VIXEN: Visual Text Comparison Network for Image Difference Captioning

Alexander Black[1], **Jing Shi**[2], **Yifei Fan**[2], **Tu Bui**[1], **John Collomosse**[1,2]

[1]CVSSP, University of Surrey
[2]Adobe Research
{alex.black|t.v.bui}@surrey.ac.uk, {jingshi|yifan|collomos}@adobe.com

Abstract

We present VIXEN – a technique that succinctly summarizes in text the visual differences between a pair of images in order to highlight any content manipulation present. Our proposed network linearly maps image features in a pairwise manner, constructing a soft prompt for a pretrained large language model. We address the challenge of low volume of training data and lack of manipulation variety in existing image difference captioning (IDC) datasets by training on synthetically manipulated images from the recent InstructPix2Pix dataset generated via prompt-to-prompt editing framework. We augment this dataset with change summaries produced via GPT-3. We show that VIXEN produces state-of-the-art, comprehensible difference captions for diverse image contents and edit types, offering a potential mitigation against misinformation disseminated via manipulated image content. Code and data are available at http://github.com/alexblck/vixen

Introduction

Image manipulation often forms the basis for fake news and misinformation. This threat may be countered by tools that encourage users to reflect upon the provenance and content of images. Given the reactionary nature of sharing, such tools should be intuitively comprehensible to enable users to make fast, informed trust decisions (Gregory 2019).

This paper contributes VIXEN – a method for intuitively summarizing the visual change between a pair of images using a short passage of text. Emerging open standards (*e.g.* C2PA (Coalition for Content Provenance and Authenticity 2023)) describe provenance frameworks that match images circulating in the wild to a federated database of originals using perceptual hashing methods (Black et al. 2021b; Nguyen et al. 2021; Black et al. 2021a; Pizzi et al. 2022). VIXEN presents a comprehensible way to review any image manipulation evidenced by such a matching (Fig. 1).

Image difference captioning (IDC) is typically addressed by representations that seek to model the spatial-semantic distribution of concepts present in a scene – for example, the relative positions of objects in CCTV footage (Jhamtani and Berg-Kirkpatrick 2018), or of primitive geometric shapes (Johnson et al. 2017). More complex kinds of manipulations require expertise to construct and thus can not be easily scaled up in volume (Tan et al. 2019). To this end, we make three technical contributions:

1. **Cross-modal image differencing.** We present a novel image differencing concept comprising a 2-branch GPT-J architecture to embed and compare facts derived from the image pair using CLIP-based image encoding. The model generates text conditioned on that comparison to explain salient changes between the image pair. We show our textual explanations to be succinct and comprehensible to non-expert users and to be quantitatively closer to ground-truth edit captions than state-of-the-art captioning methods.
2. **Synthetic Edit Training.** We propose a synthetic pairwise training framework for our VIXEN leveraging recent prompt2prompt (P2P) and language-based image editing (LBIE) approaches to supervise fine-tuning on generative image content, showing good generalization to unseen content.
3. **Augmented IP2P Dataset.** We release an augmentation of the recent InstructPix2Pix (IP2P) dataset with synthetic change captions generated via GPT-3 as a basis for training and evaluating VIXEN.

We demonstrate that VIXEN achieves higher performance than prior image difference captioning methods and is able to generalize to multiple datasets.

Related Work

Image difference captioning (IDC) is closely related to image captioning and visual question answering, both requiring a visual understanding system to model images and a language understanding system capable of generating syntactically correct captions. The revolution of IDC in recent years depends heavily on the advent of visual and text modeling approaches, together with cross-domain learning techniques that bridge the representation gap between them.

Early visual content modeling approaches employ global CNN features such as VGG (Donahue et al. 2015) and ResNet (Rennie et al. 2017) as input signals to the text generation models, leveraging the semantically rich and compact representations deliverable from these models. To better capture multi-object representations and their relation, regional modeling methods are developed (Lu et al. 2017; Gu et al. 2018; Anderson et al. 2018; Huang et al. 2019). In

"The image has been changed to a stormy night. There is now rain and lightning in Paris at sunset."

"The cat has been replaced with a tiger."

"The baby's face has been edited to look like a villainous character from an action movie."

Figure 1: Visual change summarization produced by VIXEN for original-manipulated image pairs. VIXEN is able to observe both background (left) and main subject (mid) changes as well as generalize to other datasets (right).

some, images are gridded into non-overlapping patches upon which CNN features are extracted, others instead use outputs from an early layer of a pretrained ResNet model to effectively capture spatial features in a grid fashion. In contrast, (Cornia et al. 2020; Anderson et al. 2018; Huang et al. 2019) employ Region Proposal Network (RPN) to extract features from potential candidate objects, offering better alignment with semantic objects mentioned in the paired captions. Alternative approaches include graph-based (Yang et al. 2019) and tree-based networks (Yao et al. 2019) to capture the relations of objects at multiple levels of granularity.

For a long time RNN/LSTM (Graves and Graves 2012) have been used to model text due to its inherent sequential properties. Single-layer RNN (Vinyals et al. 2015; Mao et al. 2015) or double-layer LSTM (Donahue et al. 2015; Anderson et al. 2018; Yao et al. 2019) are employed along with various techniques to integrate image features deeper into the recurrent process, including additive attention (Stefanini et al. 2022). During inference, captions are generated in an autoregressive fashion – the prediction of a word is conditioned on all previous words. While this improves linguistic coherence, RNN/LSTM-based approaches struggle in modeling long captions. This problem is levitated in recent transformer-based approaches thanks to its full-attention mechanism (Luo et al. 2021; Wang et al. 2021; Cornia et al. 2020). More advanced transformer-based approaches such as BERT (Devlin et al. 2018), GPT (Brown et al. 2020) and LLaMA (Touvron et al. 2023) have been successfully applied in various visual-language tasks (Hu et al. 2022; Mokady, Hertz, and Bermano 2021; Gao et al. 2023; Zhang et al. 2021; Li et al. 2020).

