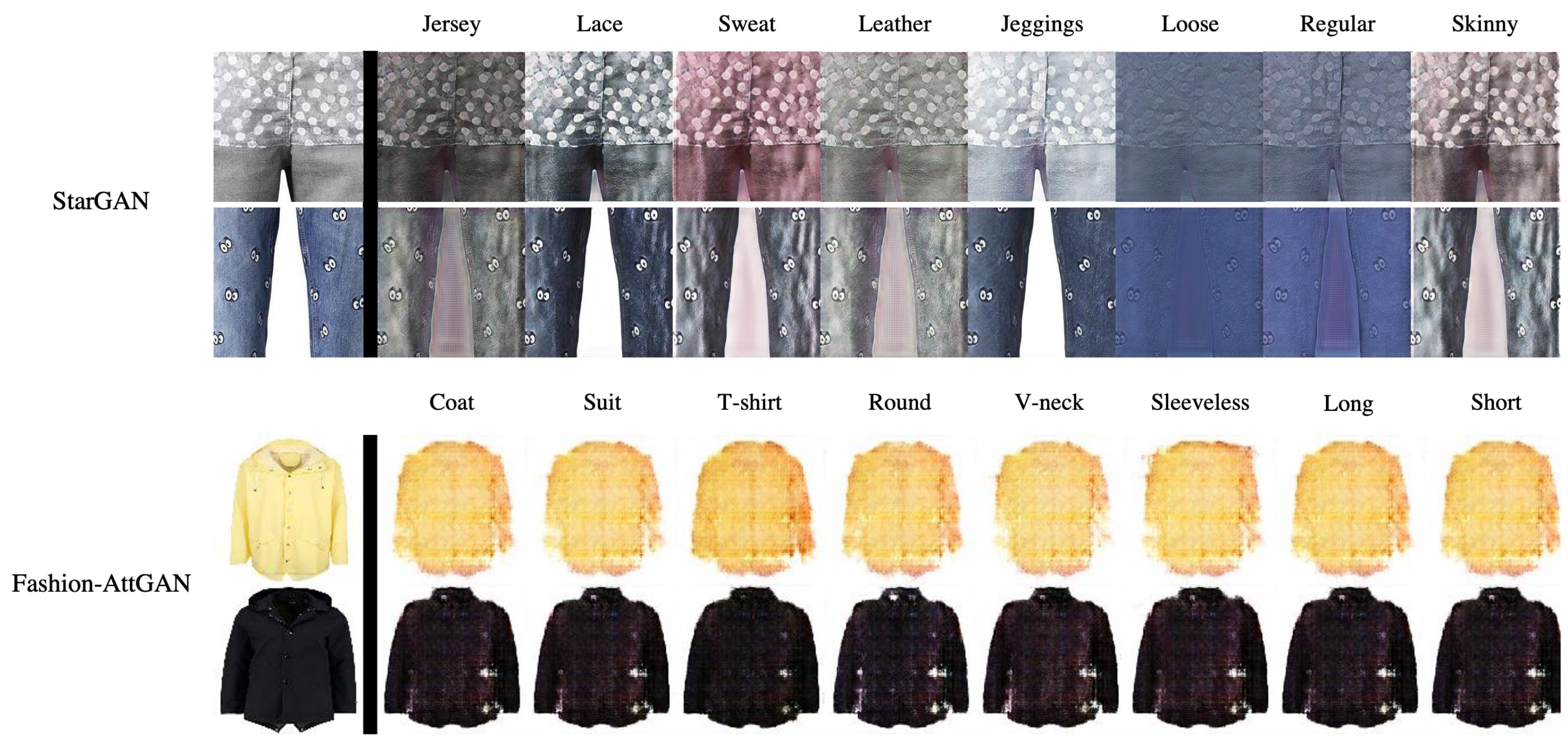
Leveraging Off-the-shelf Diffusion Model for Multi-attribute Fashion Image Manipulation

Chaerin Kong, Donghyeon Jean, Ohjoon Kwon, Nojun Kwak
Seoul National University, NAVER

