$\boldsymbol{I^2 AM}$: Interpreting Image-To-Image Latent Diffusion Models via Bi-Attribution Maps

Junseo Park and Hyeryung Jang
Dongguk University, South Korea
{mki730, hyeryung.jang}@dgu.ac.kr

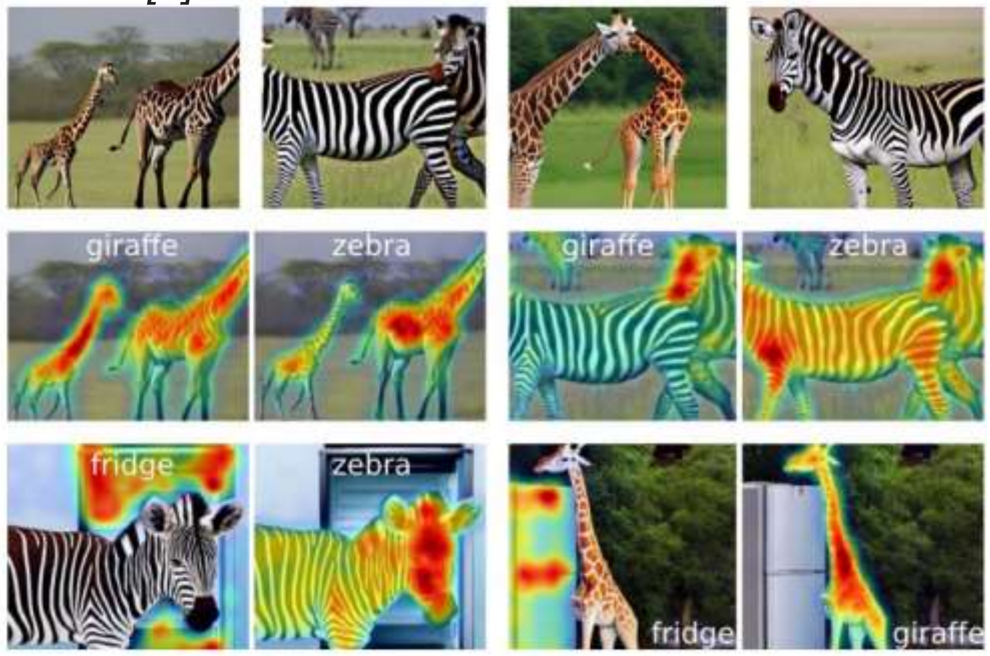
dongguk UNIVERSITY

ICLR 2025