Visual language modeling aims to bridge the gap between image/video and text representations for specific tasks such as joint embedding (*e.g.* CLIP (Radford et al. 2021) and LIMoE (Mustafa et al. 2022) for cross-domain retrieval), text-to-image (*e.g.* Stable Diffusion (Rombach et al. 2022) for text-based image generation, InstructPix2Pix (Brooks, Holynski, and Efros 2022) for image editing) and image-to-text (*e.g.* visual question answering (Alayrac et al. 2022; Wang et al. 2021), visual instructions (Gao et al. 2023; Driess et al. 2023)). In the context of image captioning, image-text mapping strategies can be categorized into two research strands. The first strand involves the early fusion of image and text features for better alignment between image objects and words (Tsimpoukelli et al. 2021; Mokady, Hertz, and Bermano 2021; Wang et al. 2021; Li et al. 2020). These methods adopt BERT-like training strategies to input a pair of image and masked caption to the masked words. At inference, the input caption is simply replaced by a start token or a prefixed phrase *e.g.* 'A picture of'. The second research strand focuses on learning a direct transformation from image to text embedding. Early CNN-based approaches feed image features as the hidden states of the LSTM text modules (Donahue et al. 2015; Vinyals et al. 2015; Yao et al. 2019; Karpathy and Fei-Fei 2015; Rennie et al. 2017) while later transformer-based methods favor cross-attention (Luo et al. 2021; Cornia et al. 2020). Recently in both research strands, there has been a trend of leveraging powerful pretrained large language and vision models to learn a simple mapping between two domains (Merullo et al. 2022; Eichenberg et al. 2021; Li et al. 2023; Tsimpoukelli et al. 2021; Mokady, Hertz, and Bermano 2021).

Image difference captioning is a form of image captioning in which the caption would ideally ignore common objects between images and rather highlight subtle changes between them. As the first work addressing IDC, Spot-the-Diff (Jhamtani and Berg-Kirkpatrick 2018) identifies potential change clusters and models them using an LSTM-based network. Their work relies on the difference between two input images at the pixel level, therefore sensitive to noises and geometric transformations. DUDA (Park, Darrell, and Rohrbach 2019) instead computes image difference at CNN semantic level, improving the robustness against slight global changes. In M-VAM (Shi et al. 2020) and VACC (Kim et al. 2021), a view-point encoder is proposed to mitigate potential view-point difference and VARD (Tu et al. 2023a) proposes a viewpoint invariant representation network to explicitly capture the change. Meanwhile, (Sun et al. 2022) uses bidirectional encoding to improve change localization and NCT (Tu et al. 2023b) aggregates neighboring features with a transformer. These methods mostly focus on image modality and take advantage of benchmark-specific properties, such as near-identical views in Spot-the-Diff (Jhamtani and Berg-Kirkpatrick 2018) or synthetic scenes with limited objects and change types (color, texture, add, drop, remove) in CLEVR (Park, Darrell, and Rohrbach 2019). More recently, IDC-PCL (Yao, Wang, and Jin 2022) and CLIP4IDC (Guo, Wang, and Laaksonen 2022) adopt BERT-like training strategies to model difference captioning language, achieving state-of-art performance.

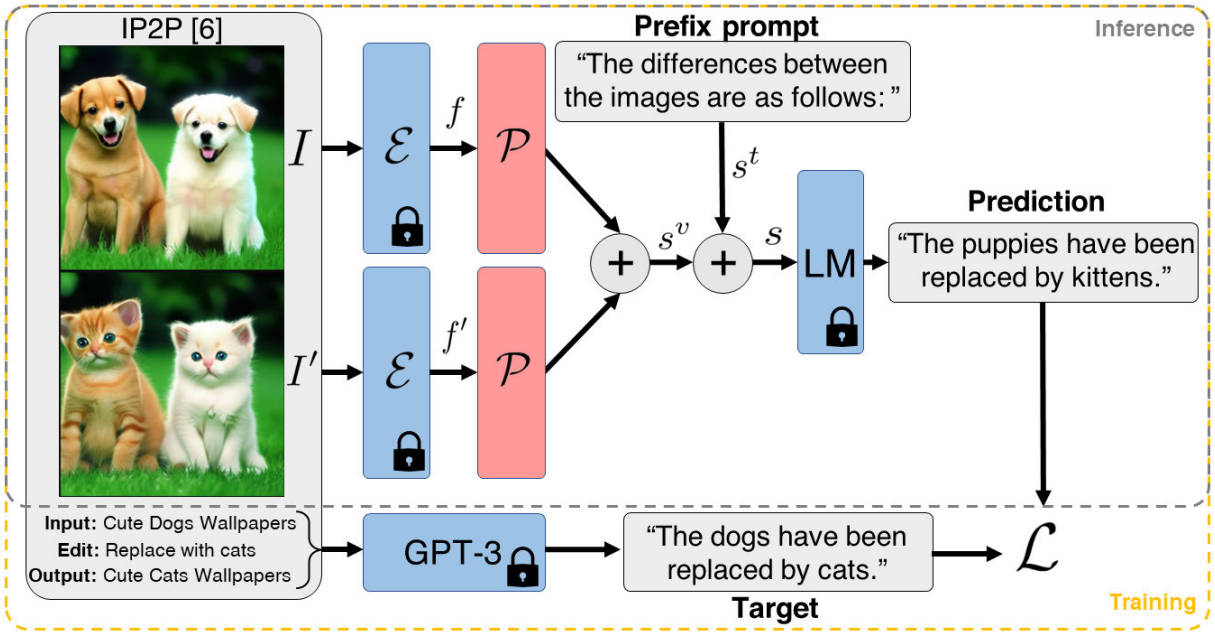


Figure 2: Model architecture and data captioning augmentation pipeline diagram. We use a pre-trained image encoder network $\mathcal{E}$ to produce a representation of two images. Both of these are projected into the input space of a large language model (LM) by a trained linear projection layer $\mathcal{P}$. Frozen layers are marked in blue, trainable in red.

