I2AM: Interpreting Image-to-Image Latent Diffusion Models via Attribution Maps

IJCAI 2024 workshop on Explainable Artificial Intelligence (XAI)

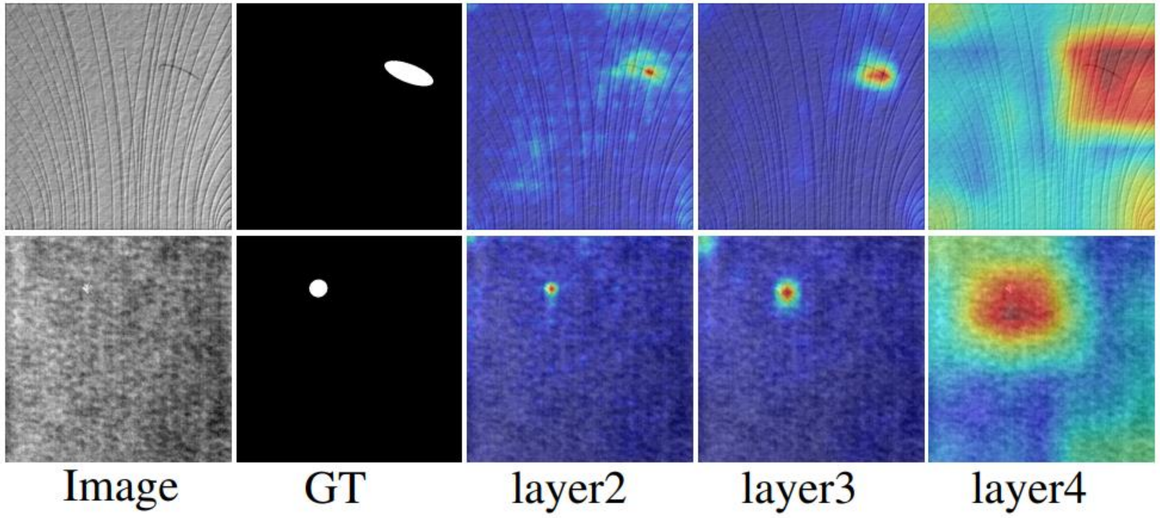
Junseo Park and Hyeryung Jang (Dongguk University, South Korea)

IJCAI
International Joint Conferences on Artificial Intelligence Organization

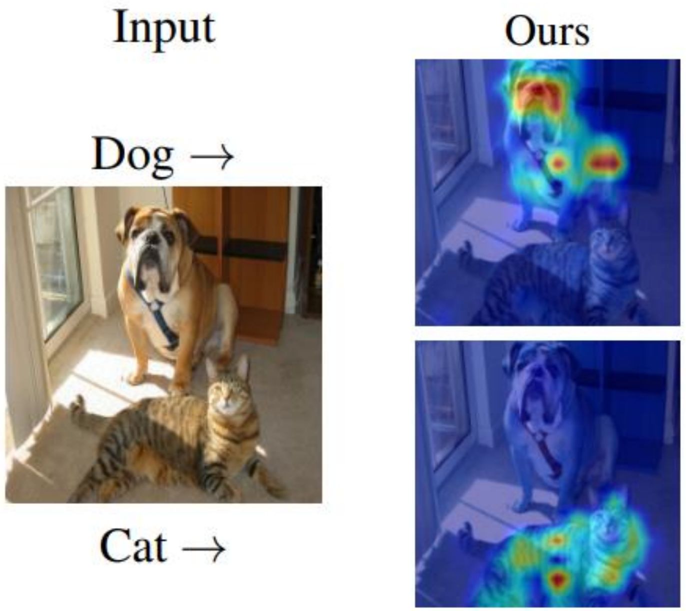
dongguk UNIVERSITY

1 Motivation

- Interpreting models using attribution map
  - Explainability is essential for enhancing trust and accountability by making AI decision-making transparent
  - Earlier efforts leveraged CNN-based image classifiers to highlight areas of interest
  - Recently, the emergence of transformers has shifted the focus towards using **attention**



CNN-based: class activation map [1]



Transformer-based: attention map [2]

1 Motivation

- Interpreting latent diffusion models **(LDMs)**
  - Analysis of text-to-image LDMs using attribution maps have advanced recently
  - There is currently a shortage of studies on image-to-image **(I2I)** LDM