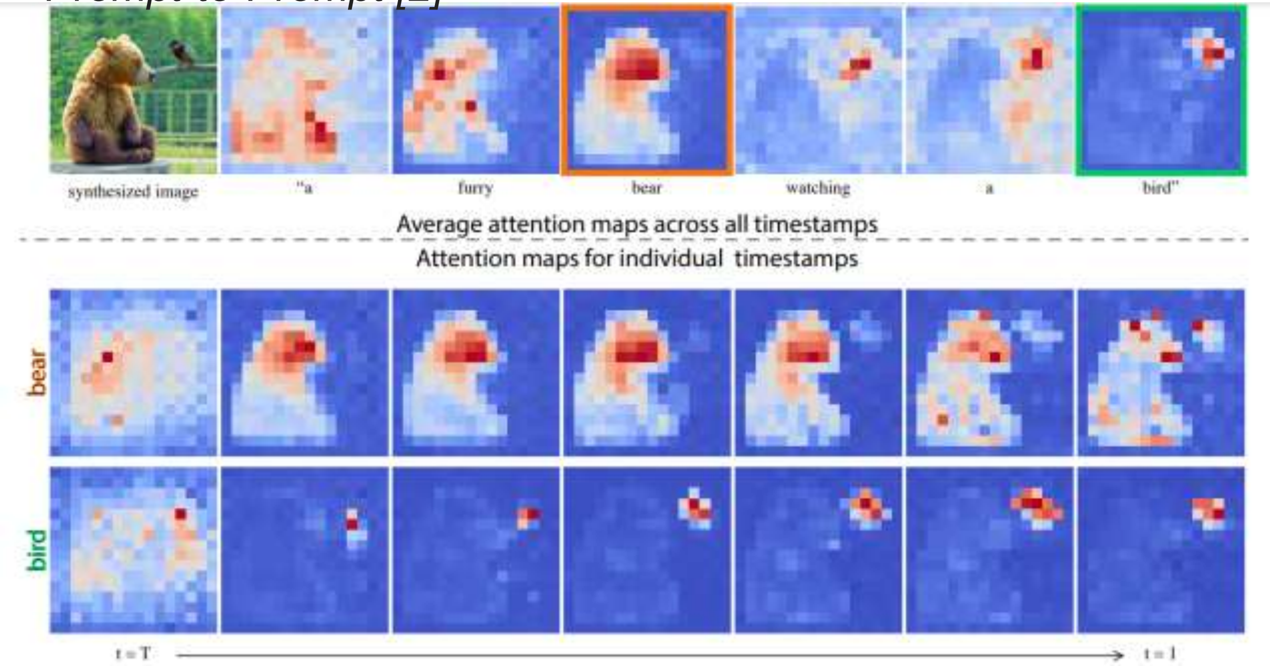
INTRODUCTION

- Recent XAI efforts on diffusion models have largely focused on text-to-image **(T2I)** models via cross-attention maps between text token and generated image patch

DAAM [1]

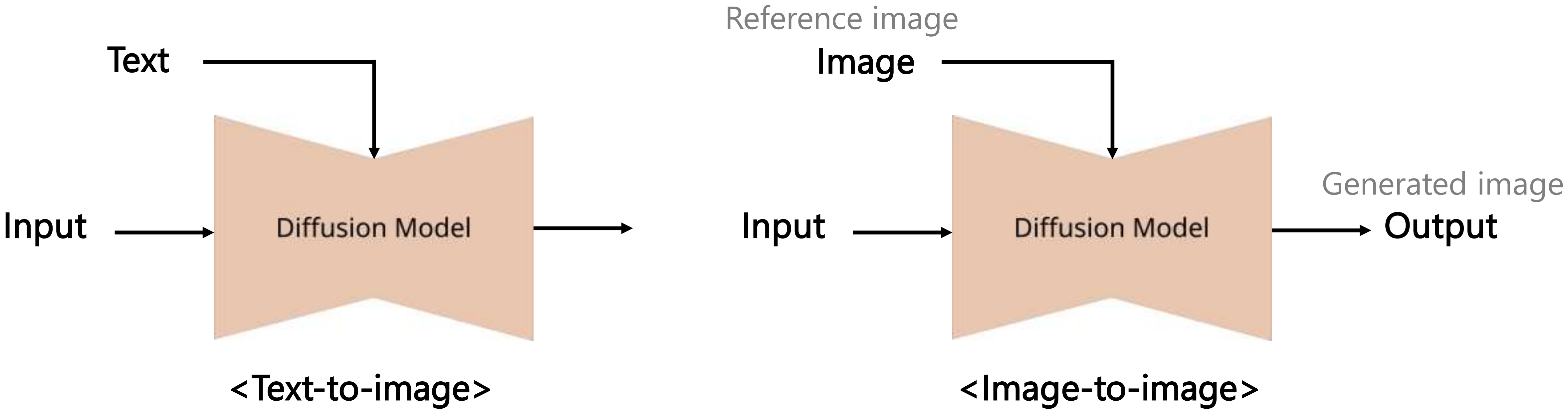


Prompt-to-Prompt [2]



