I^2AM: Interpreting Image-to-Image Latent Diffusion Models via Bi-Attribution Maps

이중 속성 맵을 통한 이미지 기반 잠재 확산 모델 해석

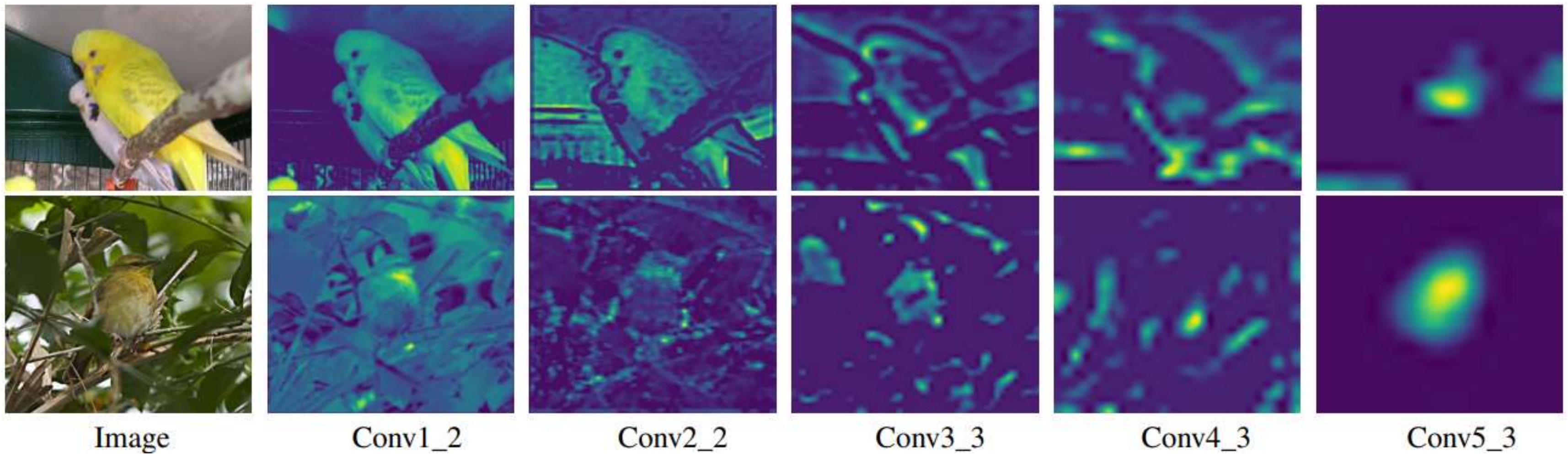
소속: 동국대학교 컴퓨터AI학과
발표자: 박준서
지도교수: 장혜령

Background

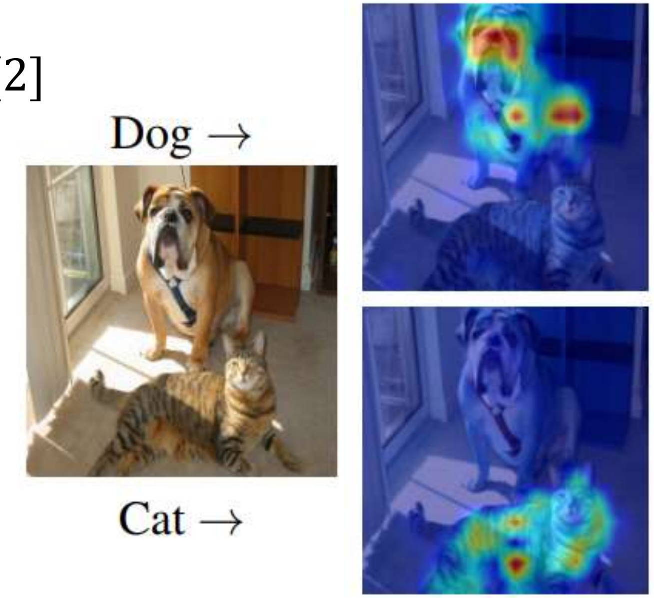
Interpretation using attribution maps

- Explainability = Trust & Accountability
  - A key element in making AI decision-making transparent.
  - Early[1]: CNN-based classifiers → visualization of regions of interest
  - Recent[2]: Shift of focus with the advent of transformers

