S01 - Image Editing via Natural Language for Complex Scene Representation

6-month internship @ CEA List

Internship context

Based in Saclay (Essonne), the LIST is one of the two institutes of CEA Tech, the technological research division of the CEA. Dedicated to intelligent digital systems, its mission is to carry out technological developments of excellence on behalf of industrial partners in order to create value.

Within the LIST, the Laboratory of Vision and Learning for Scene Analysis (LVA) conducts research in the field of computer vision and artificial intelligence for the perception of intelligent and autonomous systems. The laboratory's research themes include visual recognition, behavior and activity analysis, large-scale automatic annotation, and perception and decision models. These technologies are applied in major sectors such as security, mobility, advanced manufacturing, healthcare, and sports…

Missions

Image generation from text has seen significant advances in recent years, thanks to the development of diffusion models. These methods have enabled the synthesis of images that are not only coherent with the natural language descriptions provided by humans (e.g., Glide [1]) but are also controlled by other modalities (e.g., ControlNet[2]).

Instruction-based image editing methods (such as InstructPix2Pix [3], MGIE [5] or EditBench [6]) have also emerged, relying on pre-trained text-to-image diffusion models and LLMs (Large Language Models). This has allowed users to effortlessly modify images using natural language instructions (see Figure 1).

However, while existing instruction-based image editing methods can handle simple requests, they are often insufficient when dealing with complex scenarios that require advanced reasoning and understanding capabilities.

The first type of complex scenario is when the original image contains multiple objects, and the instruction modifies only one of these objects by altering specific attributes (such as location, relative size, color, etc.). The second type of complex scenario arises when world knowledge is needed to identify the object to be modified (e.g., the object that can display the time, the clothes worn by the person carrying a bag standing next to the red car). Both scenarios are complex in terms of scene understanding (see Figure 2) and reasoning (see Figure 3). Addressing these scenarios is crucial for natural language-based image editing (personalizing content, anonymizing content, enriching a database for training image-based AI models, enabling automatic annotations...). Some early methods have begun exploring these complex cases (e.g., SmartEdit [4]).

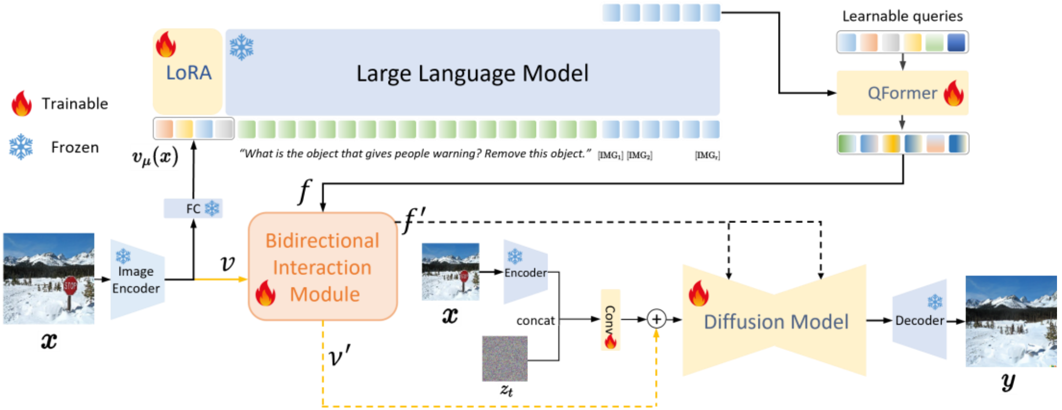
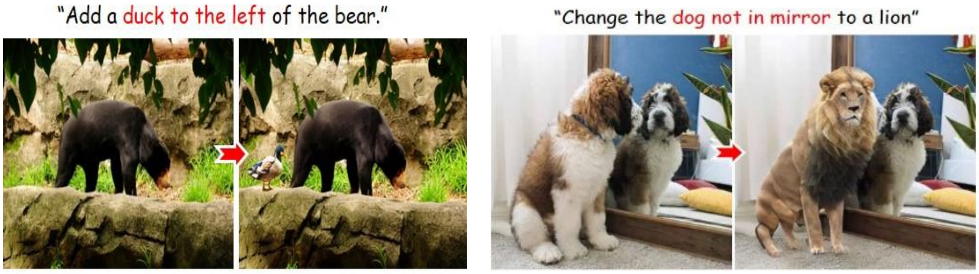


