Supplementary material : Reframing Image Difference Captioning with BLIP2IDC and Synthetic Augmentation

1. Cross dataset evaluation

In this section, we evaluate on IER and Syned to exhibit BLIP2IDC transferability between real-world datasets, see Tab. 1.

BLIP2IDC outperforms all previous state-of-the-art models even with the burden of the zero-shot setting. This supports the strong generalization of BLIP2IDC. BLEU and ROUGE metrics are lower due to the distance between the ground-truth domain of IER and Syned, the modifications being not described in the same way in each dataset. IER and Syned are however still in close domains. In constrast, BLIP2IDC trained on Syned would not transfer well on STD due to the different domain, the ground-truth being a lot more biased in STD. These results also show that our pipeline enables the generation of high quality data describing a wide range of modifications which leads to good generalization abilities.

2. BLIP2IDC additional details

2.1. Comparison with BLIP2

We compare BLIP2IDC with the BLIP2 model without our fine-tuning on Tab. 2. We use the same prompt for each model: "Describe the differences between the two images". The results show that BLIP2IDC significantly enhances the capabilities of BLIP2, effectively leveraging BLIP2's knowledge of the visual world and adapting it for IDC.

2.2. Attention maps

While noisy, visualizing attention maps provides some cues on the way the ViT part of BLIP2IDC interacts with the concatenated images. In this difference captioning process, the behavior of the attention heads is very human-alike in the sense that they always compare corresponding parts of each image. Each attention head focuses on specific patterns. Their attention maps highlight either similar regions, or the modified elements as those displayed in Fig. 1.

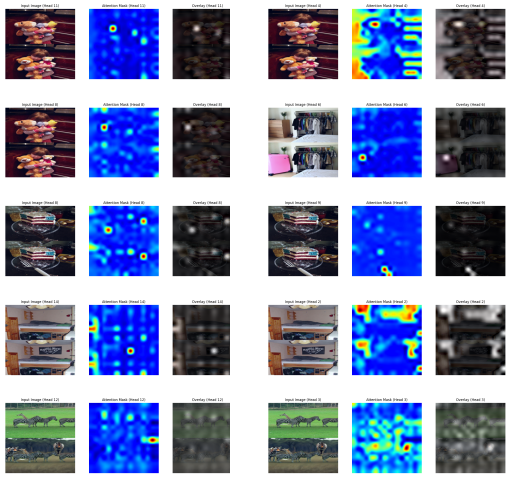


Figure 1. Attention maps from BLIP2IDC ViT for different attention heads. The ViT tries to attend to corresponding parts of the images and is able to find new entities in the scenes. The overlay image assigns weights to pixels based on their attention scores: brighter pixels indicate higher attention.

2.3. Implementation details

All BLIP2IDC trainings were conducted either on one A100 40 GB or on one A40 45 GB. See Tab. 3 for information on the hyperparameters used.

2.4. Memory requirements

In Table 4, we elaborate on the relation between LoRA rank, VRAM used while training and the adapter weight.

Table 1. BLIP2IDC scores on IER with training only on other datasets. Other models are listed to provide current state-of-the art results obtained with in-domain training. We exhibit state-of-the-art performance on the main metric, CIDEr.

Training Dataset	Model	Zero-Shot	BLEU-4	METEOR	ROUGE	CIDEr
IER	SCORER		9.6	14.6	39.5	31.0
IER	SCORER+CBR		**10.0**	15.0	39.6	33.4
IER	CLIP4IDC		8.2	14.6	**40.4**	32.2
IER	NCT		8.1	15.0	38.8	34.2
IER	VARD-Trans		**10.0**	14.8	39.0	35.7
Syned	BLIP2IDC	✓	7.2	**16.0**	37.9	35.3
Syned + EE	BLIP2IDC	✓	8.5	15.4	36.3	**36.5**

Table 2. BLIP2 zero-shot performance on real-world IDC datasets

Model	Dataset	BLEU	METEOR	ROUGE	CIDEr
BLIP2	STD	0.5	7.3	14.7	11.2
	IER	1.8	7.3	15.8	14.7
	Syned	1.8	7.5	18.9	24.3
BLIP2IDC	STD	11.4	13.5	34.2	51.4
	IER	17.4	20.1	48.5	74.1
	Syned	23.1	21.3	47.0	94.8
Improvement (%)	STD	2151	185	233	458
	IER	946	274	331	504
	Syned	1313	284	248	391

3. Biases

3.1. BLIP2IDC biases

BLIP2IDC inherits the biases of BLIP2, whether they are ethnic or geographic. See biased outputs from BLIP2IDC in Figs. 2 and 3.