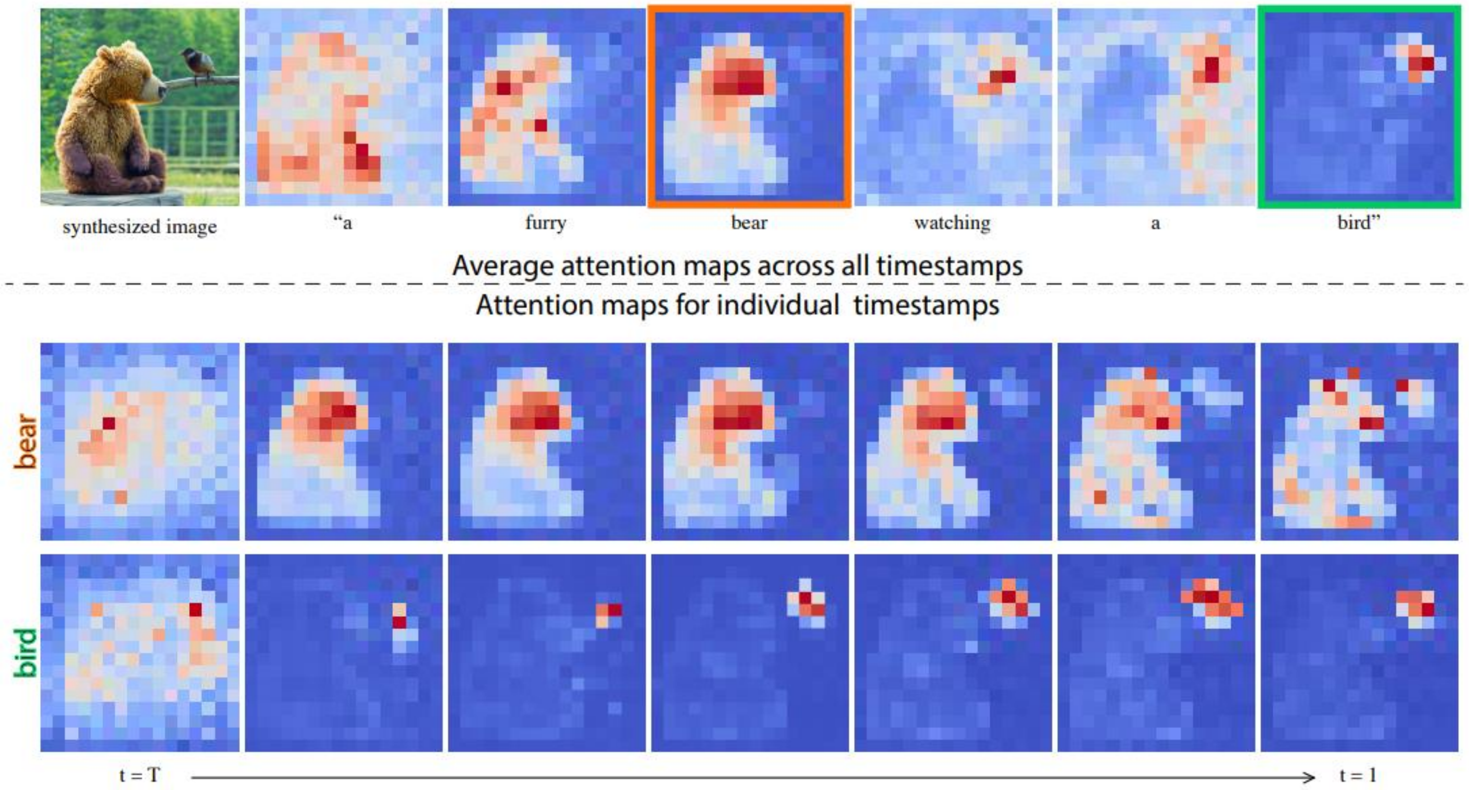


Image source [3]

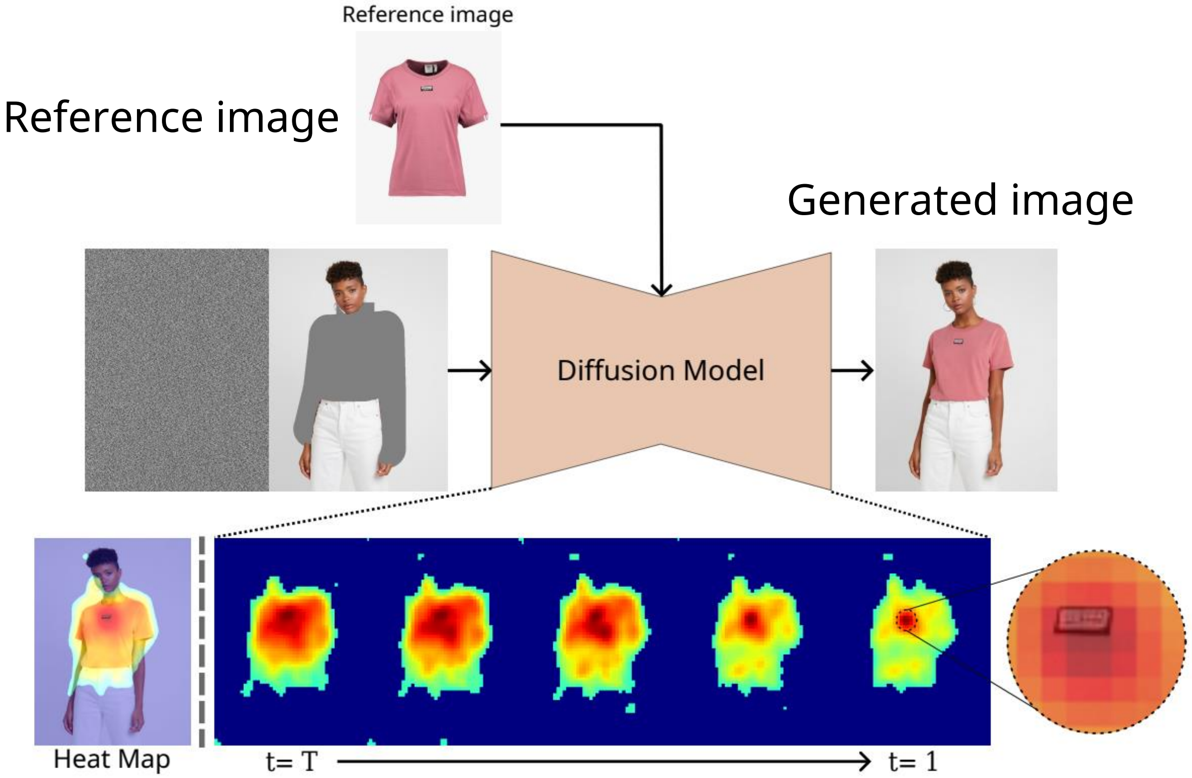
Image source [4]