Methodology

Our proposed method relies on synthetically generated image pairs and associated difference captions. We describe the creation process of visual and textual components of the dataset, details of the architecture of our proposed approach and training details necessary to reproduce the results.

Data Generation

To train our proposed approach, we require a large dataset of image pairs, each annotated with a summary of the changes between them. We propose using images generated by stable diffusion (Rombach et al. 2022) and edited with prompt-to-prompt (Hertz et al. 2022) using the pipeline presented in InstructPix2Pix (Brooks, Holynski, and Efros 2022) (IP2P). One of our contributions is the introduction of difference summary captions to IP2P images, generated using GPT-3 (Brown et al. 2020) in a few-shot learning fashion.

The InstructPix2Pix dataset is generated using the prompt-to-prompt editing framework, which provides text-based editing capabilities for synthesized images by injecting the attention maps associated with a specific word in the prompt to control the attention maps of the edited image. Therefore, all that is required to generate an image pair is two textual prompts with slight differences. IP2P uses a fine-tuned GPT-3 language model to generate plausible edits based on real input captions from LAION (Schuhmann et al. 2022). In addition to the image pairs and captions the dataset also contains an instruction that describes what edits have to be applied in order to generate the output image.

While these instructions are sufficient for the original InstructPix2Pix task of text-based image editing, they often omit the information regarding the input content. For example, for the prompt pair "a photo of a cat"/"a photo of a dog", the edit instruction might be "as a dog" or "turn it into a dog". We aim to summarize the changes by referencing both the original and edited image contents, therefore the desirable edit summarization caption would be "the cat has been replaced by a dog". To achieve this, we use GPT-3 language model in a few-shot learning fashion by including several examples of input-output-instruction-summary quadruplets where summary captions are constructed manually. While IP2P uses a fine-tuned GPT-3 to generate the instruction and second image captions, we have found the fine-tuning unnecessary in our case. Since our task does not require creativity from the model, but rather summarization of the input information, the pre-trained 'davinci' version of GPT-3 is enough to produce the captions needed.

Architecture

Our image captioning approach is inspired by (Merullo et al. 2022), which uses a trainable linear mapping between the image encoder and a large language model. However, instead of passing the projected embeddings of a single image to the language model, we project the embeddings of two images and concatenate them before feeding them into the language model. This architecture is illustrated in Figure 2. Given a source image I and its edited version I' we use an image encoder $\mathcal{E}$ to extract image feature maps

$$f = \mathcal{E}(I);\ f' = \mathcal{E}(I') \in \mathbb{R}^{k \times h}, \tag{1}$$

where h is the size of feature maps and k is the prompt sequence length. We use a fully-connected layer $\mathcal{P}$ to linearly

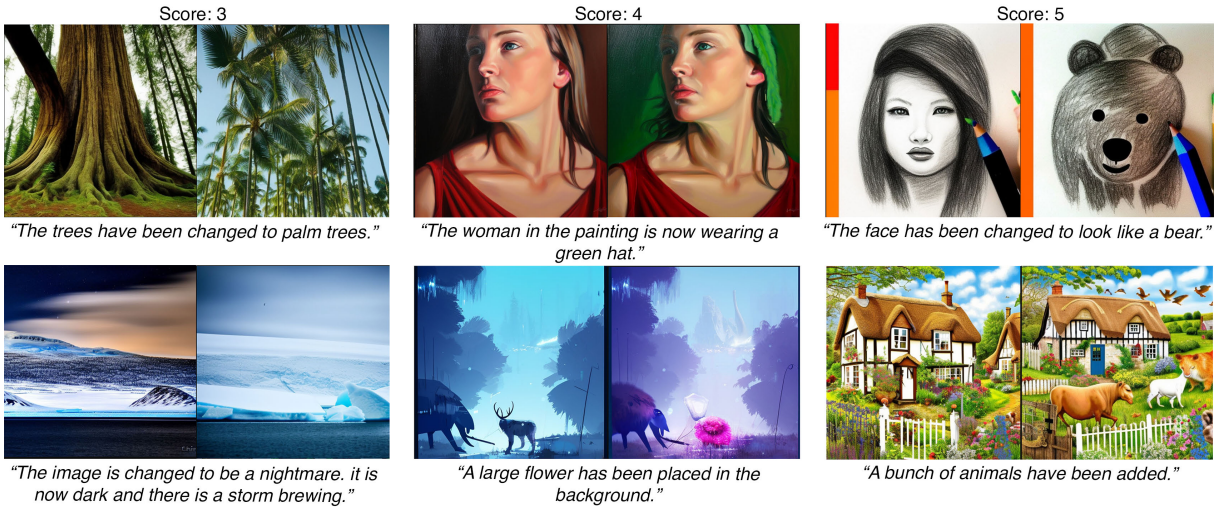


Figure 3: Image-caption pairs with an average correspondence score of 3 (left): may contain global changes when only local ones are expected (top) or fail to produce desired edits due to vague captioning (bot); 4 (mid): partially satisfy the caption, occasionally only some properties are realized correctly (top) or an existing object is replaced rather than added to the background (bot); 5 (right): mostly faithful to the depicted edits.

project the image features into dimensionality of a language model input e, creating a soft prompt s^v:

$$s^v = [\mathcal{P}(f), \mathcal{P}(f')] \in \mathbb{R}^{2k \times e}, \tag{2}$$

where $[,]$ denotes concatenation. Finally, we append a prefix s_t made of embedding of tokens for *"The differences between the images are as follows: "/"Edit instructions:"* to the visual prompt s^v to obtain the final prompt $s = [s^v, s^t]$ used for generating the summarization text.