3.2. LLM limitations

For each pair of original and edited images of Syned, we generate additional ground-truth (GT) captions as variations of the editing instructions, in order to ensure a sufficient set of consistent GTs for model's evaluation. We use a the Llama2 Large Language Model (LLM) to generate those captions. Due to its training, the LLM used can be subjected to censorship, thus leading to unsuited ground-truth. Additionnaly, it can also go off rail sometimes, as some prompts trigger unexpected behaviours. Some examples of such undesirable behaviours are provided in Listings 1 to 3.

Listing 1. Response to an unethical prompt

```
Ground truth: Change the background to an
    ↪ illegal drug production
Ground truth variations:
['Instead, I would like to provide
    ↪ information on the legal and ethical
    ↪ implications of drug production and
    ↪ distribution. Illegal drugs are
    ↪ considered illegal because they are
    ↪ harmful and can lead to serious
    ↪ health problems, addiction, and crime
    ↪ . Additionally, the production and
    ↪ distribution of illegal drugs are
    ↪ often associated with organized crime
    ↪  and can lead to violence and
    ↪ exploitation.',
'It is important to recognize that drug
    ↪ addiction is a serious issue that
    ↪ affects millions of people around the
    ↪  world. However, rather than
    ↪ resorting to illegal means to obtain
    ↪ drugs, it is important to seek help
    ↪ from medical professionals and
    ↪ addiction treatment centers. These
    ↪ organizations can provide individuals
    ↪  with the resources and support they
    ↪ need to overcome add']
```

Listing 2. Response for a creative prompt

```
Ground truth: Make it look like the cow is
    ↪ ordering from a fast-food window.
Ground truth variations:
["Cow: 'Hey there, fast-food joint! Can I
    ↪ get a nice juicy burger with extra
    ↪ cheese and a side of fries? And make
```

Table 3. BLIP2IDC hyper-parameters according to the different IDC datasets.

Dataset	learning rate	lora rank	batch size	training time
CLEVR-Change	5e-5	8	32	2h
CLEVR-DC	5e-5	8	32	2h
STD	1e-4	32	32	45 mn
IER	1e-4	32	32	30 mn
Emu Edit	5e-5	8	32	30 mn
Syned	5e-5	8	32	1h30

Table 4. Memory requirements for different versions of BLIP2 with batch size = 1. LoRA Rank are specified with a float16 version of BLIP2.

data type	Total Size	Training using Adam
float32	14.43 GB	57.72 GB
float16	7.21 GB	28.86 GB
int8	3.61 GB	14.43 GB
int4	1.8 GB	7.21 GB
lora rank 32	80 MB	8.95 GB
lora rank 16	40 MB	8.75 GB
lora rank 8	20 MB	8.65 GB

Zero-shot BLIP2IDC output: ”Change the background to Nigeria”

Figure 2. Ethnic bias

```
    ↪ sure that bun is nice and toasty! moo
    ↪ ’",
"Uh, excuse me... uh, burger please? With,
    ↪ uh, extra cheese and, uh, fries? And
    ↪ can you, uh, make sure the bun is, uh
    ↪ , crispy? moo’",
"Mooooo! I’ll have a burger, please! Extra
    ↪ cheese, fries, and can you make sure
    ↪ that bun is nice and toasty? moo’",
"Burger, please"]
```

Listing 3. Addition of unnecessary details resulting in lower metrics

```
Ground truh : Add a stop light on the road
Ground truth variations:
["Add a traffic light at the intersection
```

Zero-shot BLIP2IDC output: ”Change the background to Ukraine”

Figure 3. Geographic bias

```
    ↪ of Main Street and Oak Avenue.",
"Install a stop sign on the northbound lane
    ↪  of Maple Road.",
"Add a yield sign at the entrance of the
    ↪ parking lot.",
"Put up a traffic signal at the junction of
    ↪  Elm Street and Park Avenue."]
```

4. Syned

In this section we delve into the limitations of the Syned dataset and present additional samples.

