations:** We collect human annotations for difference summarization at the 5th, 10th, and 15th step of the manipulation sequence.

Evaluation

Model	Images	Text	B4	C	M	R	L (M)	L (H)
			METS					
GPT3.5 [5]	0	yes	1.6	8.6	10.4	15.1	16.2	0.6
GPT4-V [14]	2	no	4.0	18.6	**14.0**	20.3	22.2	2.6
GPT4-V[14]	2	yes	1.3	0.3	11.5	13.5	19.7	0.9
GPT4-V[14]	4	no	3.0	15.1	13.4	19.9	26.9	1.9
GPT4-V[14]	4	yes	1.4	0.4	11.6	12.9	24.1	1.2
ImProvShow-2 (ours)	2	no	5.8	20.7	11.4	23.1	22.6	9.4
ImProvShow-2T (ours)	2	yes	7.8	25.8	13.0	26.0	24.3	11.0
ImProvShow-4 (ours)	4	no	6.6	23.5	12.3	24.3	22.6	9.6
ImProvShow-4T (ours)	4	yes	**8.2**	**25.9**	13.4	**26.3**	**30.1**	**12.4**
			MagicBrush					
ImProvShow-2 (ours)	2	no	4.9	29.4	13.3	28.1	-	-
ImProvShow-4 (ours)	4	no	**6.8**	**44.5**	**15.6**	**31.2**	-	-

Performance evaluation shows BLEU-4 (**B4**), CIDEr (**C**), METEOR (**M**), ROUGE-L (**R**) and LLM as judge medium (**L (M)**) and high (**L (H)**) scores. See paper for more results.

Conclusion

- The **METS dataset** is released publicly and is available for download through this **QR code**!
- This work was supported by DECaDE under UKRI/EPSRC grant EP/T022485/1

datasets/AlexBlck/mets