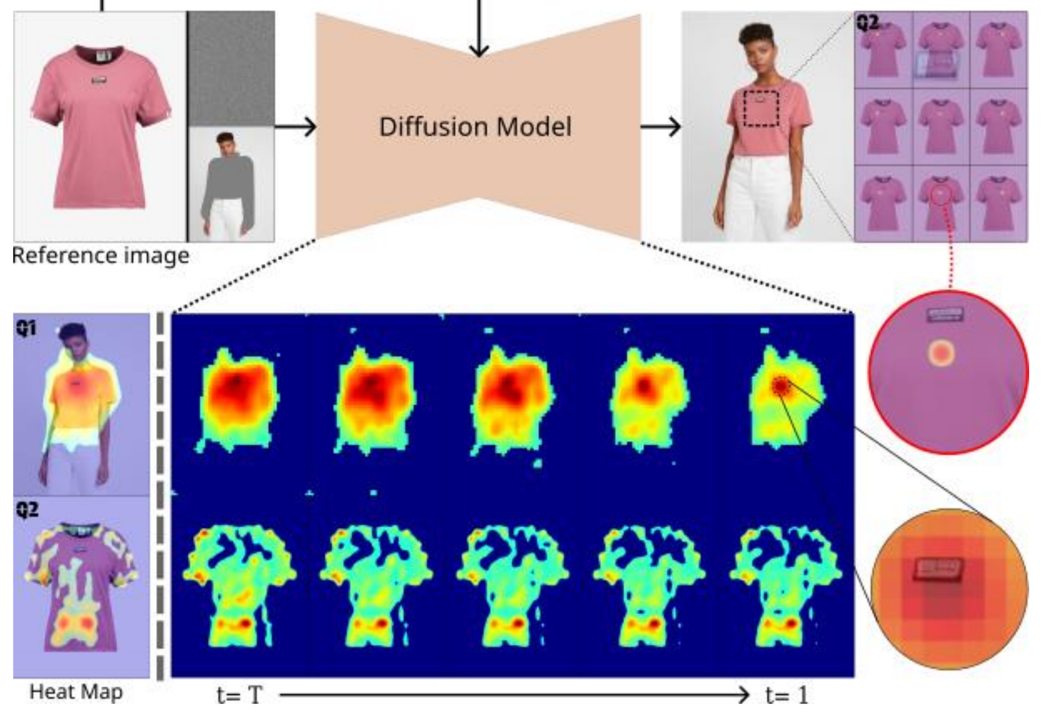
# $I^2AM$: Interpreting Image-To-Image Latent Diffusion Models via Bi-Attribution Maps

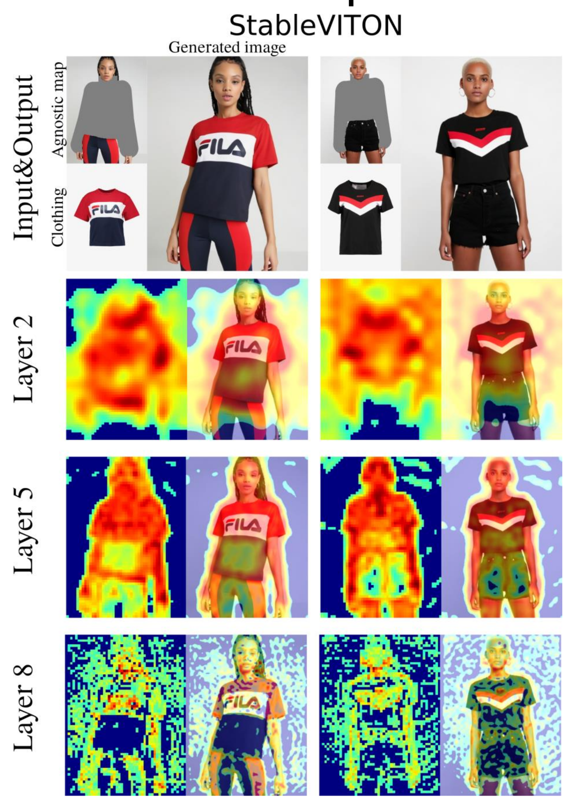
**Junseo Park and Hyeryung Jang**
Dongguk University, South Korea

dongguk UNIVERSITY    ICLR 2025

## Introduction

- Recent XAI studies on diffusion models focus on Text-to-Image (T2I) via cross-attention, while Image-to-Image **(I2I)** interpretability remains underexplored.




## Challenge

**Q.** **Can existing XAI methods for interpreting T2I models be directly applied?**

**A.** While text can be visualized at the token level through independent separation, images exhibit spatial and contextual continuity, making individual interpretation more challenging.