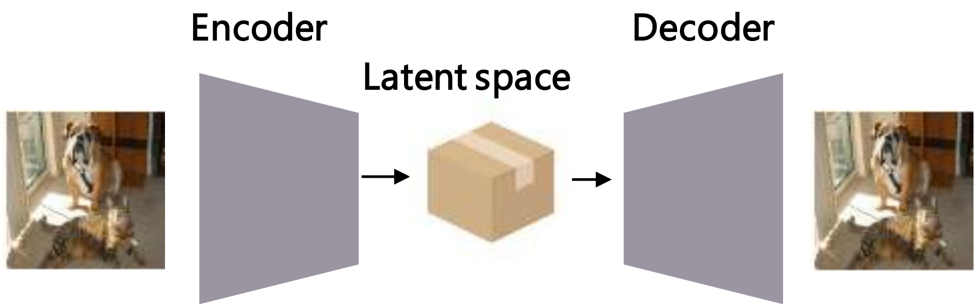
[1]



[2]



Background

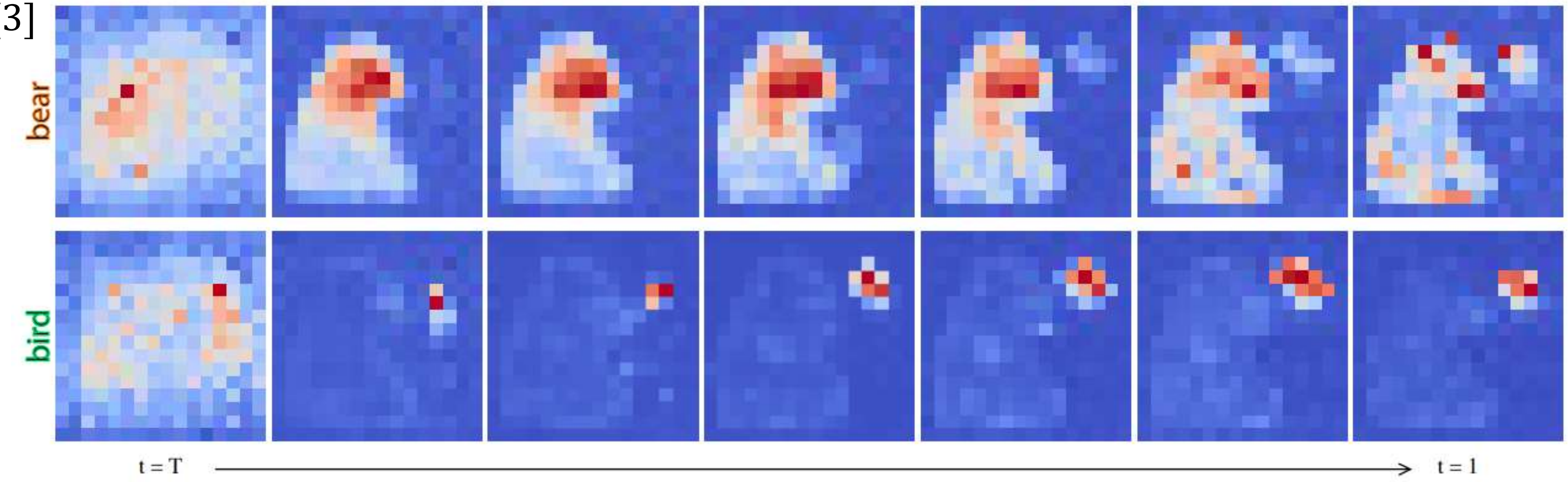


Interpreting Latent Diffusion Models

- Latent Diffusion Model (LDM)
  - Image generation in latent space
  - Complex model architecture → increased need for interpretation
  - Techniques required to understand the model's decision-making process