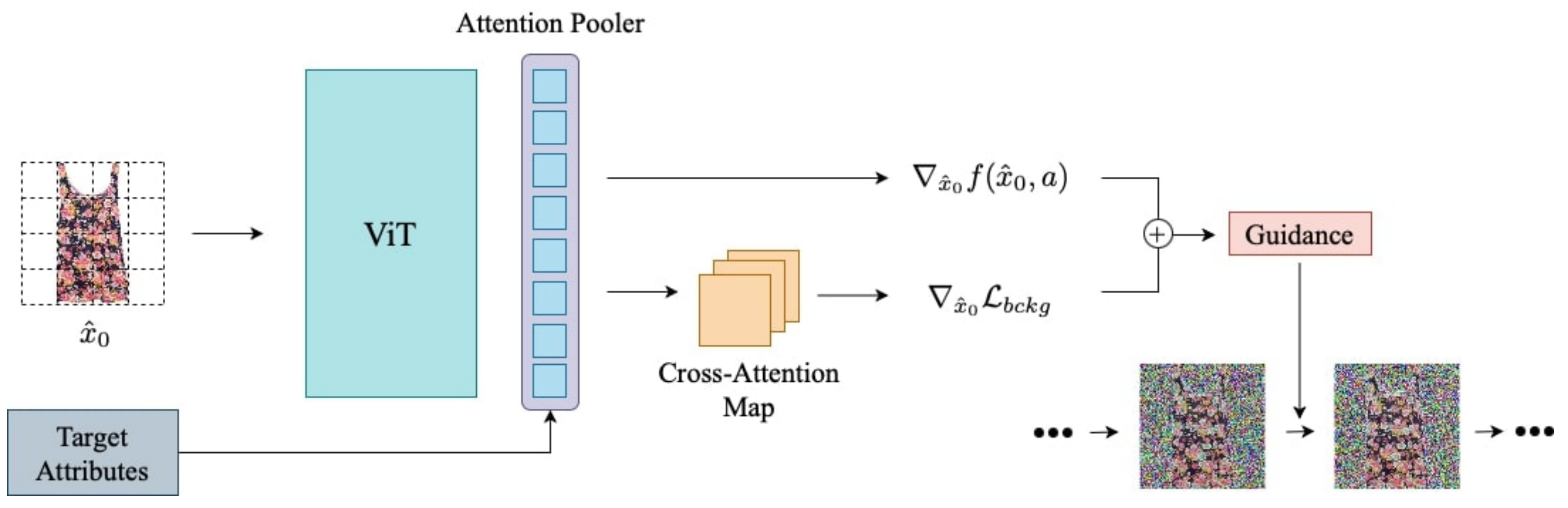
WACV WAIKOLOA HAWAII JAN 3-7 • 2023

Motivation

- GAN-based attribute-editing models only support few attributes and lack scalability.
- Diffusion models have demonstrated superior performances both in sample quality and diversity.
- While properly annotated fashion dataset is rare, numerous generic diffusion models have their pretrained checkpoints publicly available.
- With limited data, training a classifier is generally easier than training a generative model.



Formulation



Pretrained ViT + Attention Pooler

- The [cls] token of a pretrained ViT only reasons about the global semantics of an image.
- For multi-attribute editing, our classifier should attend to different local regions of an image for each attribute.

Local Image Editing with Patch-level Attention

- Use the cross-attention map for each attribute token to identify the salient region (and the background).
- Impose typical classifier guidance as well as background preservation loss to only modify the relevant area.

Evaluation

Empirical Findings from ViT Adaptation

		Category	Fabric	Sleeve Length	Pattern	Gender	Fit	Collar	Neckline	Average
Random Init.										
	End-to-End	30.1	56.6	51.8	50.3	66.1	57.1	31.0	65.1	51.0
Imagenet-pretrained										
	Attention-Pool Only	85.4	58.7	84.8	76.1	95.0	66.4	**91.4**	**84.4**	80.3
	Last2	67.0	52.8	78.9	48.5	74.6	59.7	81.6	78.1	67.6
	Last4	44.8	52.8	60.8	45.0	82.0	58.6	76.1	76.2	62.0
	Last6	43.1	52.1	73.0	44.0	77.4	54.9	76.0	71.1	61.5
CLIP-pretrained										
	Attention-Pool Only	86.3	60.2	**84.9**	80.4	95.8	69.9	91.0	83.3	**81.5**
	Last6	**87.2**	**67.9**	84.4	**87.2**	**97.4**	**70.9**	75.3	78.0	81.0
	Last12	85.6	67.8	83.2	86.4	97.1	69.9	78.5	76.5	80.6
	Last18	82.3	66.8	79.7	83.2	95.9	68.4	72.9	73.7	77.8
	Last24	51.7	61.9	70.4	61.5	84.7	61.9	41.6	65.1	62.3