We explore two options for $\mathcal{E}$. Firstly, following (Merullo et al. 2022; Eichenberg et al. 2021), we use CLIP RN50x16 as $\mathcal{E}$. The feature map before the pooling layer has dimensions $12 \times 12 \times 3072$, flattened to $k \times h = 144 \times 3072$. Secondly, we use ViT-g, followed by a Q-Former from BLIP-2 (Li et al. 2023). In this case sequence length $k = 257$. We refer to CLIP and Q-Former versions of VIXEN as **VIXEN-C** and **VIXEN-Q**, respectively. For the language model, we use GPT-J(Wang and Komatsuzaki 2021), which has input space dimensionality $l = 4096$. Consequently, for both configurations of $\mathcal{E}$, our linear projection layer $\mathcal{P}$ has input and output dimensions $h = 3072$ and $l = 4096$, respectively. The loss for the captioning task objective is defined as

$$\mathcal{L} = -\sum_{i=1}^{m} l(s^v, s_1^t, \ldots, s_i^t), \tag{3}$$

where m is a variable token length and l is next-token log-probability conditioned on the previous sequence elements

$$l(s^v, s_1^t, \ldots, s_i^t) = \log p(t_i | x, t_1, \ldots, t_{i-1}). \tag{4}$$

Training

During training, we may provide distractor image pairs with no changes present by providing the same image as both inputs $I = I'$. The frequency of the presence of distractor images is determined by probability p_d. In such cases, the target difference summary text is chosen at random from a list of pre-defined sentences, all synonymous with "there is no difference". For all our models we first train with $p_d = 0$ for two epochs, followed by two more epochs with $p_d = 0.5$. Total training time is approximately 100 hours on a single A100 GPU. We use gradient accumulation to train with an effective batch size of 2048 and optimize the loss using AdamW optimizer with $\beta_1 = 0.9, \beta_2 = 0.98$ and weight decay 0.05. For baseline approaches CLIP4IDC and IDC, we implemented dataloaders for our dataset, precomputed all necessary supporting data (e.g., ResNet-101 features, negative samples, and a vocabulary dictionary for IDC) and followed their standard two-step training pipeline with default hyperparameters specified in the GitHub repos.

Experiments

Data

We perform our main evaluation on a subset of the InstructPix2Pix (Brooks, Holynski, and Efros 2022) dataset, unseen by models during training. To ensure a high quality of the synthetically generated image-caption pairs, we score their correspondence via a user study. Additionally, we crowdsource annotations for a subset of images from the PSBattles (Heller, Rossetto, and Schuldt 2018) dataset and fine-tune and evaluate on Image Editing Request (Tan et al. 2019).

InstructPix2Pix dataset presents challenges due to its synthetically generated nature, as some of the edit summarization captions fail to accurately describe the changes made to the image pairs. This is mainly due to prompt-to-prompt occasionally generating images that do not depict the desired change accurately enough. This is further discussed in the limitations section below and illustrated in Figure 5 (mid). To ensure a reliable evaluation, we conducted a

Method	MPNet			B@4			C			M			R		
Instruct Pix2Pix															
	@3	@4	@5	@3	@4	@5	@3	@4	@5	@3	@4	@5	@3	@4	@5
VIXEN-Q (ours)	56.9	59.1	**62.3**	16.5	**18.5**	**20.8**	80.3	93.9	**134.9**	17.1	18.4	20.6	38.0	40.1	**42.2**
VIXEN-C (ours)	**59.3**	**61.4**	61.5	**16.8**	18.2	19.2	**96.6**	**107.0**	126.1	17.6	18.6	19.6	**39.2**	**40.8**	39.3
CLIP4IDC	56.8	58.3	60.7	15.8	17.3	17.7	58.8	71.0	114.7	**20.9**	**22.5**	**23.3**	33.3	35.1	34.0
IDC	38.3	38.6	37.4	8.2	8.8	7.7	4.4	5.0	5.6	16.0	16.8	16.5	29.1	30.0	27.7
PSBattles															
VIXEN-Q (ours)	**45.1**			**5.8**			7.5			**11.0**			**22.2**		
VIXEN-C (ours)	40.3			4.5			**7.7**			9.5			20.5		
CLIP4IDC	32.7			3.2			5.0			10.1			21.7		
IDC	27.0			1.0			0.7			9.2			19.5		
Image Editing Request															
VIXEN-Q (ours, FT)	50.1			7.9			35.4			14.4			33.5		
VIXEN-C (ours, FT)	**52.5**			8.6			**38.1**			**15.4**			**42.5**		
VARD	-			**10.0**			35.7			14.8			39.0		
CLIP4IDC	-			8.2			32.2			14.6			40.4		
NCT	-			8.1			34.2			15.0			38.8		
BiDiff	-			6.9			27.7			14.6			38.5		
DUDA	-			6.5			27.8			12.4			37.3		
rel-att	-			6.7			26.4			12.8			37.4		

Table 1: Image difference captioning performance on IP2P, PSBattles and Image Editing Request datasets. Evaluated on semantic similarity (MPNet), BLEU-4 (B@4), CIDEr (C), METEOR (M) and ROUGE-L (R). For IP2P, performance is reported for subsets at image-caption correspondence thresholds of 3, 4, 5.

