KU LEUVEN

Text-based Control for Image Manipulation

Maria Trusca

Faculty of Engineering Science

Department of Computer Science

Language Intelligence and Information Retrieval

European Research Council
Established by the European Commission

Introduction

- Aim: develop a method that capture relational structure of language and the geometric aspects of visual data to perform image editing.
- Text-based semantic image editing
  - *Current limitations*:
    - The background is usually altered;
    - The long and elaborate prompts are difficult to implement as instructions for image editing.

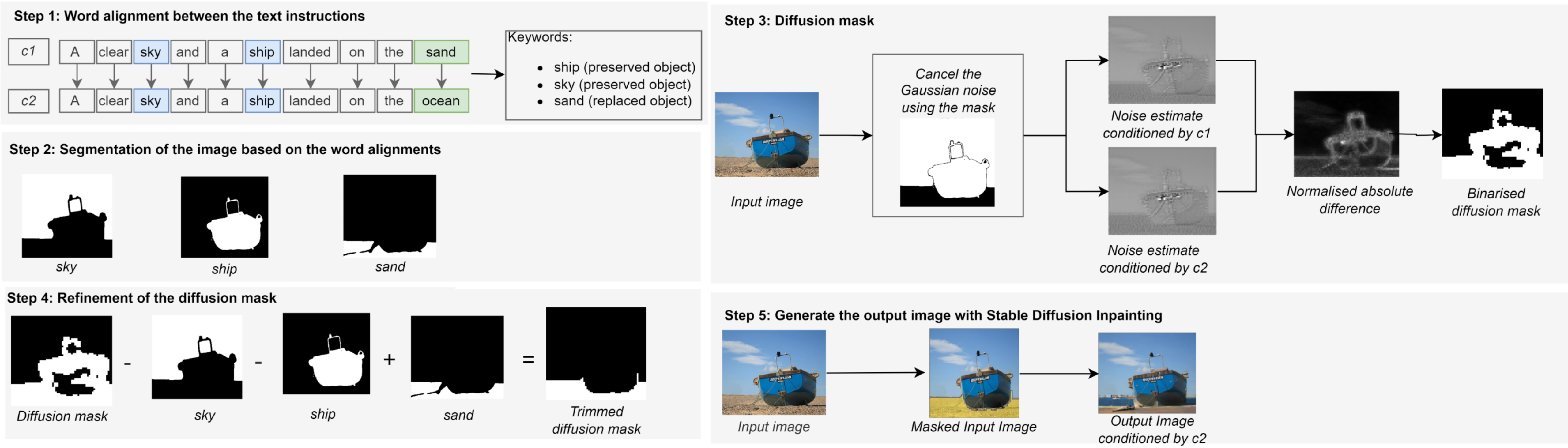
Two cats sitting on a bench

Two cats sitting on a sofa

Overview

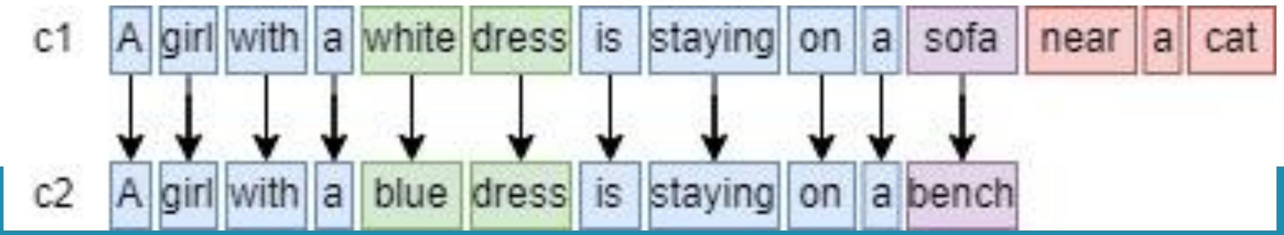
- *Proposed model* for addressing the current limitations;
- *Experimental Setup*:
  - Baselines;
  - Datasets;
  - Metrics;
- *Results*:
  - Qualitative and quantitative evaluation;
- *Conclusion.*