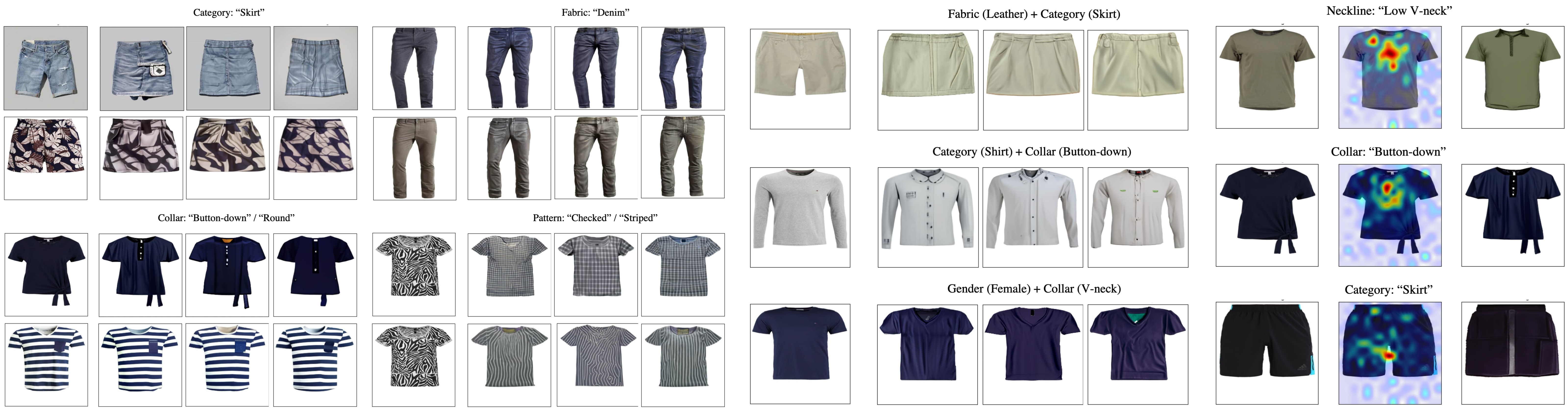
		Category	Fabric	Sleeve Length	Pattern	Gender	Fit	Collar	Neckline	Average
Imagenet-pretrained										
	No Aug.	**85.4**	**58.7**	**84.8**	**76.1**	**95.0**	**66.4**	**91.4**	**84.4**	**80.3**
	Random Aug.	81.5	57.3	83.3	73.7	92.7	65.2	91.2	82.0	78.4
CLIP-pretrained										
	No Aug.	86.3	59.8	**85.5**	79.6	**96.6**	67.7	89.5	82.5	80.9
	Random Aug.	86.3	**60.2**	84.9	**80.4**	95.8	**69.9**	**91.0**	**83.3**	**81.5**

Rich Attribute Set

Attribute	#Classes	Class
Category	16	Coat, Jacket, Suit, Shirt, T-shirt, Jumper, Shorts Trouser, Jean, Swimming, Jumpsuit, Pyjamas, Tracksuit, Bottoms, Tracksuit, Skirt, Dress
Collar	17	Buttondown, Cutaway, High, Hood, Kent, Lapel, Lined, Mandarin, Polo, Round, Shawl, Turndown, V-neck, Peter Pan, Volant, Shirt, Chin
Fabric	14	Canvas, Crocheted, Denim, Fleece, Hardshell, Jersey, Jersey Lace, Lace, Mesh, Mesh Jersey, Rib, Softshell, Sweat, Leather
Fit	15	Bootcut, Flared, High waist, Jeggings, Large, Loose, Low, Oversize, Regular, Skinny, Slim, Small, Straight, Tailered, Tapered
Gender	2	Male, Female
Neckline	11	Boat, Backless, Cache-coeur, Henley, Low v-neck, Low round, Off-the-shoulder, Round, Square, V-neck, Envelope
Pattern	16	Animal, Burnout, Camouflage, Checked, Marl, Color gradient, Colorful, Floral, Herringbone, Paisley, Photo, Pinstriped, Plain, Polka dot, Print, Striped
Sleeve Length	9	3/4, Spaghetti, Sleeveless, Elbow, Extra long, Extra short, Long, Short, Strapless

