# Guiding Instruction-based Image Editing via Multimodal Large Language Models

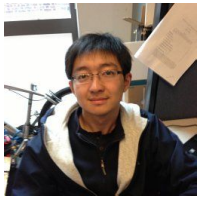


**Tsu-Jui Fu**


Wenze Hu


Xianzhi Du


William Wang


Yinfei Yang


Zhe Gan

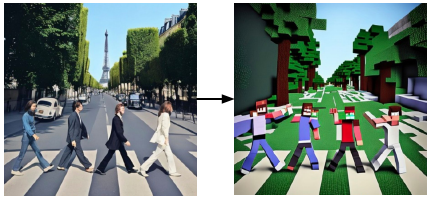



ICLR'24 (Spotlight)


https://arxiv.org/abs/2309.17102

# Instruction-based Image Editing

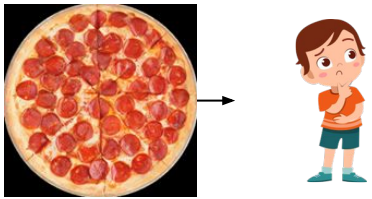
- Support straightforward human command
  - Visual perception + **instruction understanding** → visual synthesis



make it as minecraft

replace mountain with city skylines

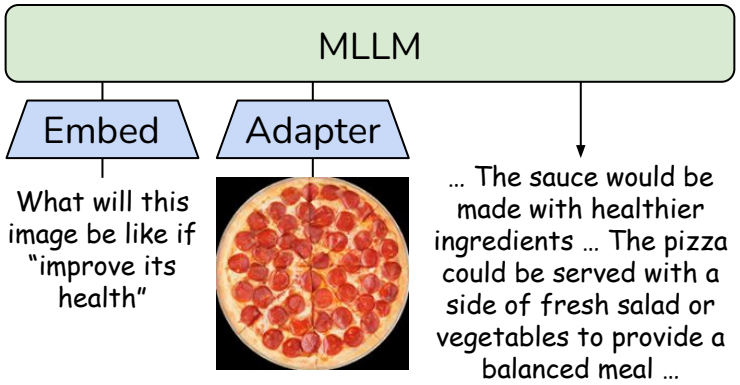
  - Challenge: **gap between guidance** of instruction and visual



improve its health

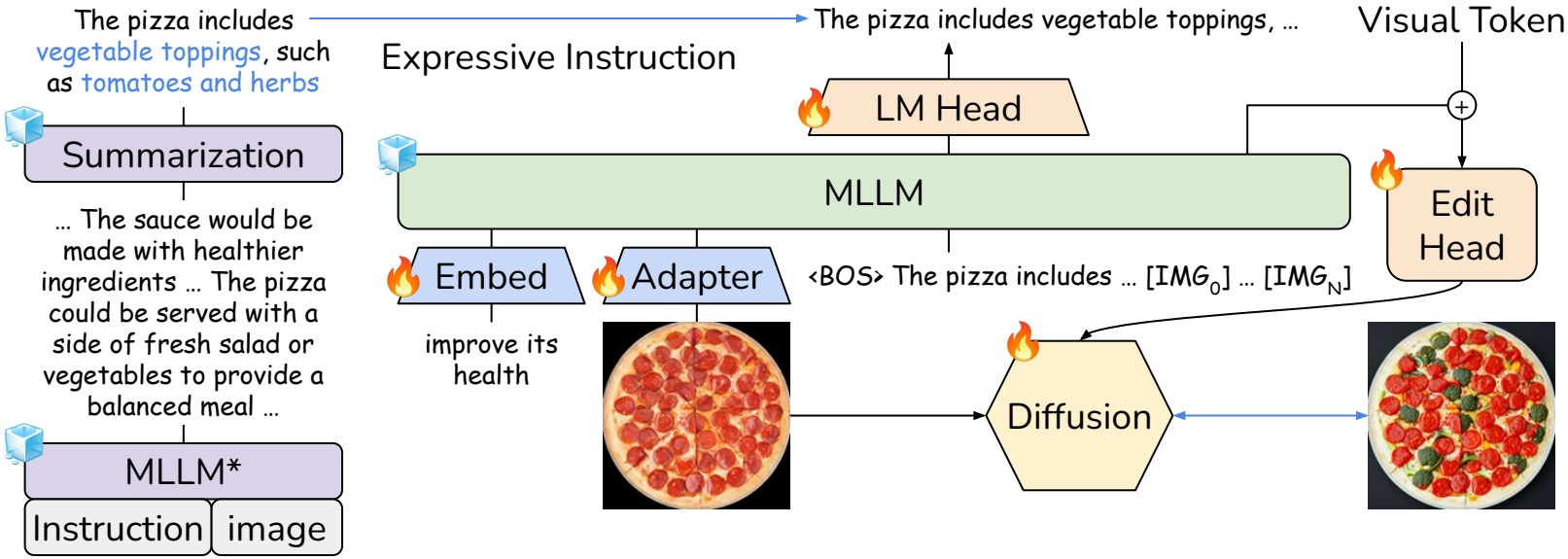
# Multimodal Large Language Model (MLLM)