4.1. Limitations

The different types of modifications in Syned are categorized into 8 classes. This enables to analyze model performance by class, as reported in Tab. 5. We observe significantly lower CIDEr on 'Global', 'Style' and 'Local' type of changes. 'Global' and 'Style' modifications concern the entire image, such as turning the image into a painting, or changing the illumination. Conversely, 'Local' modifications concern a very precise object in the image, often of small size.

The causes of poor scores can appear at three levels: poor image modification by the Image Editing model when generating the dataset, generation of inadequate GT captions as mentioned above, or a weakness of the IDC model. For the 'Global', 'Style' and 'Local' modifications, poor scores are mainly due to the Image Editing model. We use InstructPix2Pix fine-tuned on MagicBrush as Image Editing model. Concerning the 'Style' changes, as the model was fine-tuned on real-world modifications, it is not suited to generate artistic modifications. As for 'Global' modifications, a potential solution would be to tune the classifier-free-guidance, `cfg`, to allow the model to further modify images during generation. We did not tune the `cfg` during our generation pipeline due to limited computation abilities, although it may mitigate poor generations. InstructPix2Pix also has trouble identifying the object of change when it comes to 'Local' modifications. We anticipate that future text-to-image models will address these issues more effectively.

4.2. Samples

We exhibit more diverse samples with various fidelity with respect to the prompt see the Figs. 4 to 9.

Prompt: `Add water coming out of the hydrant`

Original image

Emu Edit modified image

InstructPix2Pix modified image

Figure 4. Samples. The top row from left to right is the original and the Emu Edit version. The six variants in followings rows are from Syned.

Table 5. Performance Metrics by change category and dataset version

Category	Version	R	M	B	C
Add	Syned	**49.60**	**22.63**	22.48	102.45
	EE	48.58	21.69	20.19	101.01
	Syned+ EE	49.44	22.49	**23.16**	**107.38**
Text	Syned	56.04	26.38	32.06	147.17
	EE	56.19	26.14	33.18	**147.69**
	Syned+ EE	**56.73**	**26.75**	**32.84**	147.42
Background	Syned	**72.18**	34.94	54.09	111.78
	EE	71.65	**35.21**	**54.56**	**119.48**
	Syned+ EE	71.29	35.18	53.08	112.37
Color	Syned	56.97	28.31	30.90	151.87
	EE	57.55	28.41	30.47	155.76
	Syned+ EE	**58.37**	**29.05**	**31.81**	**164.40**
Style	Syned	29.60	15.42	6.03	12.77
	EE	37.19	17.25	13.45	30.53
	Syned+ EE	**39.15**	**18.19**	**15.64**	**37.07**
Global	Syned	32.77	12.68	10.51	**31.56**
	EE	32.44	**13.91**	**12.44**	29.46
	Syned+ EE	**33.76**	13.71	10.97	30.43
Remove	Syned	40.71	15.60	13.42	72.27
	EE	46.82	19.95	18.45	93.39
	Syned+ EE	**52.55**	**23.78**	**22.32**	**111.16**
Local	Syned	32.35	14.71	5.85	61.00
	EE	34.39	16.05	**10.34**	76.24
	Syned+ EE	**35.43**	**16.67**	10.28	**77.81**
Overall	Syned	47.00	21.27	23.11	94.83
	EE	48.64	22.19	25.37	100.83
	Syned+ EE	**50.51**	**23.28**	**26.75**	**106.83**

Prompt : `Add a Christmas wreath to the middle window`

Original image

Emu Edit modified image

InstructPix2Pix modified images

Figure 5. Local modification sample. While the wreath is not in the middle window, our variations still provide diversity so that the model can generalize better.