Method	MPNet	B@4	C	M	R
Instruct Pix2Pix					
VIXEN-C (ours)	**59.3**	**16.8**	**96.6**	**17.6**	**39.2**
VIXEN-C p=0	54.4	15.4	88.5	16.1	35.9
PSBattles					
VIXEN-C (ours)	**40.3**	**4.5**	**7.7**	**9.5**	**20.5**
VIXEN-C p=0	37.8	4.2	7.2	8.9	19.2

Table 2: Impact of distractor images on performance of the model evaluated on semantic similarity (MPNet), BLEU-4 (B@4), CIDEr (C), METEOR (M) and ROUGE-L (R).

user study using Amazon Mechanical Turk (MTurk) on a sample of 5,000 images from the dataset. This results in a 837,466/93,052/5,000 train/validation/test splits. The study involved three participants per image-caption pair (95 unique participants) and aimed to rate the degree of correspondence between the image pair and its associated caption, using a scoring system from 1 (low) to 5 (high). The distribution of scores is 1: 5%, 2: 13%, 3: 26%, 4: 33%, 5: 24% . Figure 3 shows random samples of the image-caption pairs for different score threshold values.

PSBattles is a dataset of images edited in Adobe Photoshop™, collected from the 'Photoshopbattles' subreddit. The dataset contains 10k original images, paired with several manipulated variants. There are 102k variants in total contributed by 31k artists. We randomly sample 100 image pairs for crowd-sourced annotation on MTurk and collect captions from 3 participants per image pair.

Image Editing Request is a dataset of realistic photographs, paintings and illustrations paired with instructions written by humans. It contains 4k images-annotations pairs and incorporates a wide variety of edits, including affine edits and crops that are not present in the other datasets.

Metrics

We evaluate the performance of difference captioning methods using both traditional N-gram-based metrics (BLEU-4 (Papineni et al. 2002), CIDEr (Vedantam, Lawrence Zitnick, and Parikh 2015), METEOR (Banerjee and Lavie 2005) and ROUGE-L (Lin 2004)), as well as semantic similarity metric based on a language transformer model. We have found that due to a larger diversity of images and edits, the generated captions need to encompass a significantly larger vocabulary to accurately describe the changes. As a result, there are instances where the captions may not align word for word with the actual image differences, but they still convey a similar meaning. To account for this, we use a semantic textual similarity metric. We define semantic textual similarity S_{sim} between the target c and generated c' summarizations

$$S_{\text{sim}} = \cos(E(c), E(c')), \tag{5}$$

where $\cos(,) = \frac{\mathbf{A}\cdot\mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|}$ denotes cosine similarity and E is a sentence transformer. We use MPNet (Song et al. 2020) as the best-performing sentence transformer to map sentences to 768-dimensional normalized embeddings.

We also assess the quality of captions via a crowd-sourced study on Amazon Mechanical Turk (MTurk). Participants are presented with both the original and edited images. For each image pair, participants are tasked to choose one of the 4 captions, arranged in a random order. In case all four captions do not summarize the differences well enough, participants may choose the 'none of the above' option. Each task is performed by 3 unique participants. The preference is considered to be given to a particular method if two or more participants have voted for it.

	Background change	Subject replacement	Material changes
Target:	The café has been replaced by a library.	The baboons have been replaced by parrots.	The art was made into a statue. It is now three dimensional and made of stone or metal.
VIXEN-Q:	The location has been changed from a cafe to an old library.	The three gorillas have been turned into a trio of cockatoos.	The image has been changed from a painting to an sculpture.
VIXEN-C:	The location of where they meet has been changed from an unspecified place to a bookstore.	The baboons have been replaced by parrots.	The image has been changed from a digital painting to an oil on canvas.
CLIP4IDC[19]:	The background has been changed from a plain white to a brick wall.	The mother has been replaced by a kangaroo.	The image has been changed from a painting to a statue.

(a) InstructPix2Pix

Target:	A shark has been added on the right.	The jumping dog has been given a cowboy hat and a copy has been placed on the dog like they are riding a horse.	An animal has been added in the distance, which has a cartoonish oversized horse's head on a four legged body with a tail.
VIXEN-Q:	The image has been edited to make it look like a shark attack.	The dog is now wearing a cowboy hat.	The eagle has been replaced by a lion.
VIXEN-C:	The image has been changed to include a shark in addition.	The dog has been turned into a cowboy. He is now wearing jeans, boots and an old hat with feathers on it.	The golden eagles have been replaced by a pack of dogs.
CLIP4IDC[19]:	The image has been made into a movie poster. the title and credits have been added, as well as a tagline. the background has been made darker	The girl in the image is now wearing a crown.	The coyote has been replaced by a cat.

(b) PSBattles

Figure 4: Examples of edit summarizations for global changes, object replacement and material changes produced by VIXEN and CLIP4IDC on InstructPix2Pix (a) and PSBattles (b) datasets. Failure case marked with a dashed red box.

