MagicMix: Semantic Mixing with Diffusion Models

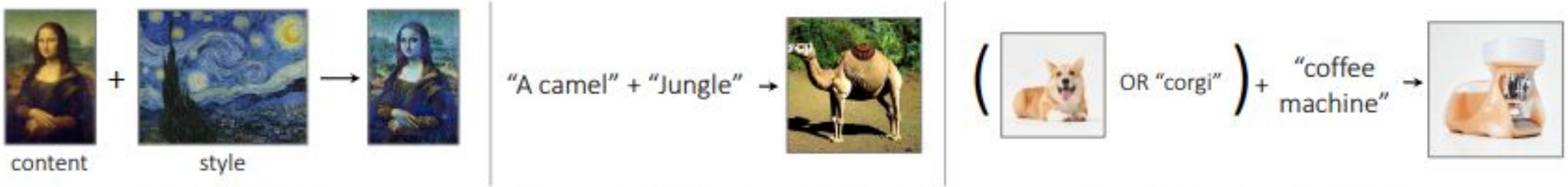
Jun Hao Liew*, Hanshu Yan*, Daquan Zhou & Jiashi Feng ByteDance Inc.

Prompt

`a mug that resembles a milk cow with four cow legs`

Introduction

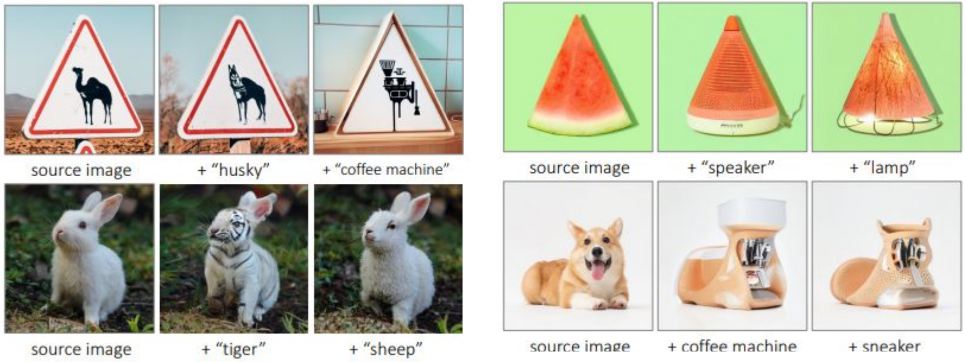
- large-scale text-conditioned image generation models
  - generating astonishing high-quality images given only text descriptions
  - DALL-E 2, Imagen, Parti, etc.,
- Style transfer, compositional generation and Semantic mixing
  - Style transfer :stylizes a image according to the given style while preserving the content.
  - Compositional generation : composes multiple individual components to generate a scene
  - Semantic mixing : fuse multiple semantics into one single novel object



(a) Style transfer (b) Compositional generation (c) Semantic mixing (ours)

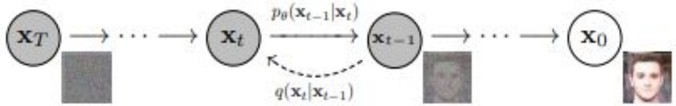
MagicMix

- Approach
  - requiring neither re-training nor user-provided masks
- Method
  - Layout semantics: corrupting a given real photo or denoising from a pure Gaussian noise from a given text prompt
  - Content semantics generation : injects a new concept and continues the denoising process until obtain the final synthesized results.