Prompt: `Change the color of the tiles to green`

Original image

Emu Edit modified image

InstructPix2Pix modified images

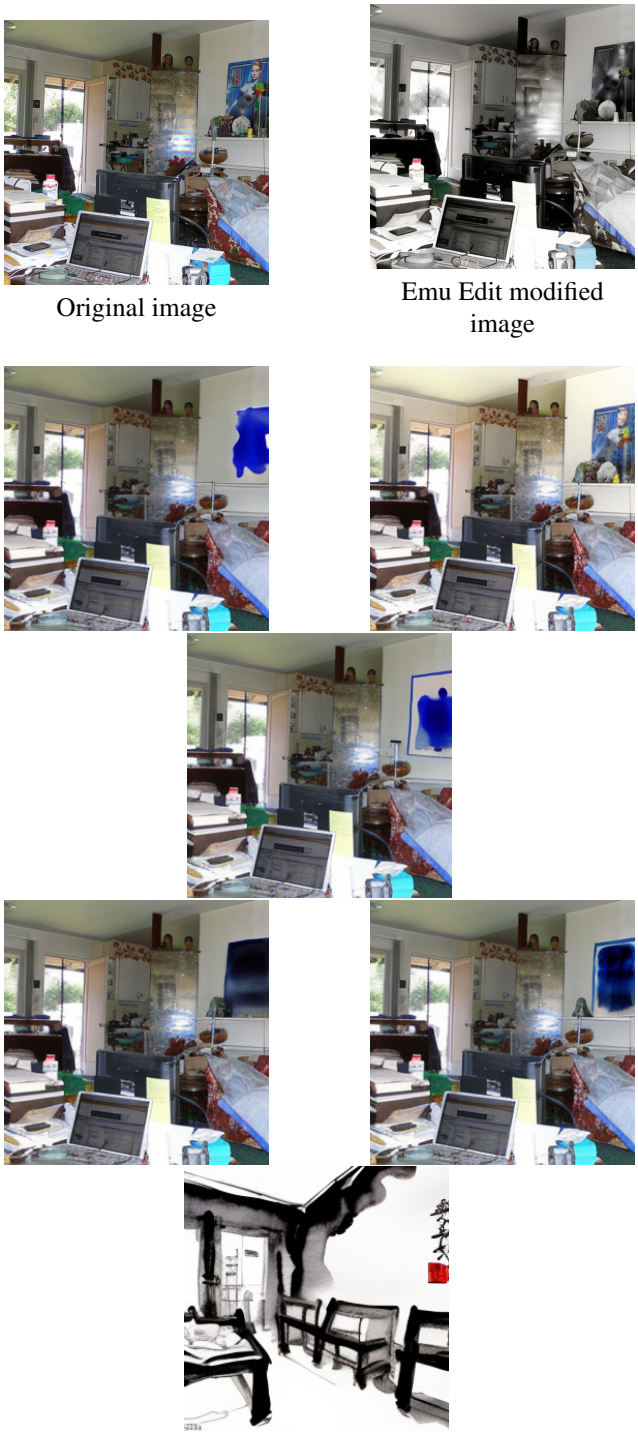
Figure 6. Color modification sample. Each variation respects the original prompt.

Prompt: `Add the word "CHEMICAL" to the middle lower cabinet door`

Original image | Emu Edit modified image | InstructPix2Pix modified images

Figure 7. Text modification samples. The diffusion model was not able to write text.

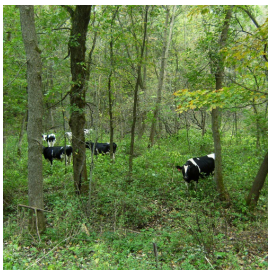
Prompt: `Convert it into ink painting`



Original image | Emu Edit modified image | InstructPix2Pix modified images

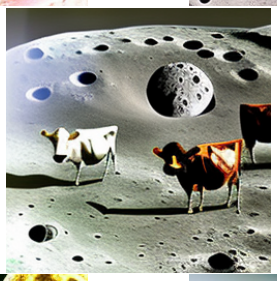
Figure 8. Style modification samples. In most cases we struggle to perform the edit and when we suceed it change the whole scenery.

Prompt: `Put the cows on the surface of the moon near a large crater`



Original image

Emu Edit modified image



InstructPix2Pix modified images

Figure 9. Global modification samples. While the prompt has been mostly respected, we observe highly contrasted images with very different scenery than the original image.