1 Motivation

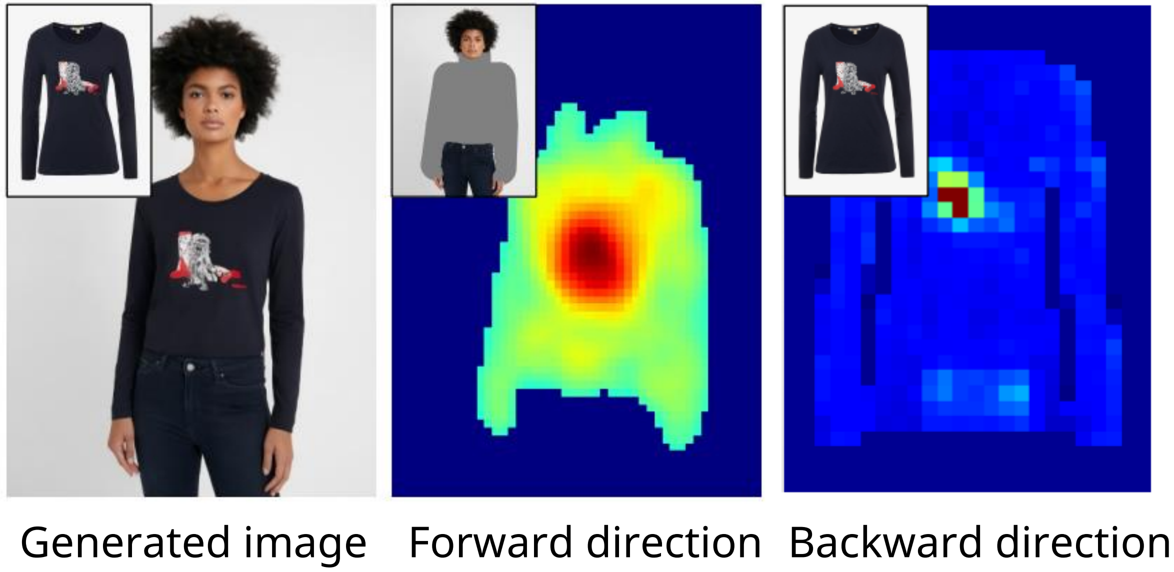
- Differences between text-to-image and image-to-image
  - Text-conditioned models
    - ✓ generate images that visually interpret provided text descriptions
    - ✓ Token-wise interpretation is practical
    - ✓ Etc.
  - Image-conditioned models
    - ✓ transform a reference image into a different visual form of the image
    - ✓ Patch-wise interpretation is less practical due to the spatial and contextual continuity
    - ✓ Etc.

2 Research Topic

- Interpreting image-to-image latent diffusion models focusing on inpainting
- Basic image-to-image LDMs performing inpainting task (VITON)
    - Input clothing **(reference image)** condition utilizing cross-attention
    - Concatenate various conditions to noisy input
    - Model (U-Net) generates more clear image by predicting the noise