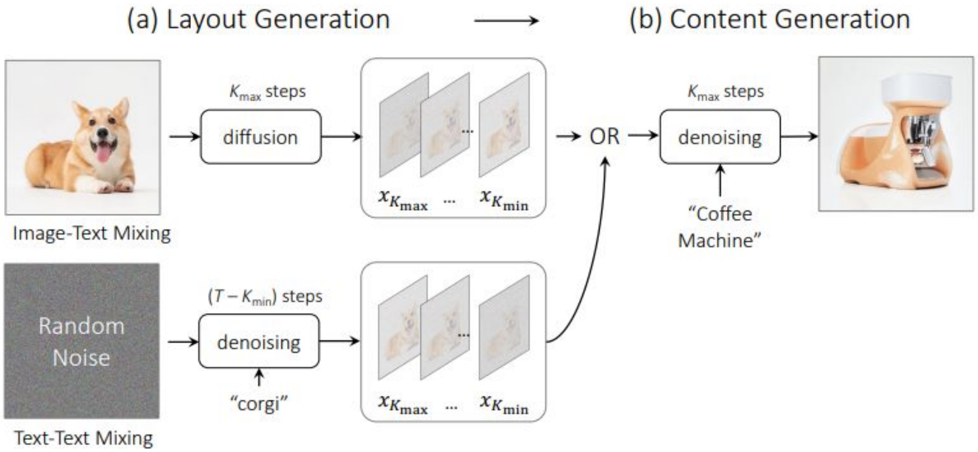


METHOD

- Denoising diffusion probabilistic model (DDPM)

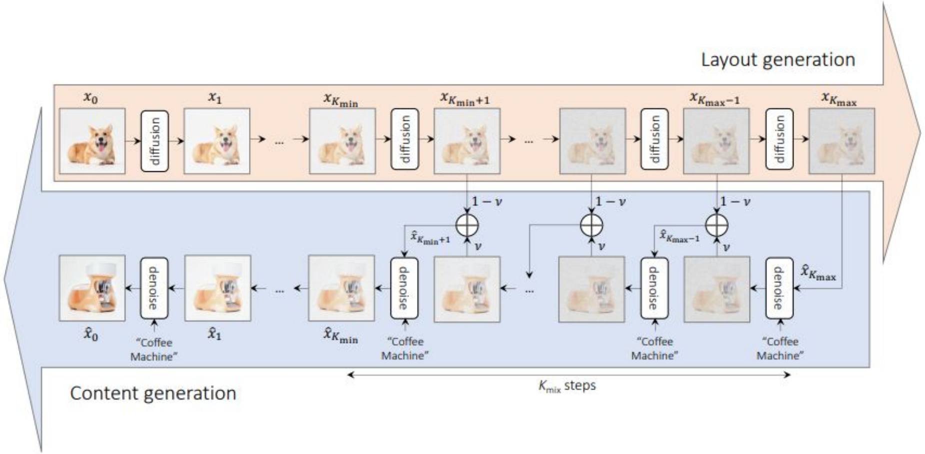


- MagicMix
  - generate images of mixed semantics by denoising the noisy layout images with a prompt
  - Image-text mixing &Text-text mixing



METHOD

- Image-text mixing
    - layout semantic: image
    - content semantic : text prompt
    - Craft its corresponding layout noises from step K_{min} to K_{max}
    - conditional generation process progressively mixes the two concepts by denoising
    - For each step $k \in [K_{min}, K_{max}]$, the generated noise of mixed semantics is interpolated with the layout noise to preserve more layout details.
- Text-text mixing
    - layout semantic : text prompt
    - content semantic : text prompt