Results

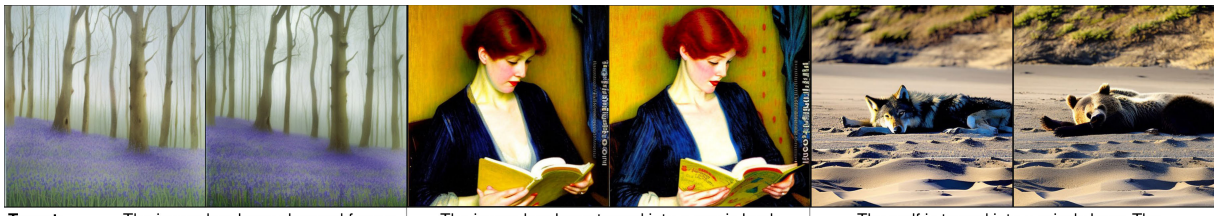
For the proposed datasets, we compare the performance of VIXEN against two baselines, IDC (Yao, Wang, and Jin 2022) and CLIP4IDC (Guo, Wang, and Laaksonen 2022). We train both of them on our augmented IP2P dataset, following the author's guidelines. For the IER dataset, we finetune on IER training set and compare against reported numbers of multiple baselines.

We report the evaluation results of evaluating both the proposed method as well as baselines in Table 1, with examples shown in Figure 4. Our methods achieve a higher score in all metrics, except METEOR (IP2P), where CLIP4IDC scores higher than both proposed architectures. This indicates that VIXEN is more tuned towards precision, rather than recall of n-grams as METEOR heavily favors recall. For IP2P, results are reported at three different correspondence thresholds. For lower threshold values, the best results are obtained by VIXEN-C. VIXEN-Q seems to benefit the most from threshold increase and outperforms other methods on pairs with a correspondence score of 5.

While all methods suffer significant performance drops when evaluated on a dataset from a different domain, VIXEN-Q shows a better ability to generalize to new data by scoring the highest on the PSBattles dataset. After finetuning the model on Image Editing Request, VIXEN-C outperforms previous methods on most metrics, except B@4 of VARD(Tu et al. 2023a).

The results of the crowd-sourced user preference study, shown in Figure 6, demonstrate that the users prefer difference captions generated by VIXEN more often than others. For the IP2P dataset, captions generated by VIXEN-Q and VIXEN-C obtained a majority vote in 32% and 26% of the cases, respectively, followed by CLIP4IDC and IDC with 24% and 15%. For the PSBattles dataset, the highest preference score is achieved by VIXEN-C with 15% of the votes. Participants chose the 'None of the above' option in 75% of the cases for PSBattles, as opposed to just 2% in IP2P. This indicates that generalization to new data domains remains a challenging task.

During inference we assume an input where one image is



Target:	The image has been changed from a painting to a photograph.	The image has been turned into a comic book.	The wolf is turned into a grizzly bear. The colors are changed from grey to brown, and the fur is made thicker.
VIXEN-Q:	The image has been changed to a foggy morning.	The woman has been turned into a cartoon character. She now looks like she is from an animated movie, with big eyes and exaggerated features.	The bear has been turned into a wolf.
VIXEN-C:	The bluebells have been changed to look like they were in a forest.	The woman has been turned into a cartoon character. She now wears glasses, and her hair is in pigtails with bows at each end of it.	The wolf has been replaced by a bear.

Figure 5: Limitations of the proposed method. Left: image captioning instead of difference captioning in case of unidentified edit. Middle: mismatch between target text-image pair and LM runoff. Right: edit described in reverse order.

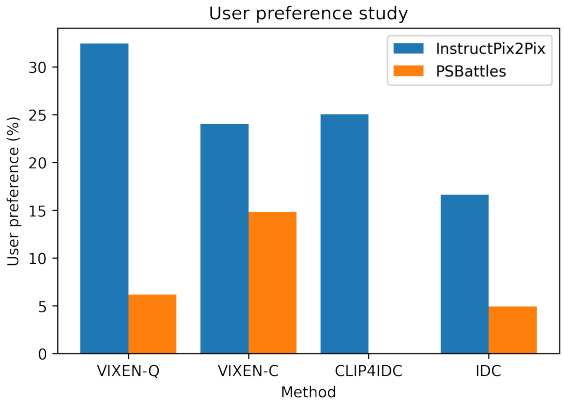


Figure 6: User preference study results. Study participants are shown an image pair and captions generated by four methods on IP2P and PSBattles datasets.

Fusion method	B@4	C	M	R
Instruct Pix2Pix				
Concatenation	**16.8**	**96.6**	**17.6**	**39.2**
Subtraction	16.4	93.7	16.9	36.8
Addition	16.2	90.8	17.3	37.5
Multiplication	14.4	82.7	15.3	35.9
Mean	10.7	63.9	12.4	33.6

Table 3: Image feature fusion ablation of VIXEN-C

an edited version of the other, but we demonstrate the benefits of having distractor same image pairs during training. The possibility of no edits case makes it harder for the model to guess the right answer by memorizing the most frequent edits within the dataset. Table 2 shows that setting the probability $p = 0$ of the same image pairs during training of VIXEN-C yields worse results on both IP2P and PSBattles datasets.

Table 3 shows performance results for different feature fusion strategies that redefine s^v in Eq 2. We have observed that concatenation leads to slightly better performance than subtraction, addition or multiplication and taking the mean of two features causes a significant performance drop. This shows that retaining the information of both image features without degradation is important for the task.

Limitations

In Figure 5 we show examples of VIXEN's failure cases. We identify and discuss three main challenges. **Left** shows an example of a very minor difference between the two images. In such cases, VIXEN occasionally resorts to captioning the image content instead of summarizing the differences. **Mid** shows a mismatch between the summary and generated images: an image pair with a slightly changed book cover, but the target caption assumes that the style of the whole image has been changed to that of a comic book. As with other LLMs, VIXEN exhibits LM runoff: having identified a concept ("cartoon character"), it might continue generating a text with a strong linguistic prior ("big eyes and exaggerated features"), absent in the images. **Right** shows that occasionally VIXEN may describe the differences between the images in a reversed order.