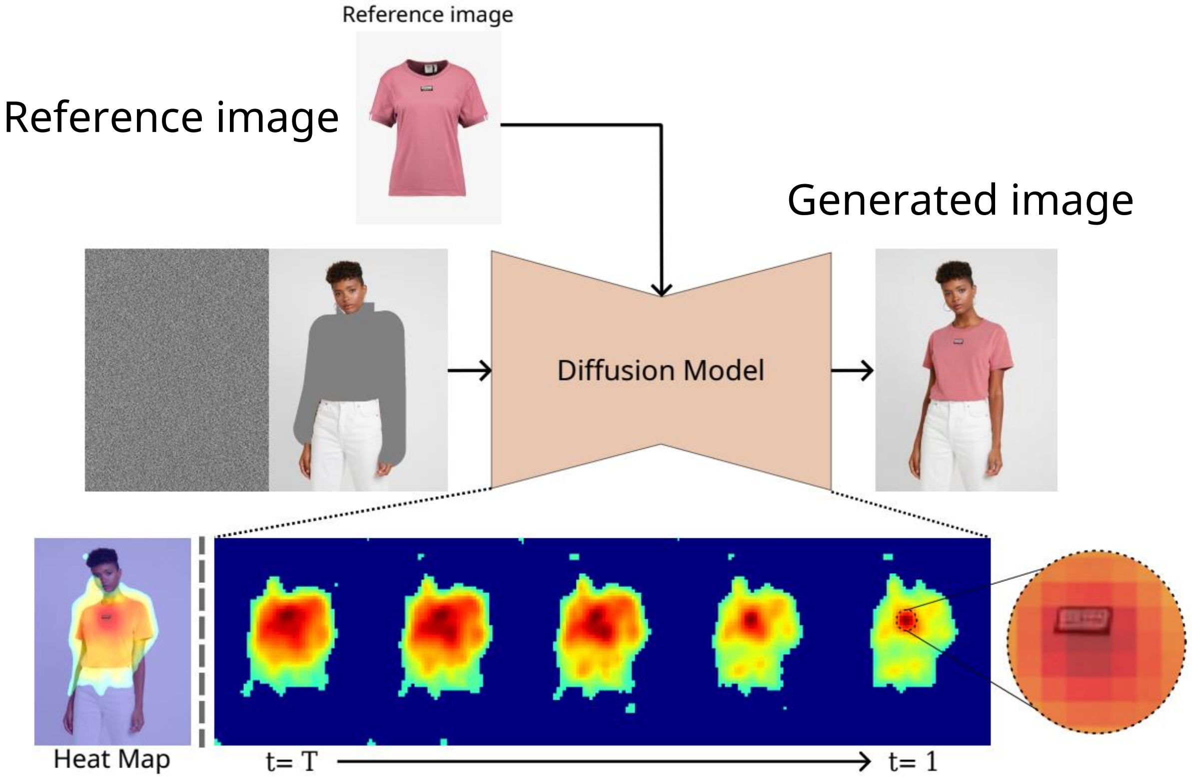


Attribution maps for generated/reference images

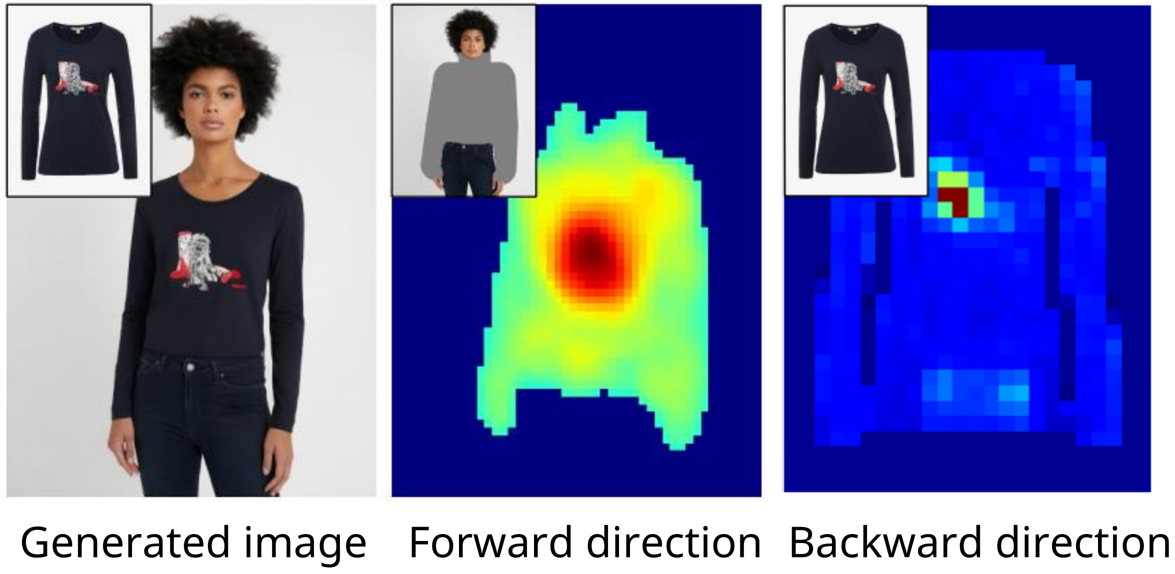


Methodology

- I^2AM: **I**mage-to-**I**mage **A**ttribution **M**aps method
  - Use cross-attention map **(attribution map)** to visualize generation process
  - Analyze generation process across time steps, attention heads
  - While text is abstract, image (e.g., clothes) maintains spatial information in latent space
  - So, we can facilitate clear visualization of the condition