## Method

- **Bi-directional** attention scores
  - **Reference-to-Generated (R2G)** attention score: influence of reference patch
  - **Generated-to-Reference (G2R)** attention score: influence of generated patch

$$\mathrm{M}^{(l)}_{g,t,n} = \mathrm{Attn\_Score}(\mathbf{W}^{(l)}_{k,n}\mathbf{c}_{\mathbf{I}}, \mathbf{W}^{(l)}_{q,n}\mathbf{f}^{(l)}_{t})$$

$$\mathrm{M}^{(l)}_{r,t,n} = \mathrm{Attn\_Score}(\mathbf{W}^{(l)}_{q,n}\mathbf{f}^{(l)}_{t}, \mathbf{W}^{(l)}_{k,n}\mathbf{c}_{\mathbf{I}})$$

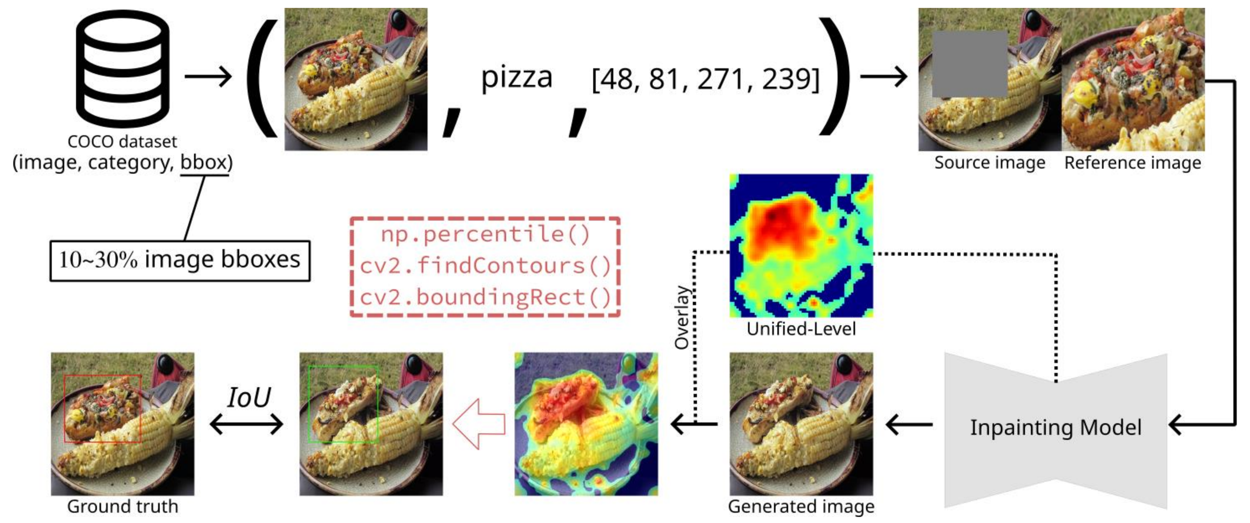
$c_I$: reference image embeddings
$f_t^{(l)}$: pre-cross-attention vectors
$W_{k,n}^{(l)}, W_{q,n}^{(l)}$ : projection matrices
$l$: cross-attention layer
$t$: time-step
$n$: attention head:

---

- We designed five types of attribution maps.
  - **Unified-level** attribution map: shows the overall generation flow
  - **Head-level** attribution map: displays score distribution for each attention head
  - **Time-level** attribution map: analyzes how the process changes over time
  - **Layer-level** attribution map: helps understand the role of each layer
  - **Specific-reference** attribution map: highlights the areas in the reference image that influenced specific patches in the generated image