Attention maps for individual timestamps

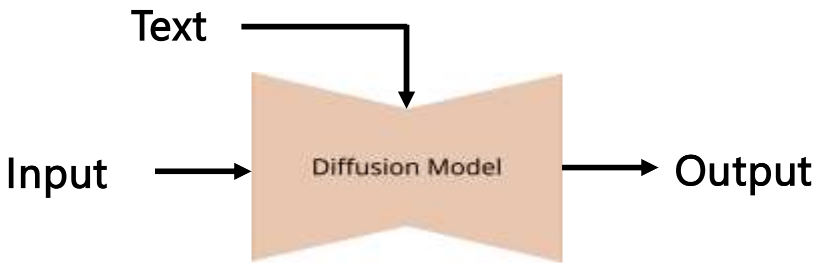
[3]



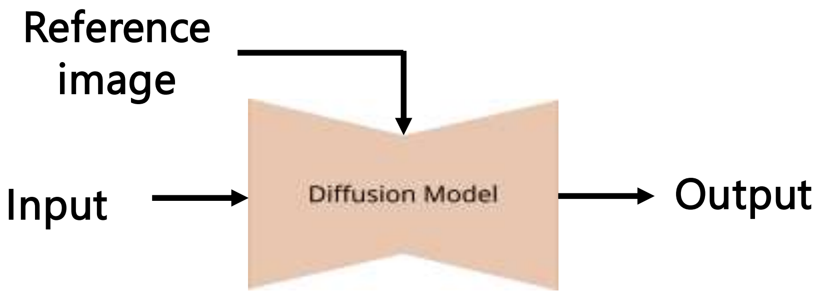
Problem Definition

Interpretability in LDMs: T2I vs I2I

- Text-to-Image (T2I)
  - Active research, with diverse methods [4]
- Image-to-Image (I2I)
  - Underexplored area, requiring research