Attribution maps for generated/reference images



3 Methodology

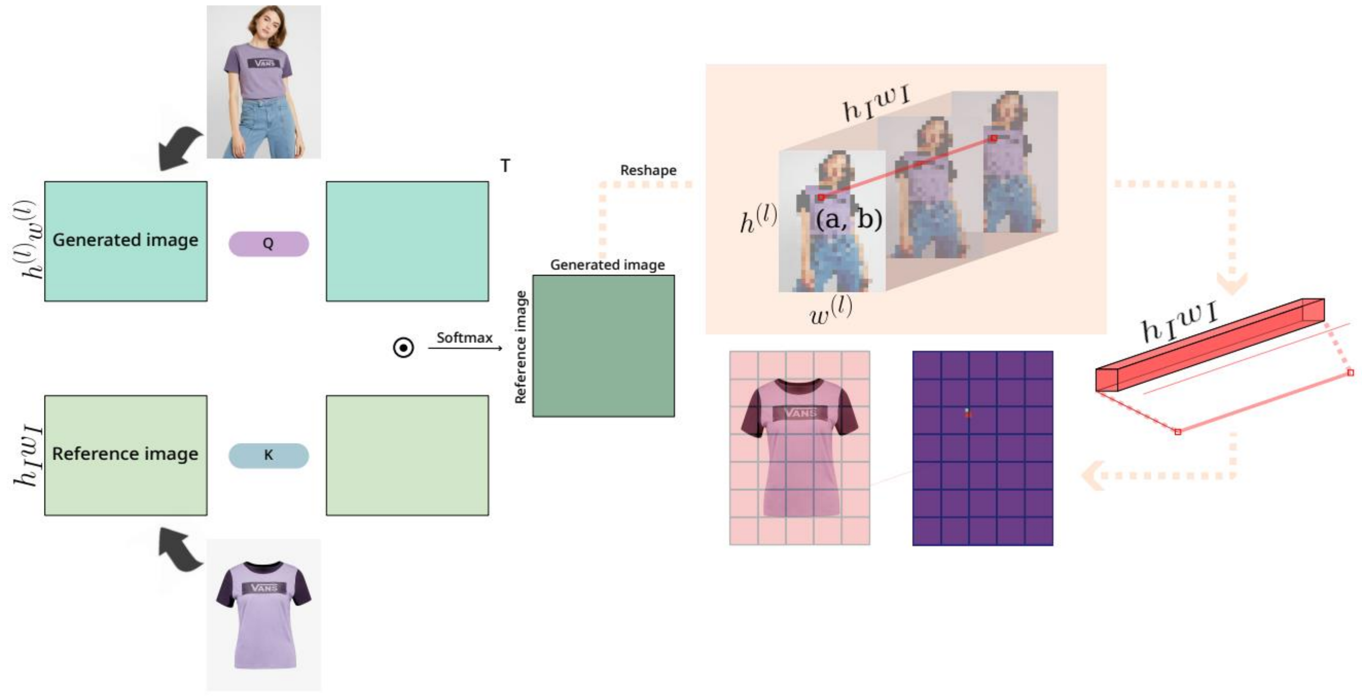
- I^2AM: **I**mage-to-**I**mage **A**ttribution **M**aps method
  - Attribution maps for generated/reference images
    - ✓ Time-and-head integrated attribution maps
    - ✓ Head/Time integrated attribution maps
    - ✓ Specific-reference attribution maps

$\mathbf{t}: [\mathbf{1}, \mathbf{T}], \textbf{attention head}: \mathbf{8}$

$f: pre - attention\ output\ vector$

$W_q, W_k: projection\ metrics\ of\ key\ and\ queries$

$c_I: embeddings\ of\ reference\ image$



$$\mathrm{Softmax}\left(\frac{(\mathbf{W}_q^{(l)}\mathbf{f}_t^{(l)})(\mathbf{W}_k^{(l)}\mathbf{c}_I)^\top}{\sqrt{d}}\right)$$

soft-max in forward direction

$$\mathrm{Softmax}\left(\frac{(\mathbf{W}_k^{(l)}\mathbf{c}_I)(\mathbf{W}_q^{(l)}\mathbf{f}_t^{(l)})^\top}{\sqrt{d}}\right)$$

soft-max in backward direction

Experimental results

- I^2AM: **I**mage-to-**I**mage **A**ttribution **M**aps method
  - Attribution maps for generated/reference images
    - ✓ Time-and-head integrated attribution maps
    - ✓ Head/Time integrated attribution maps
    - ✓ Specific-reference attribution maps
- Models
  - Paint-by-example [5]
  - DCI-VTON [6]
  - StableVITON [7]
- Sampler
  - DDIM