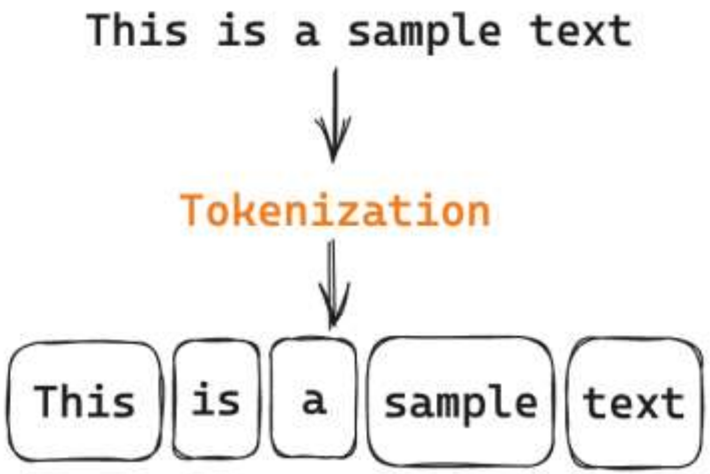
INTRODUCTION

- The interpretability in image-to-image **(I2I)** diffusion models remains underexplored

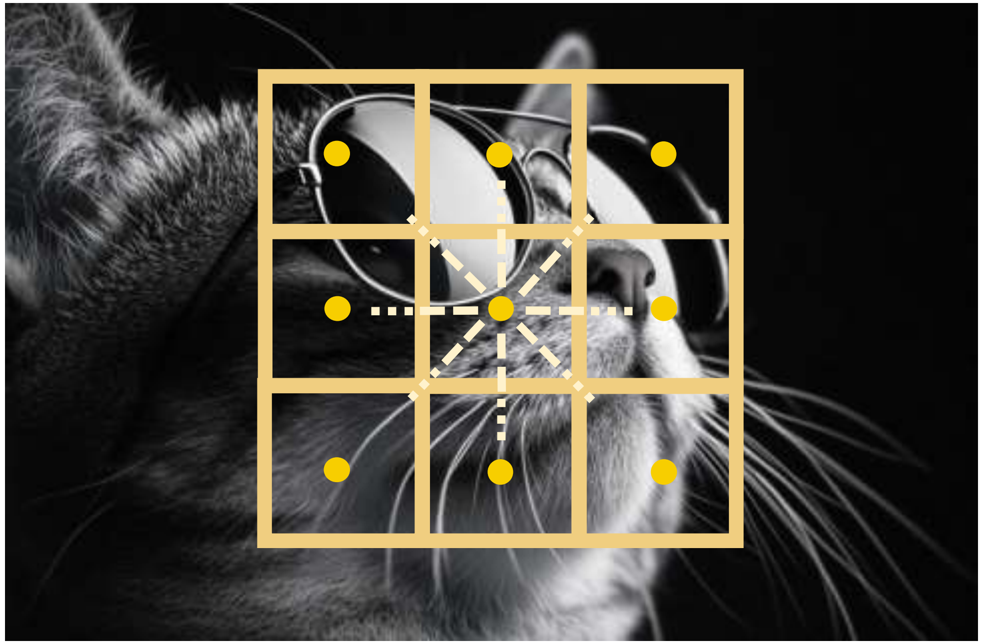


CHALLENGING ISSUES

- Text-to-Image (T2I): independent separation of text (tokenization)
- Image-to-Image (I2I): spatial and contextual continuity of reference image



VS



METHOD: I^{2}AM

- The shared image domain between reference and generated images

Uni-directional visualization:	Text	$\longrightarrow$	Image
Bi-directional visualization:	Image	$\longleftrightarrow$	Image