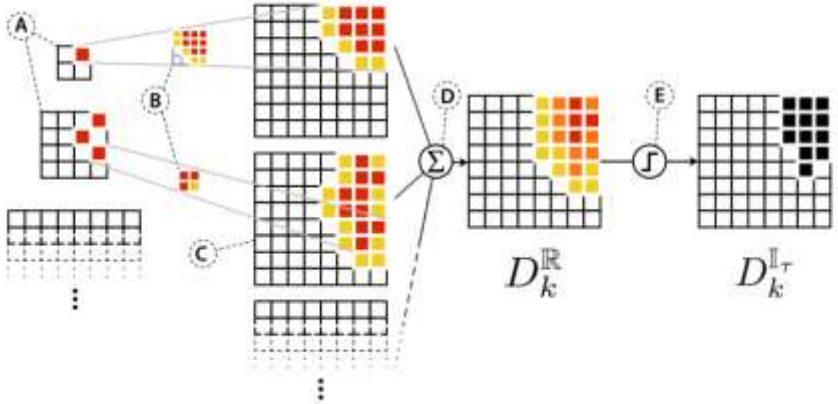


<Text-to-image>



<Image-to-image>

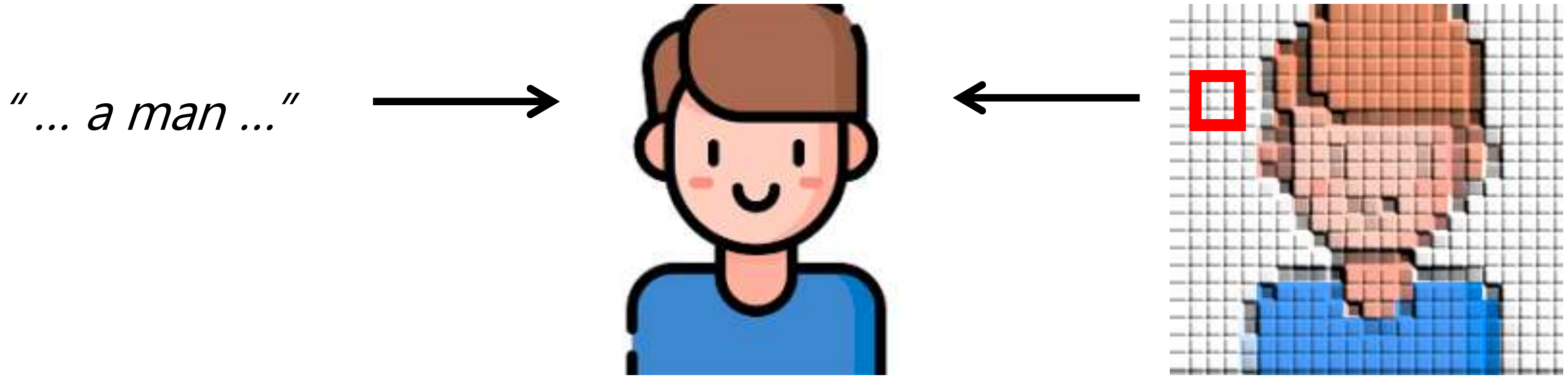
[4]



Problem Definition

Text-to-Image vs Image-to-Image

- Text-to-Image (T2I)
  - Visually interprets input text to generate images
  - Enables token-wise interpretation
- Image-to-Image (I2I)
  - Transforms reference images into different visual forms (e.g., inpainting)
  - Patch-wise interpretation is challenging → spatial and contextual continuity



Method

$\boldsymbol{I^2 AM}$: Image-to-Image **A**ttribution **M**aps method

- Visualization of the generation process using cross-attention maps
  - Text is abstract, but images retain spatial information in latent space
  - Patch-wise interpretation is difficult, yet generation can be visualized bidirectionally using image-domain features
- Objective
  - analyze the I2I latent diffusion models by time steps t, attention heads n, and layers l

Uni-directional visualization: Text $\longrightarrow$ Image

Bi-directional visualization: Image $\longleftrightarrow$ Image

Method

I^2AM: Image-to-Image Attribution Maps method

- Bi-directional attention scores
  - Reference-to-Generated(R2G) attention map: influence on generated image
  - Generated-to-Reference(G2R) attention map: contribution to generated image

R2G

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

soft-max in backward direction

G2R

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

soft-max in forward direction

$\boldsymbol{c_I}$: reference image embeddings

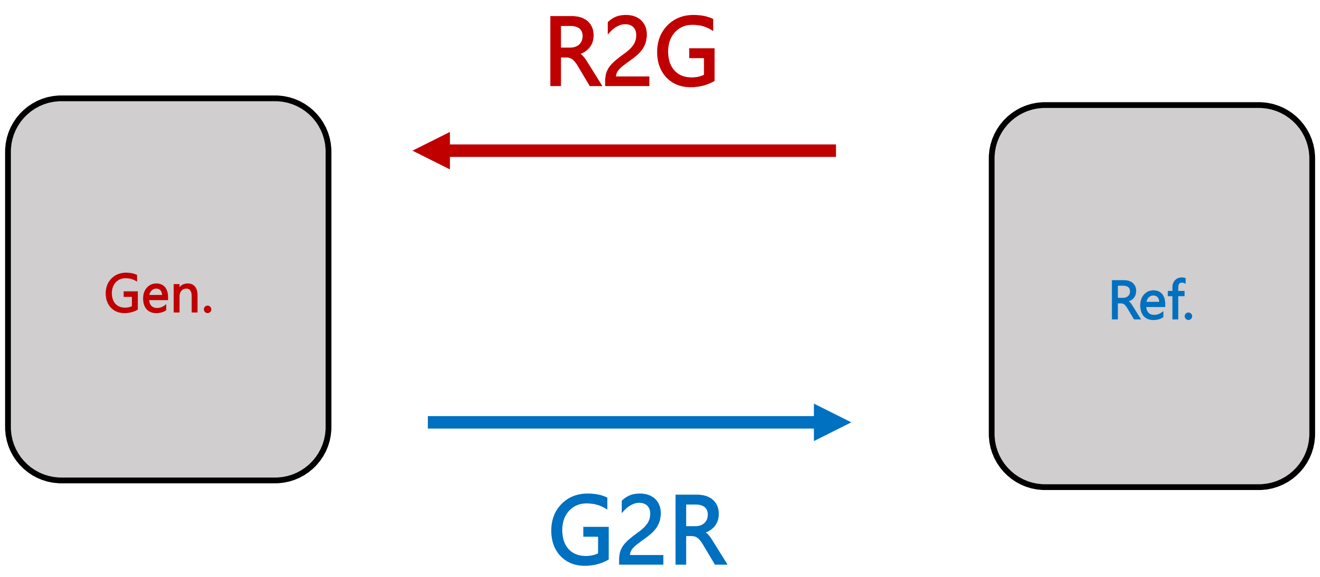
$\boldsymbol{f}_t^{(l)}$: pre-cross-attention vectors