Conclusion

We presented VIXEN – an image difference captioning approach that provides textual descriptions of the manipulations applied to an image. We have augmented the InstructPix2Pix dataset of generated images with difference summarization captions generated by GPT-3 in order to train and evaluate VIXEN. We have shown that VIXEN achieves higher performance than other image difference captioning methods. We have also demonstrated that, while VIXEN shows better generalizability to other datasets, there is still a performance gap when switching from synthetic to real data. Future works might alleviate this by including a varied spread of manipulations types into the training set, including insertion, deletion and text edits, which the current generative pipelines struggle with.

References

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a visual language model for few-shot learning. *NeurIPS*, 35: 23716–23736.

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proc. CVPR*, 6077–6086.

Banerjee, S.; and Lavie, A. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proc. ACL WS Intr. Extr. Eval. Measures Machine Trans. Summarization*, 65–72.

Black, A.; Bui, T.; Jenni, S.; Swaminathan, V.; and Collomosse, J. 2021a. Vpn: Video provenance network for robust content attribution. In *Proc. CVMP*, 1–10.

Black, A.; Bui, T.; Jin, H.; Swaminathan, V.; and Collomosse, J. 2021b. Deep Image Comparator: Learning to Visualize Editorial Change. In *Proc. CVPR WS*, 972–980. IEEE.

Brooks, T.; Holynski, A.; and Efros, A. A. 2022. Instructpix2pix: Learning to follow image editing instructions. *arXiv preprint arXiv:2211.09800*.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *NeurIPS*, 33: 1877–1901.

Coalition for Content Provenance and Authenticity. 2023. Technical Specification 1.3. Technical report, C2PA.

Cornia, M.; Stefanini, M.; Baraldi, L.; and Cucchiara, R. 2020. Meshed-memory transformer for image captioning. In *Proc. CVPR*, 10578–10587.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Donahue, J.; Anne Hendricks, L.; Guadarrama, S.; Rohrbach, M.; Venugopalan, S.; Saenko, K.; and Darrell, T. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *Proc. CVPR*, 2625–2634.

Driess, D.; Xia, F.; Sajjadi, M. S. M.; Lynch, C.; Chowdhery, A.; Ichter, B.; Wahid, A.; Tompson, J.; Vuong, Q.; Yu, T.; Huang, W.; Chebotar, Y.; Sermanet, P.; Duckworth, D.; Levine, S.; Vanhoucke, V.; Hausman, K.; Toussaint, M.; Greff, K.; Zeng, A.; Mordatch, I.; and Florence, P. 2023. PaLM-E: An Embodied Multimodal Language Model. In *arXiv preprint arXiv:2303.03378*.

Eichenberg, C.; Black, S.; Weinbach, S.; Parcalabescu, L.; and Frank, A. 2021. MAGMA–Multimodal Augmentation of Generative Models through Adapter-based Finetuning. *arXiv preprint arXiv:2112.05253*.

Gao, P.; Han, J.; Zhang, R.; Lin, Z.; Geng, S.; Zhou, A.; Zhang, W.; Lu, P.; He, C.; Yue, X.; et al. 2023. LLaMA-Adapter V2: Parameter-Efficient Visual Instruction Model. *arXiv preprint arXiv:2304.15010*.

Graves, A.; and Graves, A. 2012. Long short-term memory. *Supervised sequence labelling with recurrent neural networks*, 37–45.

Gregory, S. 2019. Ticks or it didn't happen. Technical report, Witness.org.

Gu, J.; Cai, J.; Wang, G.; and Chen, T. 2018. Stack-captioning: Coarse-to-fine learning for image captioning. In *Proc. AAAI*, volume 32.

Guo, Z.; Wang, T.-J.; and Laaksonen, J. 2022. CLIP4IDC: CLIP for Image Difference Captioning. In *Proc. Conf. Asia-Pacific Chapter Assoc. Comp. Linguistics and Int. Joint Conf. NLP*, 33–42.

Heller, S.; Rossetto, L.; and Schuldt, H. 2018. The PS-Battles Dataset – an Image Collection for Image Manipulation Detection. *CoRR*, abs/1804.04866.

Hertz, A.; Mokady, R.; Tenenbaum, J.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2022. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*.

Hu, X.; Gan, Z.; Wang, J.; Yang, Z.; Liu, Z.; Lu, Y.; and Wang, L. 2022. Scaling up vision-language pre-training for image captioning. In *Proc. CVPR*, 17980–17989.

Huang, L.; Wang, W.; Xia, Y.; and Chen, J. 2019. Adaptively aligned image captioning via adaptive attention time. *NeurIPS*, 32.

Jhamtani, H.; and Berg-Kirkpatrick, T. 2018. Learning to Describe Differences Between Pairs of Similar Images. In *Proc. Conf. Empirical Methods NLP*, 4024–4034.

Johnson, J.; Hariharan, B.; Van Der Maaten, L.; Fei-Fei, L.; Lawrence Zitnick, C.; and Girshick, R. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proc. CVPR*, 2901–2910.

Karpathy, A.; and Fei-Fei, L. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proc. CVPR*, 3128–3137.

Kim, H.; Kim, J.; Lee, H.; Park, H.; and Kim, G. 2021. Agnostic change captioning with cycle consistency. In *Proc. ICCV*, 2095–2104.

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 preprint arXiv:2301.12597*.