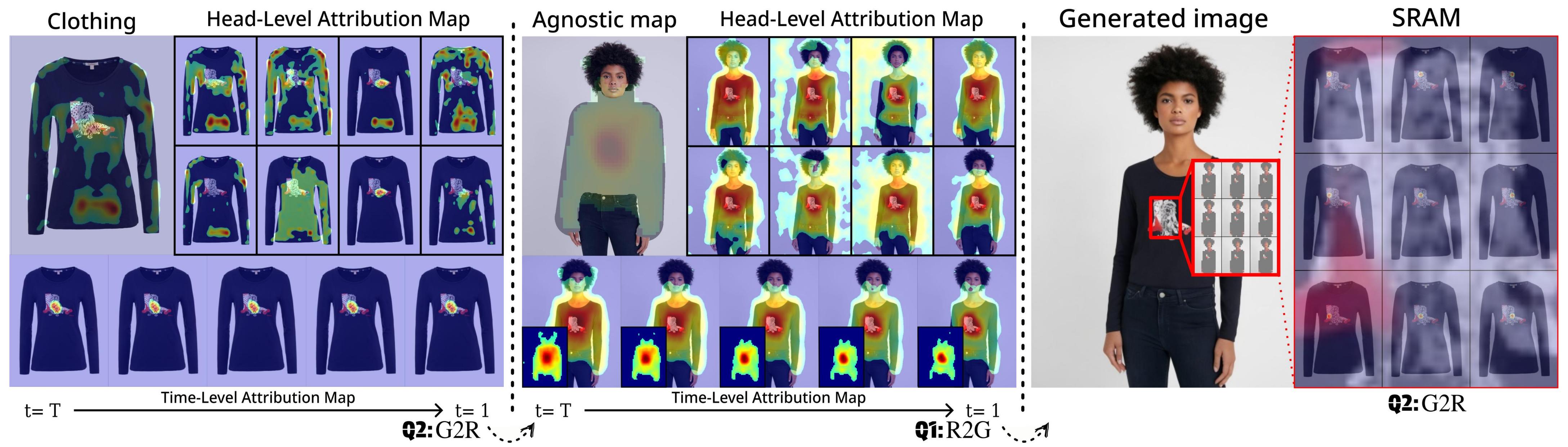
## Experiments

- We assess on object detection task how effectively $I^2AM$ captures and visualizes critical object features in both reference and generated images, even in unseen scenarios.



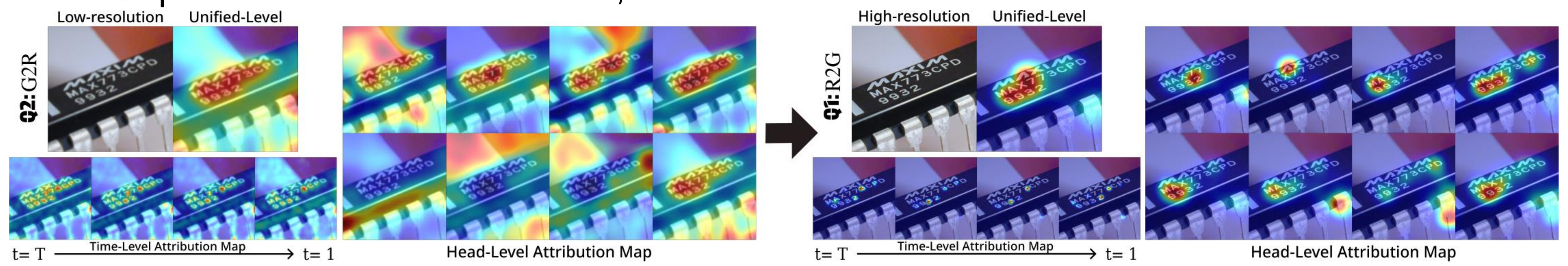

| Method | $\mathrm{mIoU}^{>0.5}_{GT}$ | $\mathrm{mIoU}^{>0.5}_{gen}$ |
|---|---|---|
| Supervised manner & Seen dataset | | |
| Faster-RCNN Ren et al. (2016) | 0.3225 | 0.2658 |
| Mask-RCNN He et al. (2017) | 0.3294 | 0.2706 |
| YOLOv3 Redmon (2018) | 0.2448 | 0.1978 |
| MaskFormer Cheng et al. (2022) | **0.3568** | **0.2932** |
| RTMDet Lyu et al. (2022) | 0.3228 | 0.2516 |
| Unsupervised manner & Unseen dataset | | |
| DAAM Tang et al. (2022) | 0.1807 | |
| Overall | 0.1807 | |
| $\mathrm{Random}_{10\sim30\%}$ | 0.2028 | |
| Ours | – | 0.2416 |

- Inpainting model: Paint-by-Example, DCI-VTON, StableVITON
- Bidirectional maps illustrate how reference details, such as clothing patterns and textures, transfer to the generated image.




---

- Super-resolution model: PASD, SeeSR




- We applied our method to model debugging and refinement.
  - Downstream metrics: FID, KID, LPIPS, SSIM
  - We trained a custom model, found that attention score variance caused inconsistent colors, and applied a new loss function to ensure stable attention and improve performance

| Method | FID ↓ | KID ↓ | LPIPS ↓ | SSIM ↑ |
|---|---|---|---|---|
| DCI-VTON Gou et al. (2023) | 13.0953 | 0.0334 | <u>0.0824</u> | 0.8612 |
| StableVITONKim et al. (2023a) | **10.6755** | 0.0064 | **0.0817** | <u>0.8634</u> |
| Custom | 11.6572 | <u>0.0042</u> | 0.1020 | 0.8396 |
| Refined custom | <u>11.5420</u> | **0.0022** | 0.0964 | **0.8644** |


Specific-Reference Attribution Map

- Comparison with T2I approach
  - Generate uninterpretable maps


## Conclusion

- We propose a method using cross-attention maps to analyze image-to-image latent diffusion models.
- $I^2AM$ produces two attribution maps: one capturing the reference image's influence on the generated image **(R2G)** and another tracing the generated image back to the reference **(G2R)**.
- Experiments on object detection, inpainting, and super-resolution demonstrate that $I^2AM$ enhances interpretability, identifies critical attribution patterns, and provides valuable insights for debugging and refinement.