Proposed Model: Overview



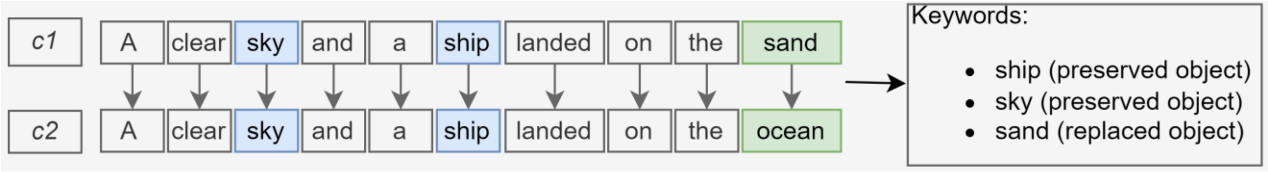
Proposed Model: DM-Align

- *Step 1*: Word alignment between the text instructions (*semi-Markov CRF* model - Lin et al. (2019))
  - Given two text instructions c_1 and c_2, we set the following assumptions:
    - Editable regions are indicated by:
      - Shared nouns with different modifiers (dress);
      - Substituted nouns (bench).
    - Non-editable regions are indicated by:
      - Shared nouns without modifiers / with identical modifiers (girl);
      - Deleted nouns (cat).



Proposed Model: DM-Align

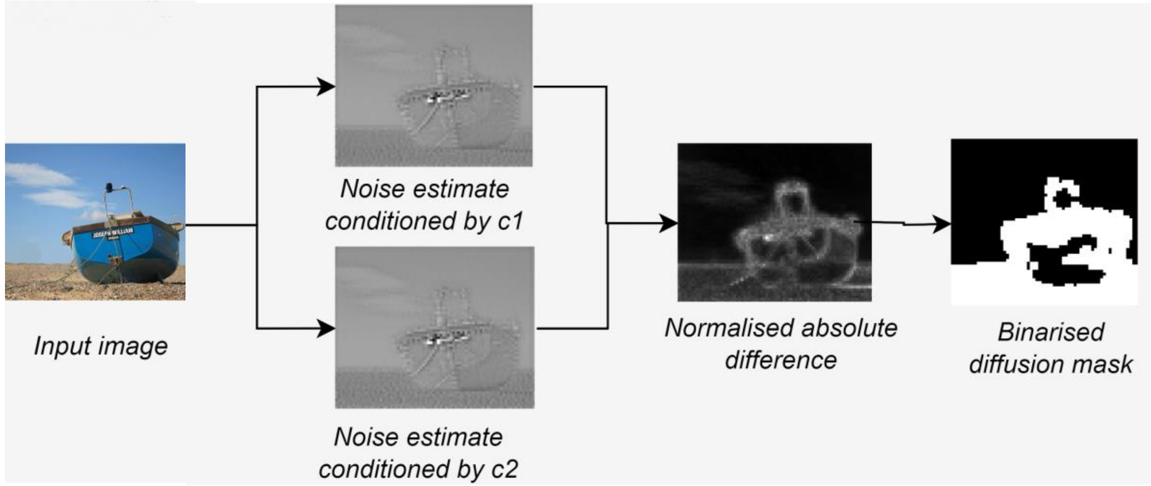
- *Step 2*: Segmentation of the image based on the word alignments
  - The word alignments computed at the first step are used to indicate the editable and non-editable objects in the image. The regions are detected using *Grounded-SAM.*
  - Considering the text instructions:



  - We use *Grounded-SAM* to identify the regions of the keywords:

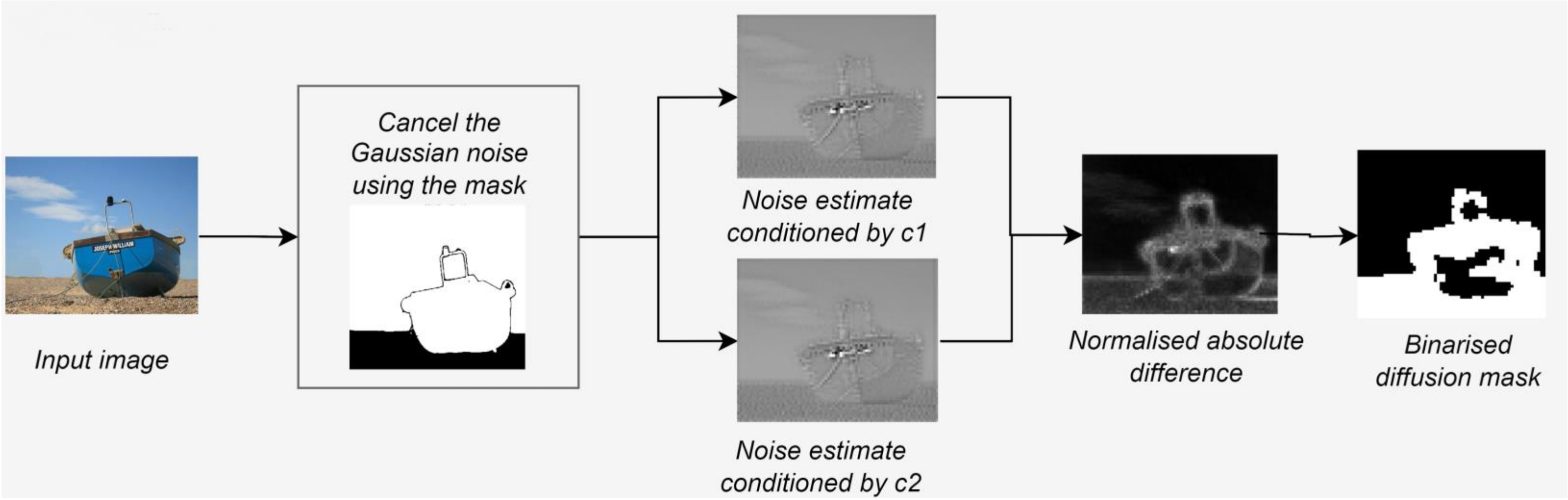
Proposed Model: DM-Align

- *Step 3*: Diffusion mask
  - Besides the editable/non-editable regions detected based on the word alignments, a diffusion mask is used to ensure:
    - Coherence of the output image with respect c_2;
    - Coping with the replacement of objects of different sizes.
  - Computation of diffusion mask:
    - Two noise estimates e_1 and e_2 are computed by running two diffusion models over the input image and each of the text instructions c_1 and c_2;
    - Diffusion mask is obtained after the normalization and binarization of the absolute difference between e_1 and e_2.



Proposed Model: DM-Align

- *Step 4.1*: Integration of the regions detected based on the word alignments in the diffusion process.
  - The noise variable of the forward process is cancelled for the non-editable regions detected based on the word alignments (ship and sky).



Proposed Model: DM-Align

- *Step 4.2*: Integration of the diffusion mask with the regions detected based on the word alignments
  - The diffusion mask with noise cancellation gives the initial context for the image editing.
  - The extension or the reduction of the diffusion mask by the regions detected based on the word alignments improves the precision of the final mask.

Non-editable regions Editable region

- *Step 5*: Use the refined diffusion mask and the Stable Diffusion Inpainting to edit the input image based on the text instruction c_2.

Experimental Setup

- Baselines: FlexIT, ControlNet, Prompt-to-Prompt, DiffEdit
- Datasets:
  - *BISON*$_{07}$
  - *DREAM*
- Evaluation Metrics:
  - *Text-based metrics*: *CLIP* score
  - *Image-based metrics*: *FID, LPIPS* and *pixel-wise Mean Square Error (PWMSE).*

Results and discussion

- How well can the *DM-Align* model edit a source image considering the complexity of the text instruction?

  - Image-based metrics: *DM-Align* outperforms all other baselines, especially for the case of long and elaborate text instructions (*BISON*$_{07}$)

  - Text-based metrics: *FlexIT* is the best baseline as the model is built on top of the *CLIP* model.

		Image-based			**Text-based**
		FID ↓	LPIPS ↓	PWMSE ↓	CLIPScore↑
BISON$_{07}$	FlexIT	72.44±0.15	0.49±0.00	42.34±0.02	**0.88±0.00**
	DiffEdit	82.46±0.26	0.46±0.00	50.96±4.07	0.79±0.00
	ControlNet	78.50±0.26	0.42±0.00	52.16±0.78	0.77±0.00
	PtP				0.77±0.00
	DM-Align	**60.05±1.35**	**0.27±0.00**	**34.72±0.55**	0.78±0.00
DREAM	FlexIT	147.56±1.34	0.71±0.00	53.49±0.01	**0.86±0.00**
	DiffEdit	125.71±1.62	0.71±0.00	53.52±0.84	0.77±0.00
	ControlNet	140.18±1.87	0.72±0.00	53.78±0.60	0.77±0.00
	PtP				0.78±0.00
	DM-Align	**110.20±0.30**	**0.69±0.00**	**50.62±0.25**	0.78±0.00

Results and discussion

- How well does the *DM-Align* model preserve the background?