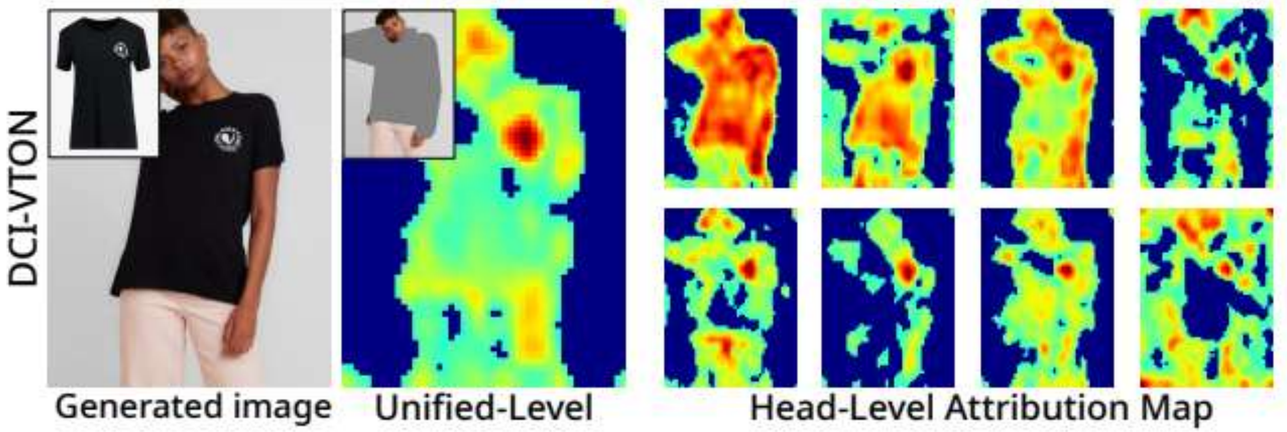
- **Bi-directional** attention scores
  - Reference-to-Generated attention score $\mathrm{M}^{(l)}_{\mathsf{g},t,n}$ luence of reference patch
  - Generated-to-Reference attention score $\mathrm{M}^{(l)}_{\mathsf{r},t,n}$ luence of generated patch

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

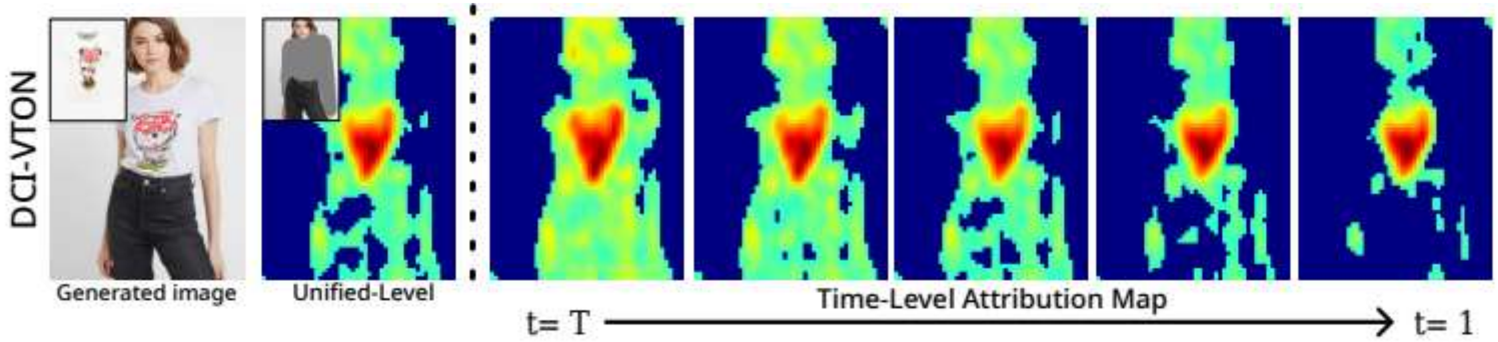
$\boldsymbol{c}_{\boldsymbol{I}}$: reference image embeddings
$\boldsymbol{f}^{(l)}_{t}$: pre-cross-attention vectors
$\boldsymbol{W}^{(l)}_{k,n}, \boldsymbol{W}^{(l)}_{q,n}$: projection matrices for queries and keys