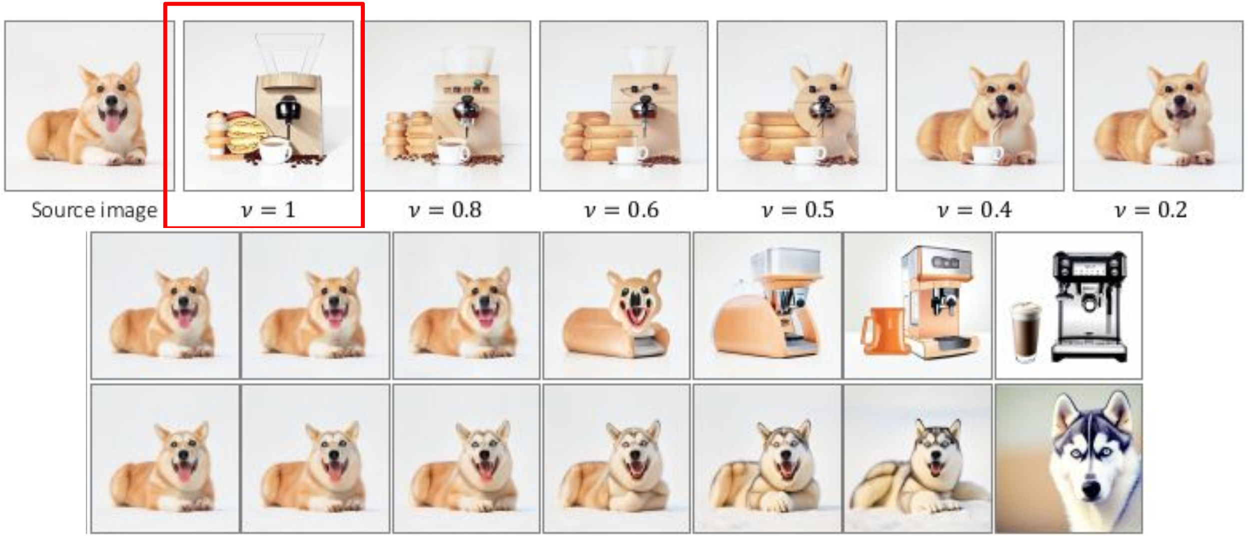


METHOD

- mixing ratio control :
  - K_{min} : the noisy layout image contains rich details from the given layout image
  - K_{max} : idestroy the irrelevant details and preserve the coarse layout.
  - Varying time-step for content injection.
    - when K is small : limited number of denoising steps only modify a small part of image content.
    - Much K is required to ensure sufficient steps for mixing (e.g., corgi and coffee machine), (e.g., corgi and husky)

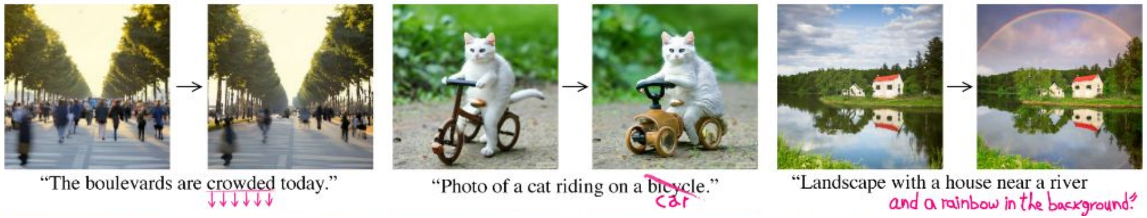
METHOD

- Preserving more layout details.
  - v controls the ratio between layout and content semantics.
- Optimal value of v.
  - determined by the semantic similarity between the two concepts
  - when two concepts has are extremely dissimilar diffusion model requires a large value of v

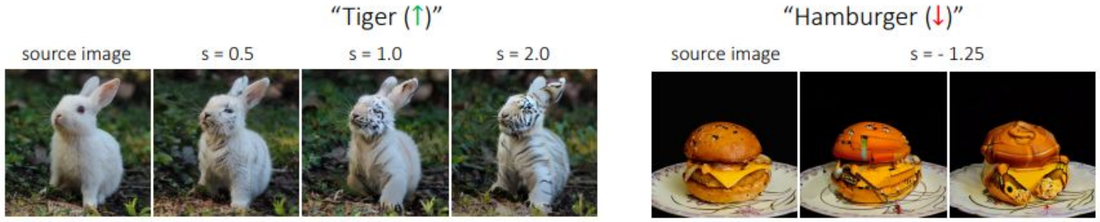


METHOD

- weighted image-text cross attention
  - Inspired by Prompt-to-Prompt
  - Concept removal
    - negative s : the diffusion models to generate an image with a layout similar to that of a text prompt while the non-text prompt object