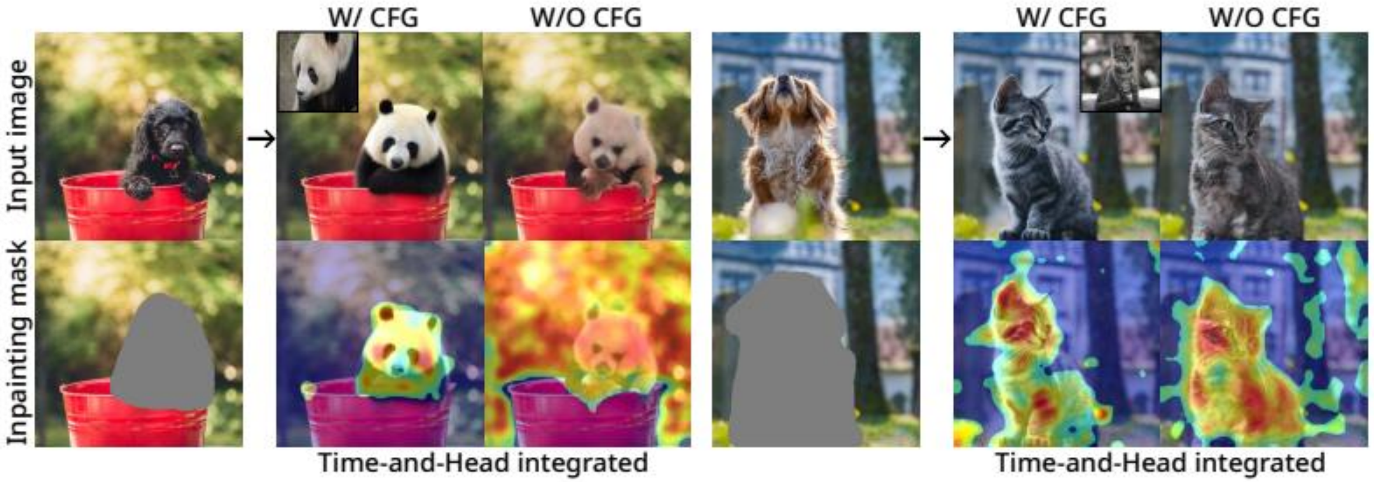
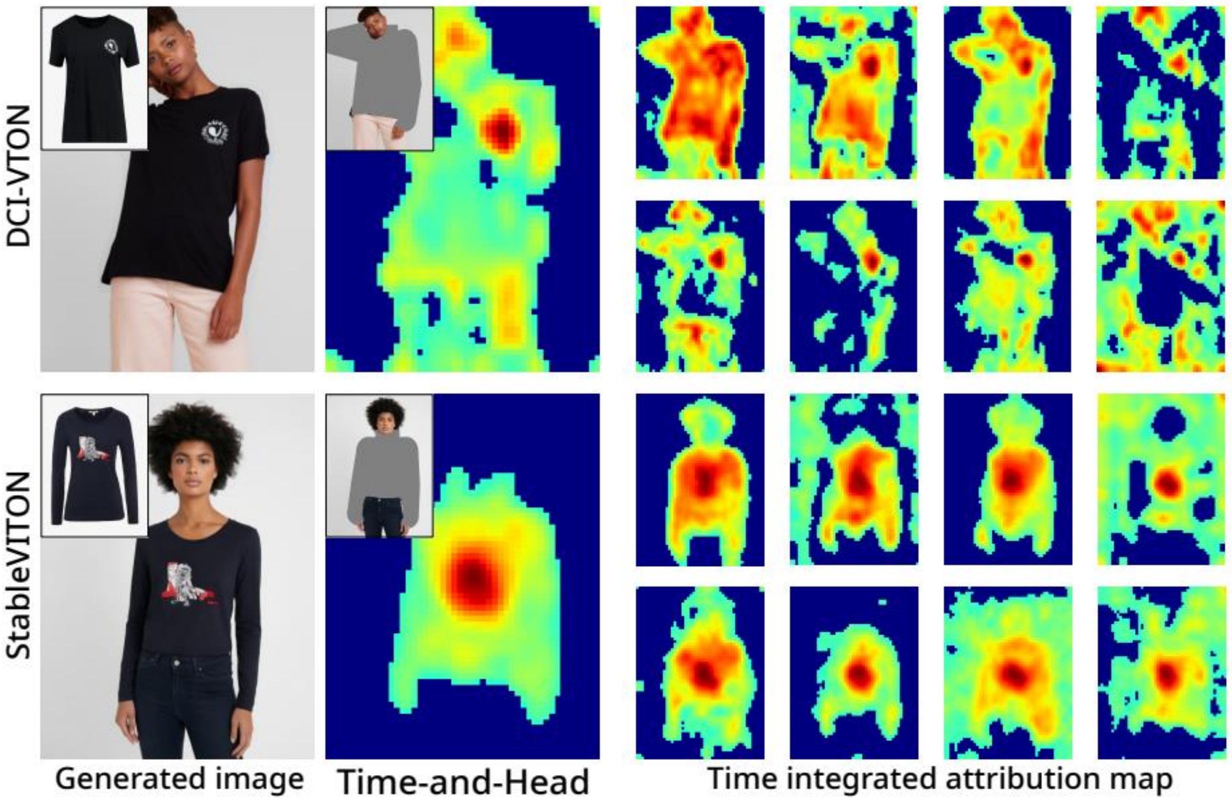
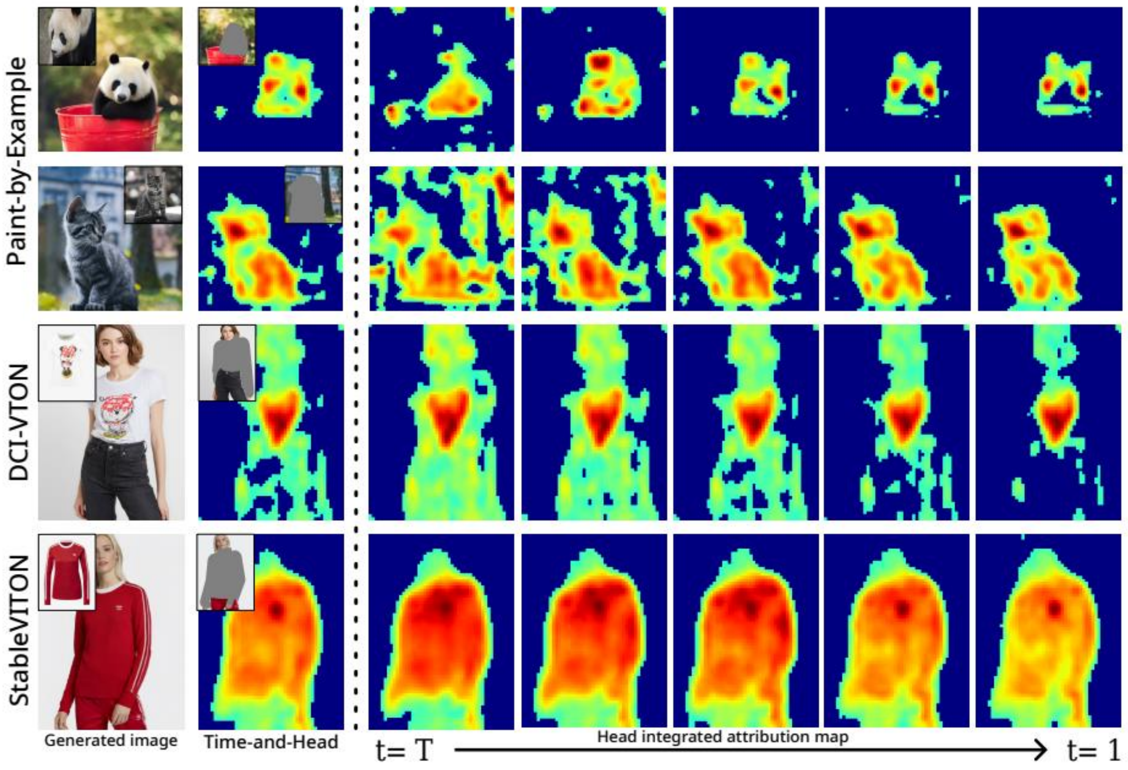