$\boldsymbol{W}_k^{(l)}, \boldsymbol{W}_q^{(l)}$: projection matrices for queries and keys

Method

I^2AM: Image-to-Image Attribution Maps method

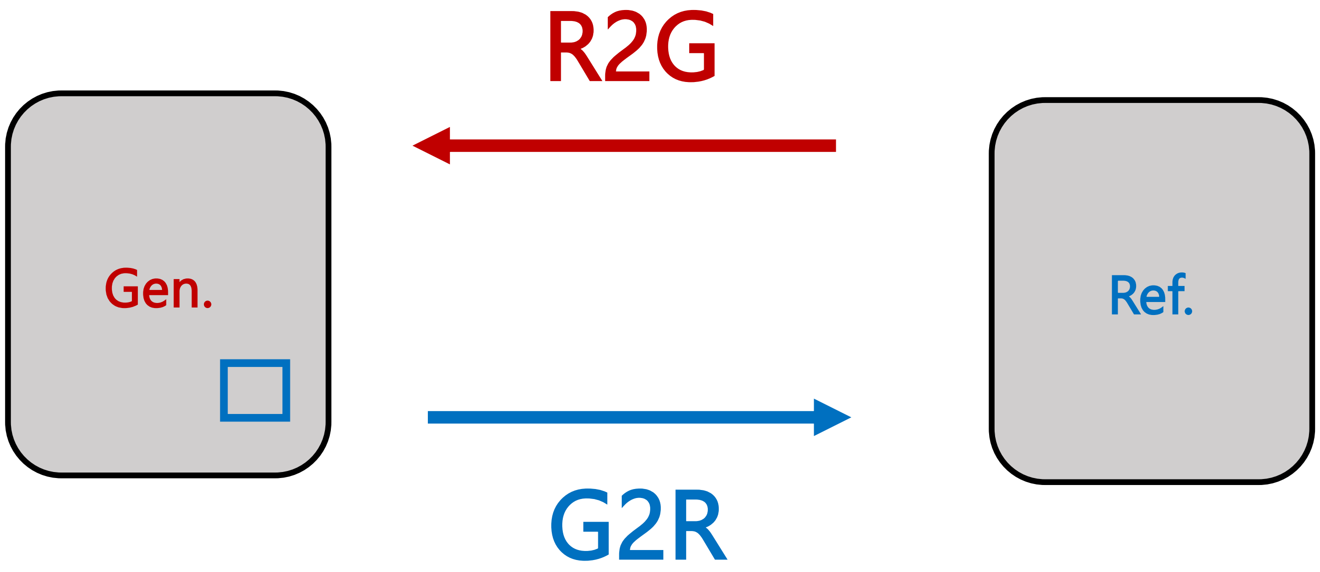
- Bi-directional attention scores
  - Reference-to-Generated(R2G) attention map: influence on generated image
  - Generated-to-Reference(G2R) attention map: contribution to generated image