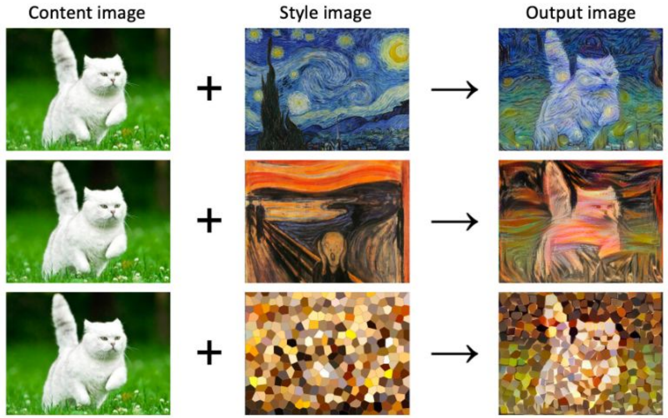


Prompt-to-Prompt

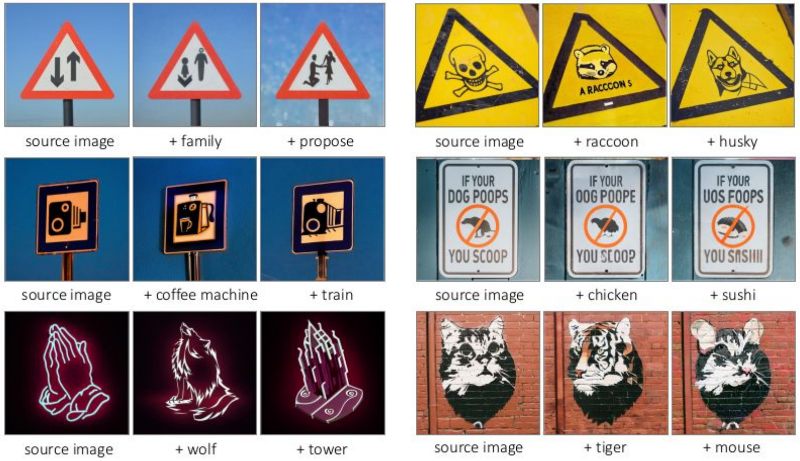


APPLICATIONS

- Semantic style transfer
    - Style transfer : the content image is stylized based on the reference style image without changing the image content
    - Allows the user to inject new semantics while preserving the spatial layout and geometry