Figure 4: Time-and-Head integrated attribution map visualization, both with and without CFG. The dispersion of attention scores exceeded the inpainting mask's range when CFG was not used.

Experimental results

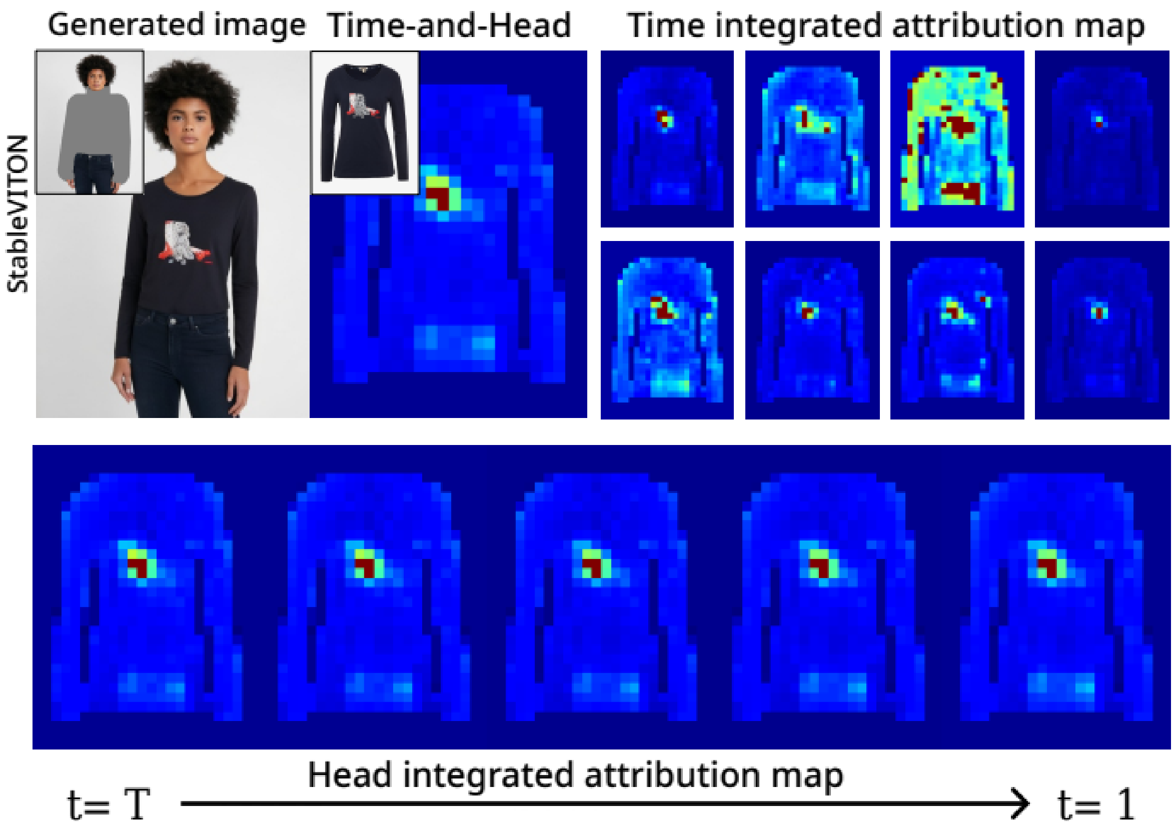
- Results of attribution maps for the generated image
  - The model gradually forms the object's structure, consistently assigning high attention scores to important features such as facial details or clothing logos.