Li, X.; Yin, X.; Li, C.; Zhang, P.; Hu, X.; Zhang, L.; Wang, L.; Hu, H.; Dong, L.; Wei, F.; et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Proc. ECCV*, 121–137. Springer.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.

Lu, J.; Xiong, C.; Parikh, D.; and Socher, R. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proc. CVPR*, 375–383.

Luo, Y.; Ji, J.; Sun, X.; Cao, L.; Wu, Y.; Huang, F.; Lin, C.-W.; and Ji, R. 2021. Dual-level collaborative transformer for image captioning. In *Proc. AAAI*, volume 35, 2286–2293.

Mao, J.; Xu, W.; Yang, Y.; Wang, J.; Huang, Z.; and Yuille, A. 2015. Deep captioning with multimodal recurrent neural networks (m-rnn). In *Proc. ICLR*.

Merullo, J.; Castricato, L.; Eickhoff, C.; and Pavlick, E. 2022. Linearly mapping from image to text space. *arXiv preprint arXiv:2209.15162*.

Mokady, R.; Hertz, A.; and Bermano, A. H. 2021. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*.

Mustafa, B.; Ruiz, C. R.; Puigcerver, J.; Jenatton, R.; and Houlsby, N. 2022. Multimodal Contrastive Learning with LIMoE: the Language-Image Mixture of Experts. In *NeurIPS*.

Nguyen, E.; Bui, T.; Swaminathan, V.; and Collomosse, J. 2021. Oscar-net: Object-centric scene graph attention for image attribution. In *Proc. ICCV*, 14499–14508.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proc. Assoc. Comp. Linguistics*, 311–318.

Park, D. H.; Darrell, T.; and Rohrbach, A. 2019. Robust change captioning. In *Proc. ICCV*, 4624–4633.

Pizzi, E.; Roy, S. D.; Ravindra, S. N.; Goyal, P.; and Douze, M. 2022. A self-supervised descriptor for image copy detection. In *Proc. CVPR*, 14532–14542.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *Proc. ICML*, 8748–8763. PMLR.

Rennie, S. J.; Marcheret, E.; Mroueh, Y.; Ross, J.; and Goel, V. 2017. Self-critical sequence training for image captioning. In *Proc. CVPR*, 7008–7024.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proc. CVPR*, 10684–10695.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; Schramowski, P.; Kundurthy, S.; Crowson, K.; Schmidt, L.; Kaczmarczyk, R.; and Jitsev, J. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. In *NeurIPS*, volume 35, 25278–25294.

Shi, X.; Yang, X.; Gu, J.; Joty, S.; and Cai, J. 2020. Finding it at another side: A viewpoint-adapted matching encoder for change captioning. In *Proc. ECCV*, 574–590. Springer.

Song, K.; Tan, X.; Qin, T.; Lu, J.; and Liu, T.-Y. 2020. Mpnet: Masked and permuted pre-training for language understanding. *NeurIPS*, 33: 16857–16867.

Stefanini, M.; Cornia, M.; Baraldi, L.; Cascianelli, S.; Fiameni, G.; and Cucchiara, R. 2022. From show to tell: A survey on deep learning-based image captioning. *IEEE TPAMI*, 45(1): 539–559.

Sun, Y.; Li, L.; Yao, T.; Lu, T.; Zheng, B.; Yan, C.; Zhang, H.; Bao, Y.; Ding, G.; and Slabaugh, G. 2022. Bidirectional difference locating and semantic consistency reasoning for change captioning. *IJIS*, 37(5): 2969–2987.

Tan, H.; Dernoncourt, F.; Lin, Z.; Bui, T.; and Bansal, M. 2019. Expressing Visual Relationships via Language. arXiv:1906.07689.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Tsimpoukelli, M.; Menick, J. L.; Cabi, S.; Eslami, S.; Vinyals, O.; and Hill, F. 2021. Multimodal few-shot learning with frozen language models. *NeurIPS*, 34: 200–212.

Tu, Y.; Li, L.; Su, L.; Du, J.; Lu, K.; and Huang, Q. 2023a. Viewpoint-Adaptive Representation Disentanglement Network for Change Captioning. *IEEE Transactions on Image Processing*, 32: 2620–2635.

Tu, Y.; Li, L.; Su, L.; Lu, K.; and Huang, Q. 2023b. Neighborhood Contrastive Transformer for Change Captioning. arXiv:2303.03171.

Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *Proc. CVPR*, 4566–4575.

Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2015. Show and tell: A neural image caption generator. In *Proc. CVPR*, 3156–3164.

Wang, B.; and Komatsuzaki, A. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. https://github.com/kingoflolz/mesh-transformer-jax.

Wang, Z.; Yu, J.; Yu, A. W.; Dai, Z.; Tsvetkov, Y.; and Cao, Y. 2021. Simvlm: Simple visual language model pretraining with weak supervision. In *Proc. ICLR*.

Yang, X.; Tang, K.; Zhang, H.; and Cai, J. 2019. Auto-encoding scene graphs for image captioning. In *Proc. CVPR*, 10685–10694.

Yao, L.; Wang, W.; and Jin, Q. 2022. Image difference captioning with pre-training and contrastive learning. In *Proc. AAAI*, volume 36, 3108–3116.

Yao, T.; Pan, Y.; Li, Y.; and Mei, T. 2019. Hierarchy parsing for image captioning. In *Proc. ICCV*, 2621–2629.

Zhang, P.; Li, X.; Hu, X.; Yang, J.; Zhang, L.; Wang, L.; Choi, Y.; and Gao, J. 2021. Vinvl: Revisiting visual representations in vision-language models. In *Proc. CVPR*, 5579–5588.