Method

$\boldsymbol{I^2AM}$: Image-to-Image **A**ttribution **M**aps method

- Bi-directional attention scores
  - Reference-to-Generated(R2G) attention map: influence on generated image
  - Generated-to-Reference(G2R) attention map: contribution to generated image
    - ✓ SRAM: influence of reference image on a specific generated patch



Experiments

Setting

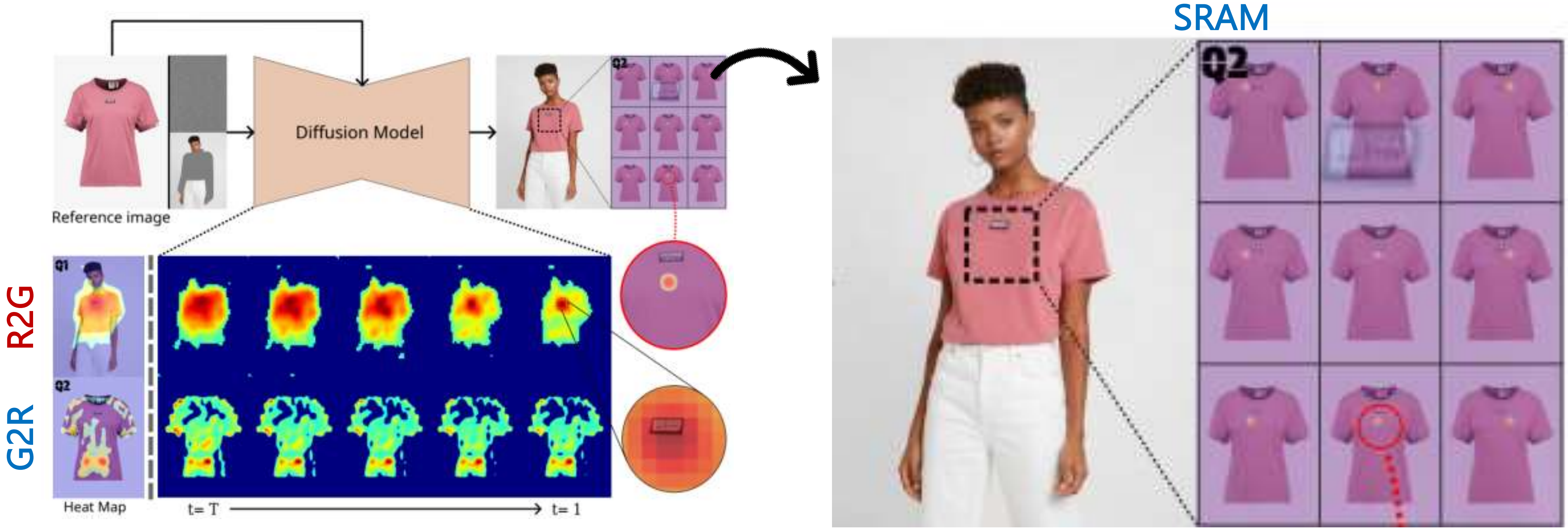
- Task
  - Image inpainting, specifically virtual try-on (VITON)
- Model
  - StableVITON [5] on VITON-HD [6]

[5]

Experiments

I^2AM: Image-to-Image Attribution Maps method

- Reference-to-Generated(R2G) attention map: influence on generated image
- Generated-to-Reference(G2R) attention map: contribution to generated image
    - SRAM: influence of reference image on a specific generated patch



Reference Papers

[1] Jiang, Peng-Tao, et al. "Layercam: Exploring hierarchical class activation maps for localization." *IEEE transactions on image processing* 30 (2021): 5875-5888.

[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] 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.

[6] Choi, Seunghwan, et al. "Viton-hd: High-resolution virtual try-on via misalignment-aware normalization." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021.