Figure 1: SmartEdit framework



"Add a duck to the left of the bear."
"Change the dog not in mirror to a lion"

Figure 2: SmartEdit on understanding scenarii

…/…

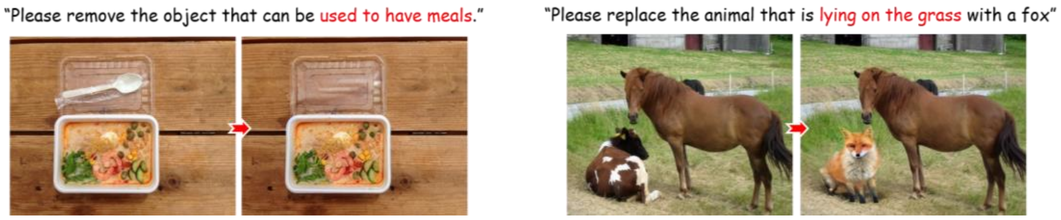

"Please remove the object that can be used to have meals."
"Please replace the animal that is lying on the grass with a fox"

Figure 3: SmartEdit on reasoning scenarii

Internship objectives

In this internship, we propose conducting a state-of-the-art review of image editing methods, testing the most promising approaches on complex scene images requiring advanced comprehension or reasoning (e.g., scene analysis, video surveillance, anomaly detection, road scenes…), and proposing a new method for complex image editing via natural language with minimal reliance on vision services (minimizing human interaction to just the prompt, without adding masks or bounding boxes for guidance, and controlling the editing process).

Mastering fine-grained image editing through natural language and reasoning would not only enable the generation of hard-to-acquire images but also facilitate the creation of images for training AI models.

References

[1] Nichol, A. Q., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., Mcgrew, B., ... & Chen, M. (2022, June). GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *ICML*.
[2] Zhang, L., Rao, A., & Agrawala, M. (2023). Adding conditional control to text-to-image diffusion models. In *ICCV*.
[3] Brooks, T., Holynski, A., & Efros, A. A. (2023). Instructpix2pix: Learning to follow image editing instructions. In *CVPR*.
[4] Huang, Y., Xie, L., Wang, X., Yuan, Z., Cun, X., Ge, Y., ... & Shan, Y. (2024). Smartedit: Exploring complex instruction-based image editing with multimodal large language models. In *CVPR*.
[5] Fu, T. J., Hu, W., Du, X., Wang, W. Y., Yang, Y., & Gan, Z. (2023). Guiding Instruction-based Image Editing via Multimodal Large Language Models. In *ICLR*.
[6] Wang, S., Saharia, C., Montgomery, C., Pont-Tuset, J., Noy, S., Pellegrini, S., ... & Chan, W. (2023). Imagen editor and editbench: Advancing and evaluating text-guided image inpainting. In *CVPR*.

Qualifications

- Students in their 5th year of studies (M2)
- Computer vision skills
- Machine learning skills (deep learning, LLM, VLM, generative AI…)
- Python proficiency in a deep learning framework (especially PyTorch or TensorFlow)

Job-related benefits

Join CEA List and LVA as an intern to:

- Work in one of the most innovative research organizations in the world (ranked in the global top 100, top 3 in France), addressing societal challenges to build the world of tomorrow
- Discover a rich ecosystem: privileged connections between the industrial and academic sectors
- Conduct research in an environment where autonomy and creativity are recognized, and where valorizing results is encouraged (publication of scientific articles, patents, and sharing of open-source code whenever possible).
- Join a young and dynamic team made up of research engineers, PhD students, post-doctoral researchers, and interns.
- Benefit from an internal computing infrastructure equipped with around 300 state-of-the-art GPUs.
- Receive a stipend between €1300 and €1400 per month.
- Have the opportunity to continue with a PhD or as a research engineer after the internship.
- Have the possibility of remote work, receive a 75% (instead of 50%) reimbursement on public transportation costs, and benefit from the "mobili-jeune" aid to reduce rent costs…

To apply, contact the laboratory with a CV and cover letter: lva-stages@cea.fr