Guiding Instruction-based Image Editing via Multimodal Large Language Models

Tsu-Jui Fu  Wenze Hu  Xianzhi Du  William Wang  Yinfei Yang  Zhe Gan

ICLR’24 (Spotlight)

https://arxiv.org/abs/2309.17102
Instruction-based Image Editing

- Support straightforward human command
  - Visual perception + **instruction understanding** → visual synthesis

- Challenge: **gap between guidance** of instruction and visual

- make it as minecraft
- replace mountain with city skylines

- improve its health
Multimodal Large Language Model (MLLM)

- MLLM contains **latent visual knowledge / creativity**
  - Explicit description and concrete intention to guide editing
  - Response is helpful but **redundant**

---

... The sauce would be made with healthier ingredients ... The pizza could be served with a side of fresh salad or vegetables to provide a balanced meal ...
MLLM-guided Image Editing (MGIE)

- Learn to derive **concise expressive instruction**
- Image editing via **latent imagination**
- **Parameter-efficient** end-to-end optimization

The pizza includes *vegetable toppings*, such as tomatoes and herbs. The sauce would be made with healthier ingredients ... The pizza could be served with a side of fresh salad or vegetables to provide a balanced meal ...

\[<\text{BOS}>\text{ The pizza includes } \text{vegetable toppings}, \text{ ...} \]

Visual Token

**Summarization**

... The sauce would be made with healthier ingredients ... The pizza could be served with a side of fresh salad or vegetables to provide a balanced meal ...

**MLLM**

**Embed**

**Adapter**

**Diffusion**

**Edit Head**

**LM Head**

**Expressive Instruction**

<IMS> The pizza includes ... [IMS] ...

**MLLM**

**Instruction**

image
Experiments

- Dataset (train on IPr2Pr only)
  - **Photoshop-style**: EVR / GIER
  - **Global optimization**: MA5k
  - **Local manipulation**: MagicBrush

- **Experiments**
  - make the barn a pagoda
  - turn the day into night
  - lake and snowy mountain
  - remove boy with red shirt
  - make it a red truck
  - give the lady a hat
  - increase the brightness
  - let laptop have a green web
Experiments

- Dataset (train on IPr2Pr only)
- Evaluation metrics
Experiments

- Dataset (train on IPr2Pr only)
- Evaluation metrics: SSIM↑ / CVS↑ / LPIPS↓ / DINO↑
- Baseline: InsPix2Pix
add lightning and make the water reflect

brighten image a lot, sharpen photo

remove the christmas tree in the background

let the donuts have strawberry glaze on them
Why MGIE is Helpful?

- Expressive instruction is **more aligned with input/goal image**
  - Provide explicit and concrete guidance

<table>
<thead>
<tr>
<th>Input CLIP-Score</th>
<th>Goal CLIP-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ins. 22.1</td>
<td>Expressive Ins. 25.9</td>
</tr>
<tr>
<td>Ins. 22.6</td>
<td>Expressive Ins. 26.5</td>
</tr>
</tbody>
</table>

- change the background as star wars
  - Add a background that is related to the Star War franchise, such as a lightsaber or a spaceship.

- remove all people except baby
  - If we remove all, we would be left with a baby sitting on the beach with a shark and a few balloons.

- desired image has more light contrast
  - Adjust between the light and dark areas to bring out the details of the leaves and the tree trunk.

- close the dog’s mouth
  - If the dog had its mouth covered, it would be as if it "play with" the stick or chew on it.
Takeaway

- Bridge the gap between guidance for **instruction-based image editing**
  - MLLM derivation + diffusion via latent imagination

Images:
- Add a storm
- Remove text
- Add contrast to simulate light
- Let the floor be made of wood
- As if the shop was a library
- Change the hair to purple color
- Make dark on RGB and sharpen
- Make the face happy

QR Codes:
- Code
- Demo