METHOD: I²AM

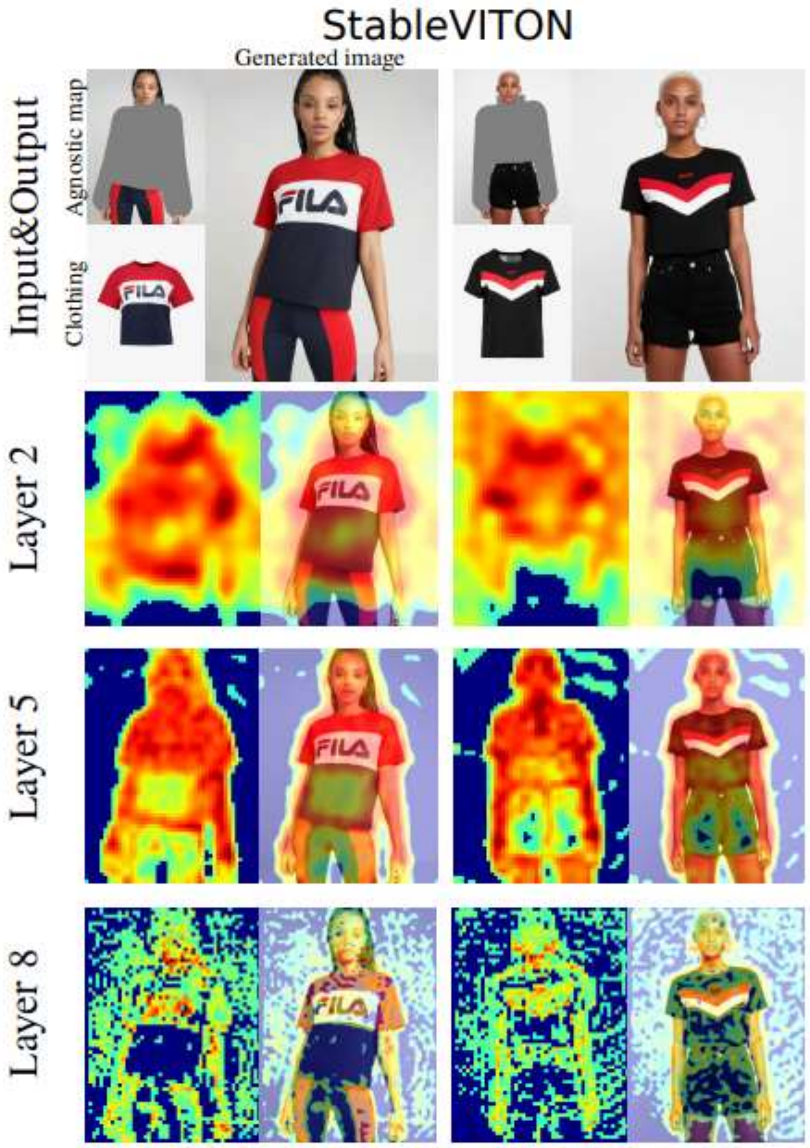
- Attribution maps tailored for diffusion models
- Unified / Layer / Head / Time -level attribution maps