  - *BISON*$_{07}$ compared with the best baselines improves:
    - *FID* by 96.26 %
    - *LPIPS* by 116.67%
    - *PWMSE* by 39.26%
  - *DREAM:*
    - *FID* by 55.64%
    - *LPIPS* by 4.51%
    - *PWMSE* by 13.22%

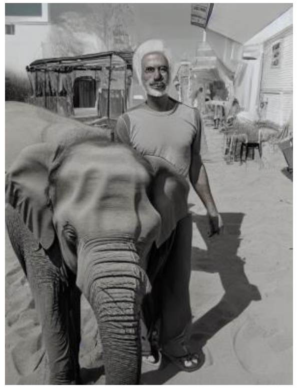
Results and discussion

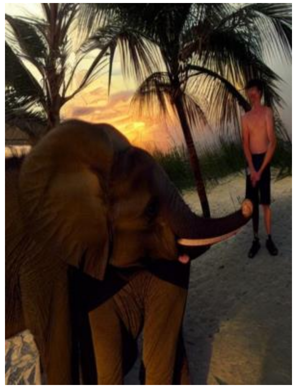
- Qualitative Evaluation

  - c_1 A man standing next to a baby elephant in the city. c_2. A man standing next to his elephant on the beach.


Initial image


DM-Align


ControlNet


DifiEdit


FlexIT

Results and discussion

- Qualitative Evaluation

  - c_1 A vase filled with red and white flowers. c_2. A vase filled with lots of colorful flowers.

Initial image

DM-Align

ControlNet

DifiEdit

FlexIT

Results and discussion

- Qualitative Evaluation

  - c_1 A man eating a hot dog next to a waterway.c_2. A man eating a hot dog at a crowded event.

Initial image

DM-Align

ControlNet

DifiEdit

FlexIT

Results and discussion

- Human Qualitative Evaluation
  - Randomly select 100 images from *BISON$_{07}$* and ask Amazon MTurk annotators to evaluate the editing process using a 5-point Likert scale based on the following aspects:
    - Q1: quality of the editing process based on the text instruction c_2;
    - Q2: preservation of the background;
    - Q3: the quality of the editing process in terms of compositionality, sharpness, distortion, color and contrast.

	Q1 ↑	Q2 ↑	Q3 ↑
FlexIT	3.77	4.12	3.83
DiffEdit	3.74	3.89	3.86
ControlNet	3.41	3.77	3.90
PtP	2.24	1.98	2.18
DM-Align	**3.89**	**4.35**	**3.95**

Conclusions and limitations

- Conclusions:
  - Due to the differentiation between the changed and unchanged content, the outputs generated by *DM-Align* have a high level of explainability.
  - Compared with the baselines, *DM-Align* demonstrate a better capability to keep the background and to edit images using elaborate and long text instructions

- Limitations:
  - While *DM-Align* can implement operations like insertion, deletion and replacement of objects, the model has difficulties when trying to change the position of objects.

Publications

- Related Calculus publications / work in progress
  - Dario Pavllo, Graham Spinks, Thomas Hofmann, Marie Francine Moens & Aurelien Lucchi (2020). Convolutional Generation of Textured 3D Meshes. In Proceedings of the Thirty-fourth Conference on Neural Information Processing Systems (NeurIPS).
  - Wolf Nuyts, Maria Trusca, Jonathan Thomm, Robert Hönig, Thomas Hofmann, Tinne Tuytelaars and Marie-Francine Moens (2024). Object-Attribute Binding in Text-to-Image Generation: Evaluation and Control (will be submitted soon).
- Current work:
  - Maria Trusca, Tinne Tuytelaars and Marie-Francine Moens (2024). DM-Align: Text-based semantic image editing using cross-modal alignments (under review).

Thank you!