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



# MLLM-guided Image Editing (MGIE)

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



# Experiments

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

make the barn a pagoda

turn the day into night

make it a red truck

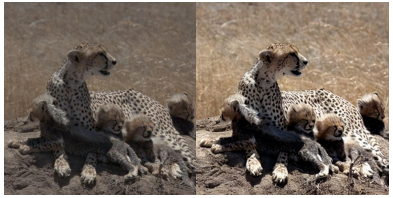
give the lady a hat

IPr2Pr

lake and snowy mountain

remove boy with red shirt

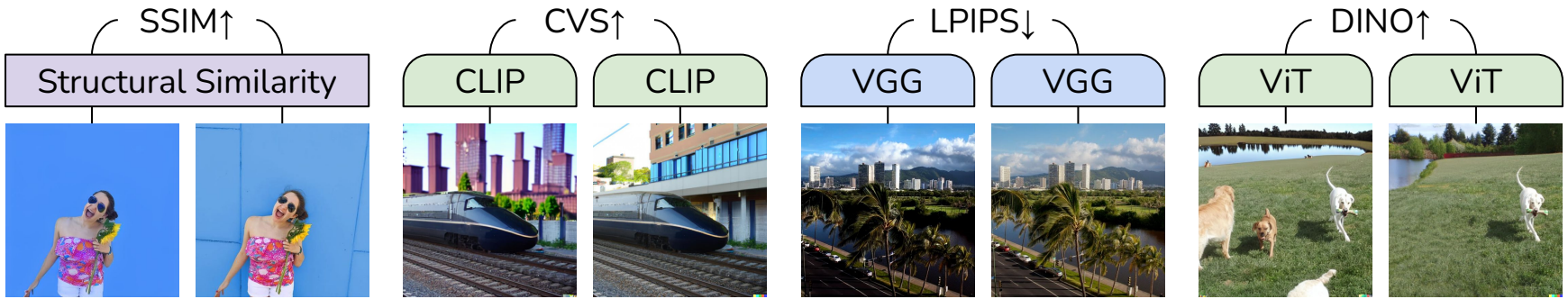
increase the brightness



let laptop have a green web

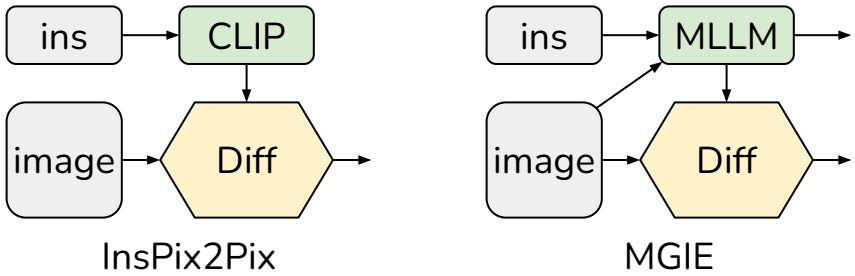
# Experiments

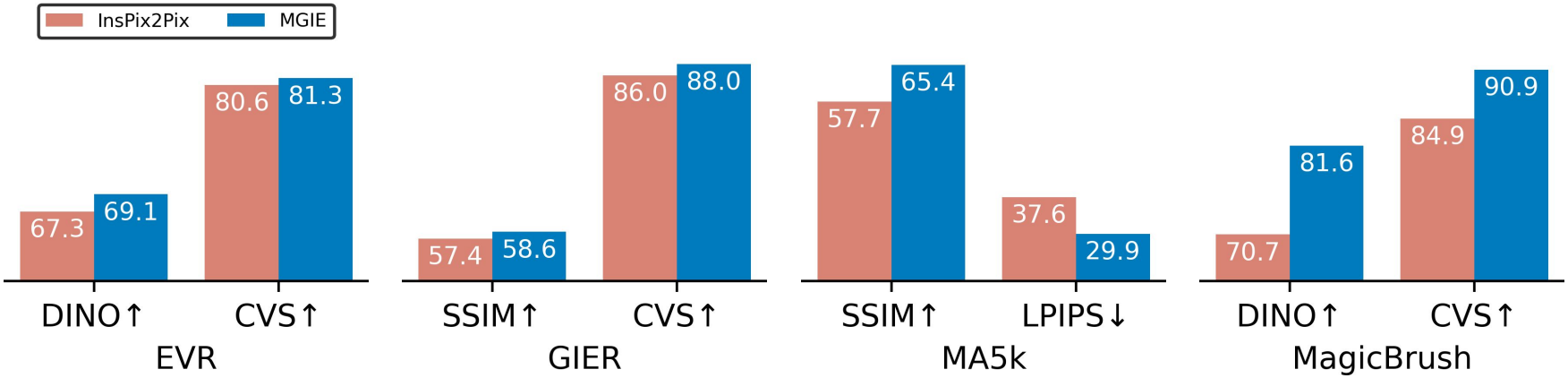
- Dataset (train on IPr2Pr only)
- Evaluation metrics



# Experiments

- Dataset (train on IPr2Pr only)
- Evaluation metrics: SSIM↑ / CVS↑ / LPIPS↓ / DINO↑
- Baseline: InsPix2Pix




InsPix2Pix
MGIE
67.3
69.1
80.6
81.3
DINO↑
CVS↑