Head-Level attribution maps



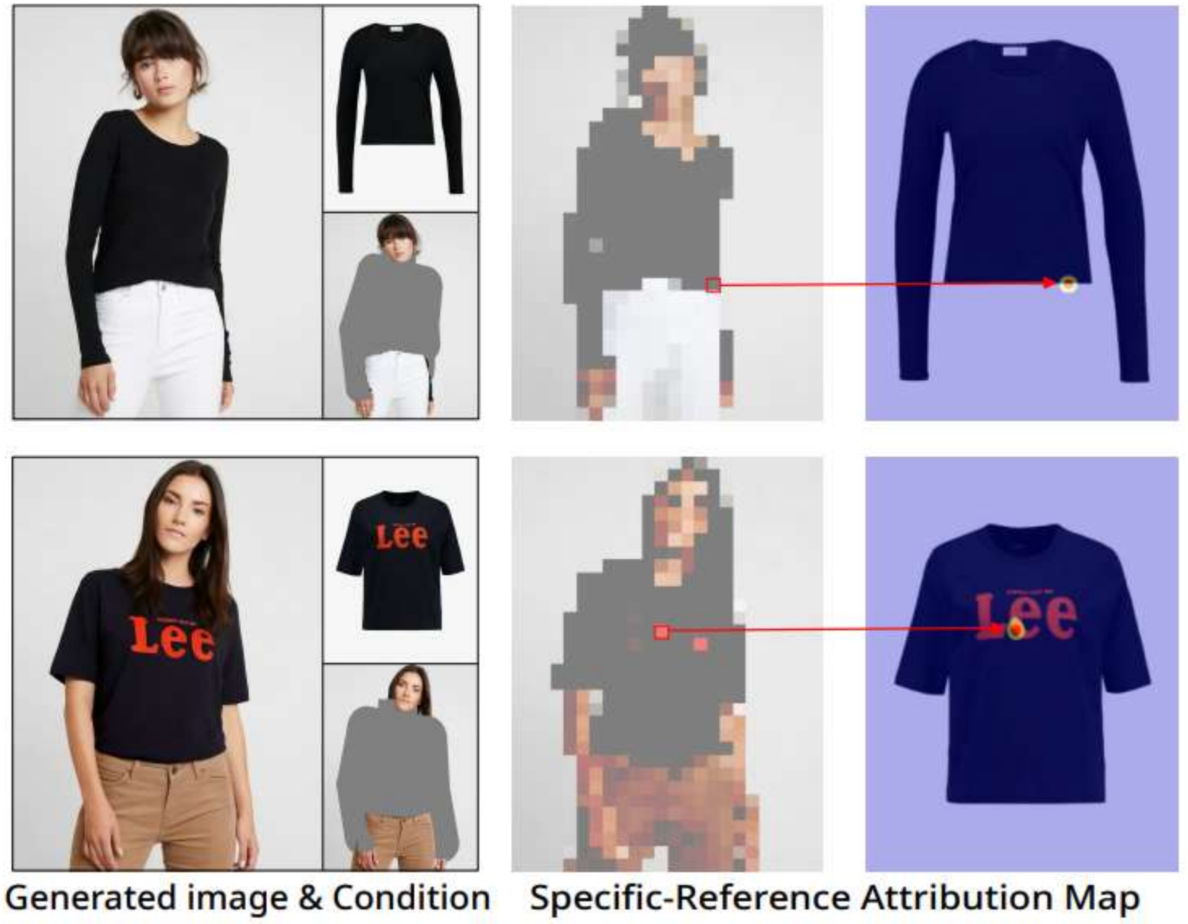
Time-Level attribution maps



Layer-Level attribution maps

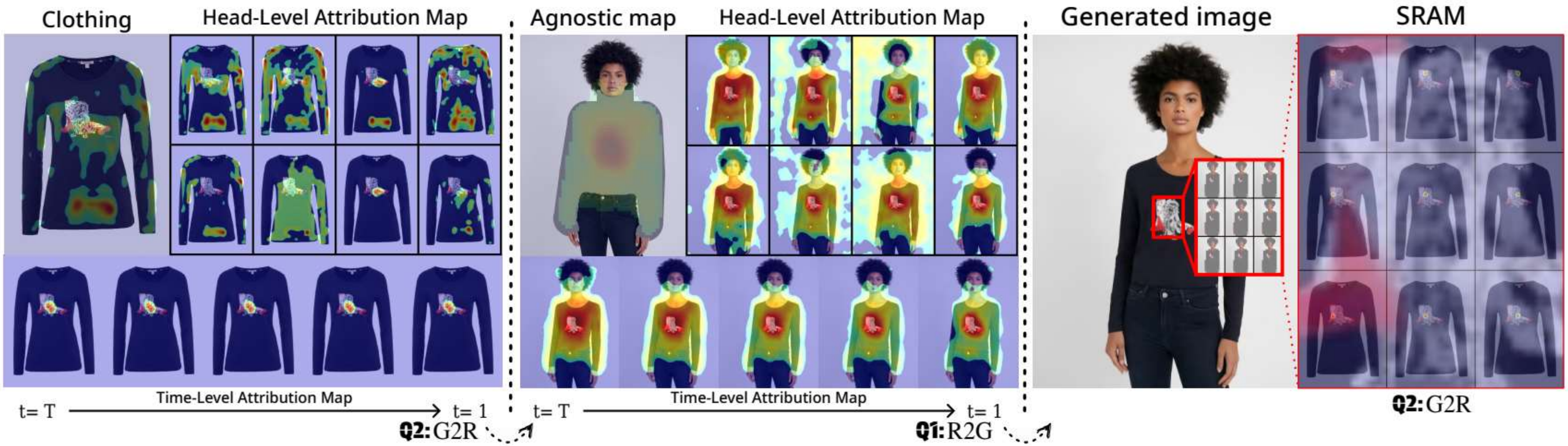
METHOD : I²AM

- Specific-reference attribution map



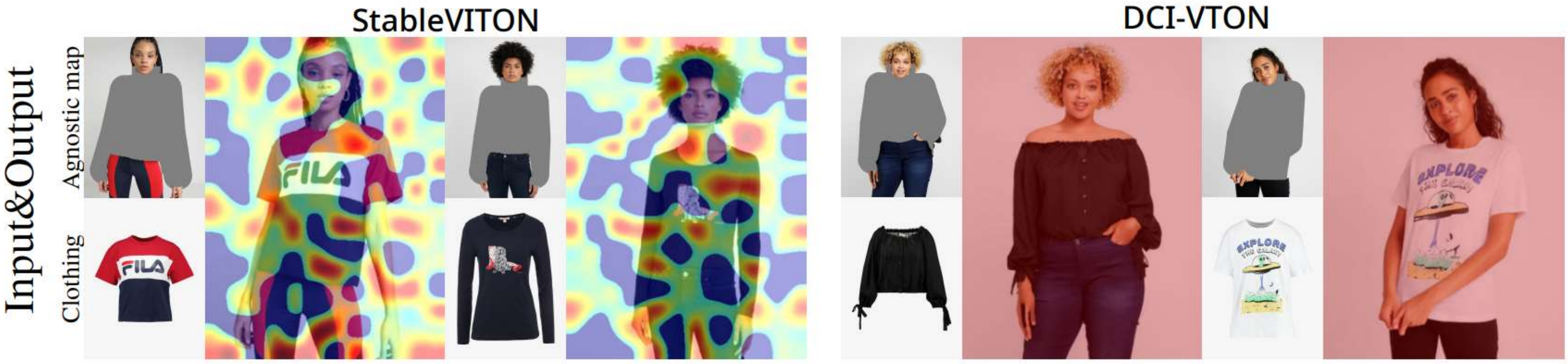
EXPERIMENTS

- Task: inpainting and super-resolution tasks
- Model
  - Inpainting: Paint-by-Example *[3]*, DCI-VTON *[4]*, StableVITON *[5]*
  - Super-resolution: PASD *[6]*, SeeSR *[7]*



EXPERIMENTS

- Comparison with other T2I attention-based method (DAAM *[1]*)



(a) All patch embeddings

(b) Only CLS embedding

EXPERIMENTS

- Model debugging and refinement
  - Utilize **I²AM** to analyze attention alignment in custom model
  - Refine custom model for better consistency and performance

Specific-Reference Attribution Map

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

References

Reference papers

[1] Tang, Raphael, et al. "What the daam: Interpreting stable diffusion using cross attention.", arxiv 2022.

[2] Hertz, Amir, et al. "Prompt-to-prompt image editing with cross attention control.", arxiv 2022.

[3] 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.

[4] 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.

[5] Kim, Jeongho, et al. "Stableviton: Learning semantic correspondence with latent diffusion model for virtual try-on." *CVPR* 2024.

[6] Yang, Tao, et al. "Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization." *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 2024.

[7] Wu, Rongyuan, et al. "Seesr: Towards semantics-aware real-world image super-resolution." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2024.

THANK YOU.