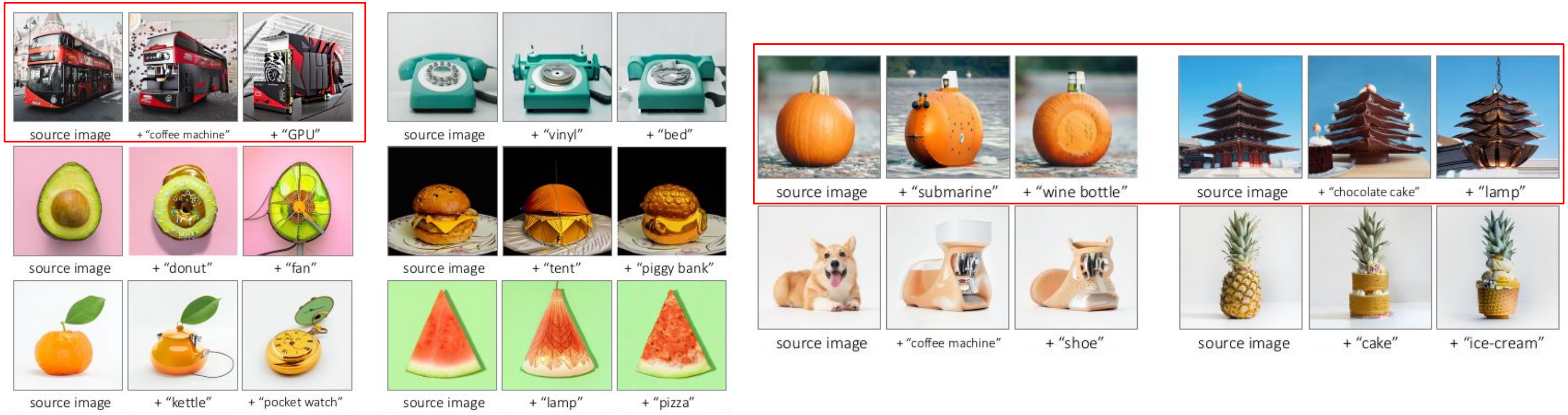


Style transfer



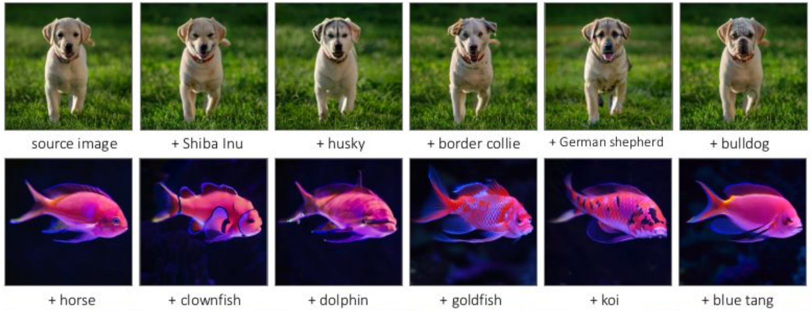
APPLICATIONS

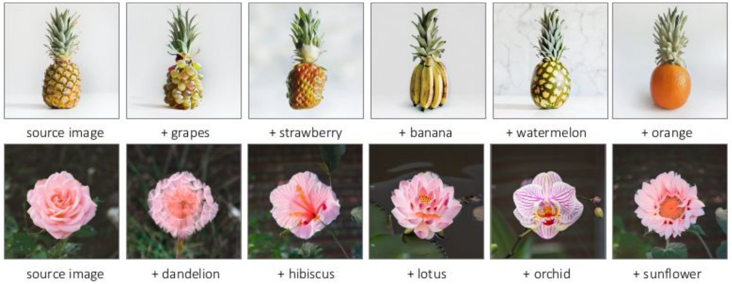
- Novel object synthesis
  - allows the synthesis of novel objects by injecting new concepts (e.g., coffee machine) into an existing object (e.g., bus).



APPLICATIONS

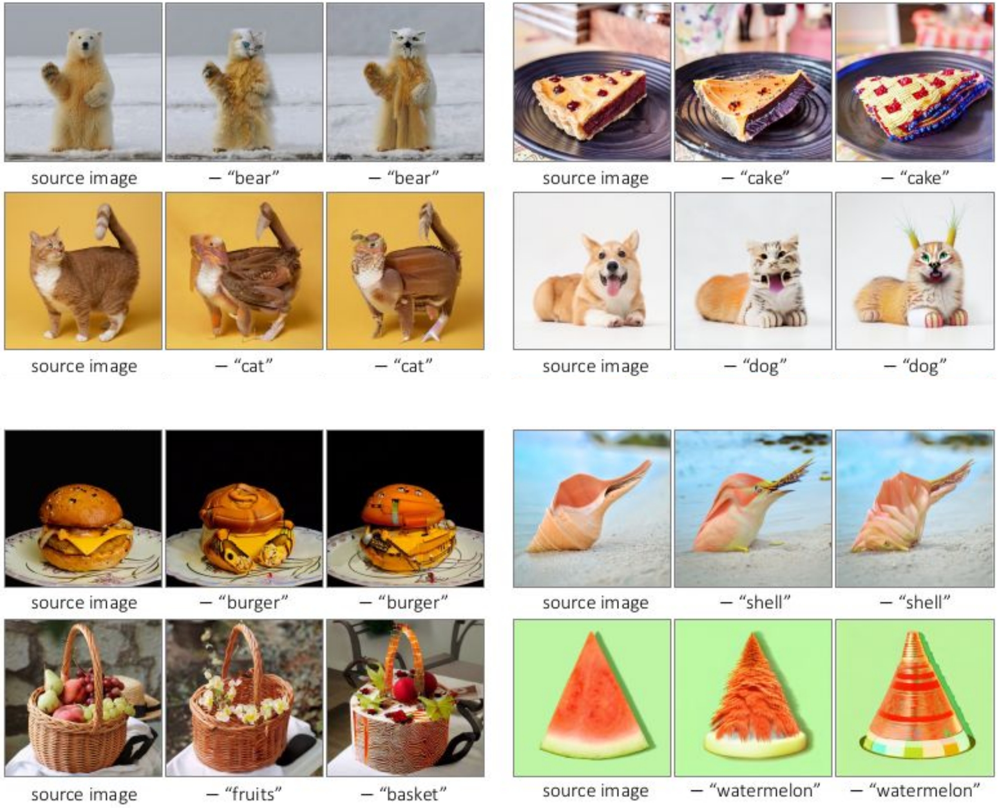
- Breed mixing
  - mixing two different species or animals





APPLICATIONS

- Concept removal
  - remove original semantic and let the model to decide what to generate aside from its original content.



APPLICATIONS

- Text-text semantic mixing
  - text-text mixing mode : the final synthesis result is unpredictable.

LIMITATIONS

- Shape similarity : two concepts cannot be mixed if they do not share any shape similarity