EVR
57.4
58.6
86.0
88.0
SSIM↑
CVS↑
GIER
57.7
65.4
37.6
29.9
SSIM↑
LPIPS↓
MA5k
70.7
81.6
84.9
90.9
DINO↑
CVS↑
MagicBrush


add lightning and make the water reflect

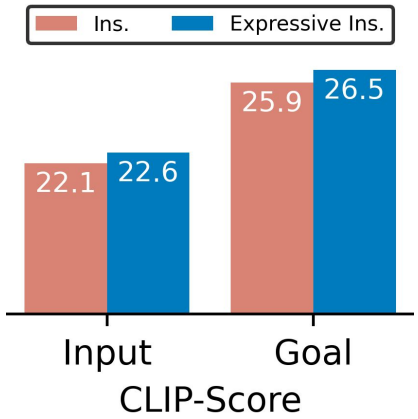

brighten image a lot, sharpen photo


remove the christmas tree in the background
Input
InsPix2Pix
MGIE
Ground-T


let the donuts have strawberry glaze on them
Input
InsPix2Pix
MGIE
Ground-T

# Why MGIE is Helpful ?

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




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

add a storm

remove text

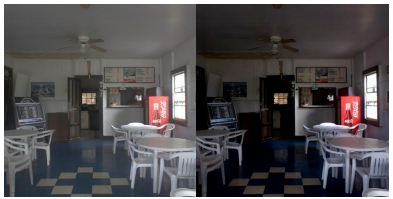

add contrast to simulate light

let the floor be made of wood

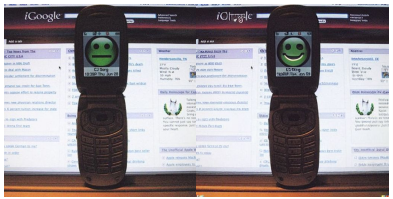

as if the shop was a library

change the hair to purple color

make dark on rgb and sharpen



make the face happy




Code

Demo