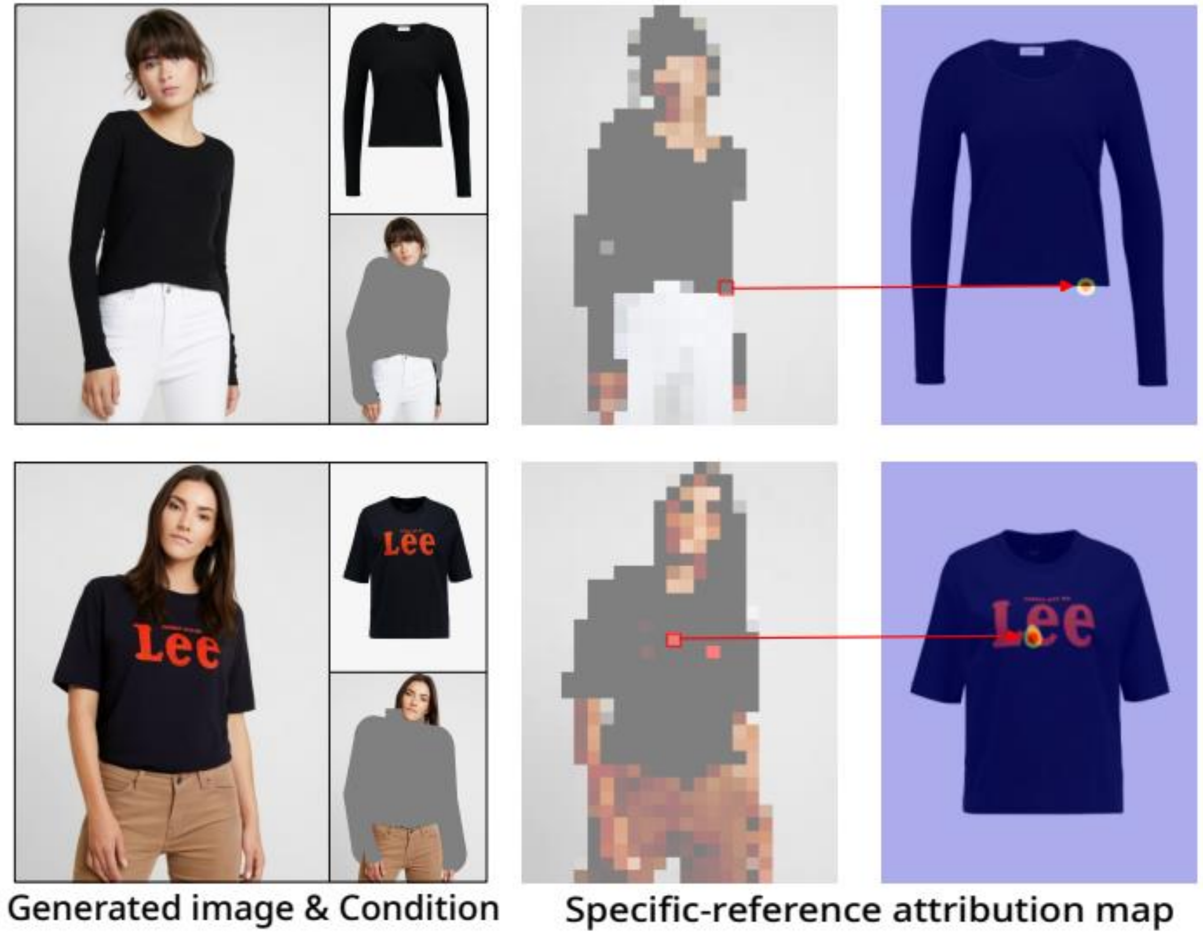




Experimental results

- Results of attribution maps for the reference image
  - To confirm whether meaningful information is extracted from the reference image for image synthesis, one needs to examine the reference attribution map





5 Conclusion

- Our contributions
  - Propose analysis and visualization methods for I2I LDMs
  - Provide insights into generation process of I2I LDMs by analyzing attribution maps at each time and attention head
  - Present attribution maps for the generated and reference images using characteristics of I2I LDMs

6 Reference papers

[1] Jiang, Peng-Tao, et al. “Layercam: Exploring hierarchical class activation maps for localization.” IEEE Transactions on Image Processing (2021)

[2] Chefer, Hila, Shir Gur, and Lior Wolf. "Transformer interpretability beyond attention visualization." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021.

[3] Hertz, Amir, et al. "Prompt-to-prompt image editing with cross attention control." *arXiv preprint arXiv:2208.01626* (2022).

[4] Tang, Raphael, et al. "What the daam: Interpreting stable diffusion using cross attention." *arXiv preprint arXiv:2210.04885* (2022).

[5] Yang, Binxin, et al. "Paint by example: Exemplar-based image editing with diffusion models." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.

[6] Gou, Junhong, et al. "Taming the power of diffusion models for high-quality virtual try-on with appearance flow." *Proceedings of the 31st ACM International Conference on Multimedia*. 2023.

[7] Kim, Jeongho, et al. "Stableviton: Learning semantic correspondence with latent diffusion model for virtual try-on." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024.

Thank you.

E-mail: mki730@dgu.ac.kr