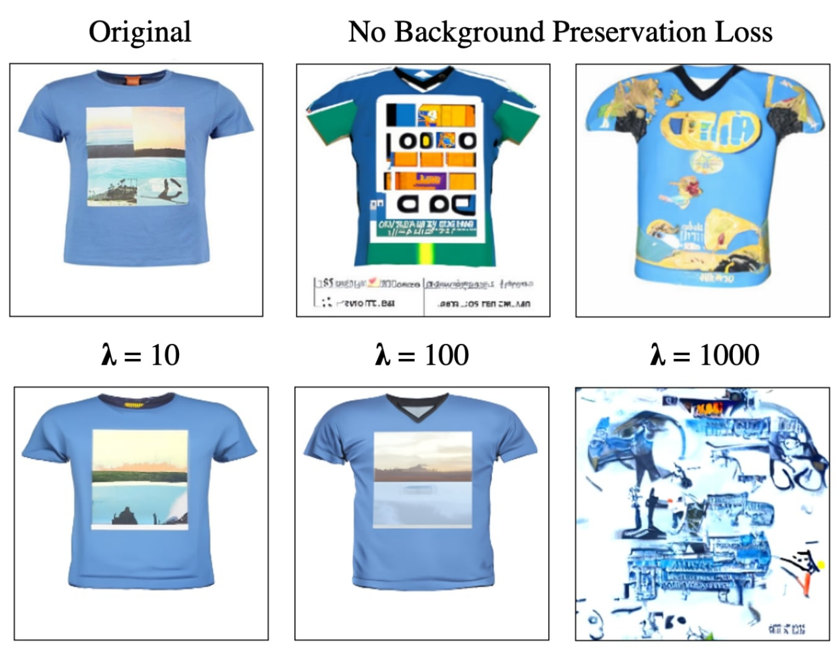
- With a relatively small (~100k) and visually distant dataset, training the attention pooler alone performs the best.

Qualitative Evaluations

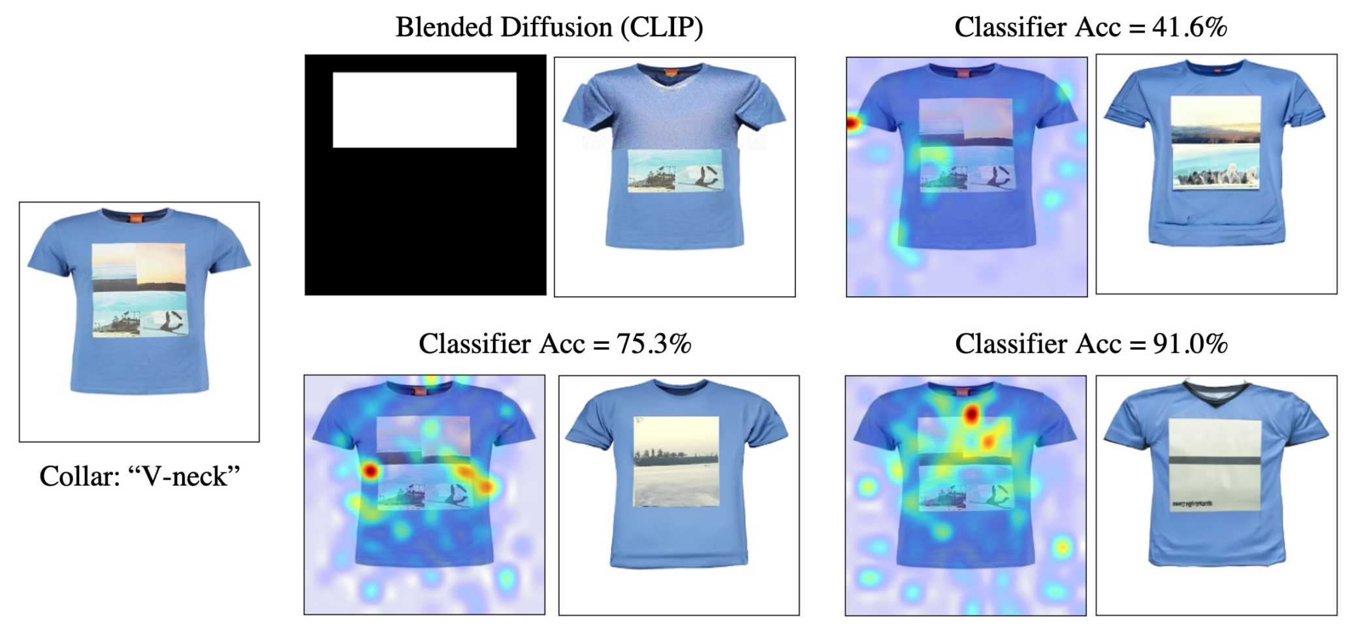

Category: "Skirt"
Fabric: "Denim"
Collar: "Button-down" / "Round"
Pattern: "Checked" / "Striped"
Fabric (Leather) + Category (Skirt)
Category (Shirt) + Collar (Button-down)
Gender (Female) + Collar (V-neck)
Neckline: "Low V-neck"
Collar: "Button-down"
Category: "Skirt"

- Our framework performs a wide range of attribute editing with a single model, producing realistic samples.

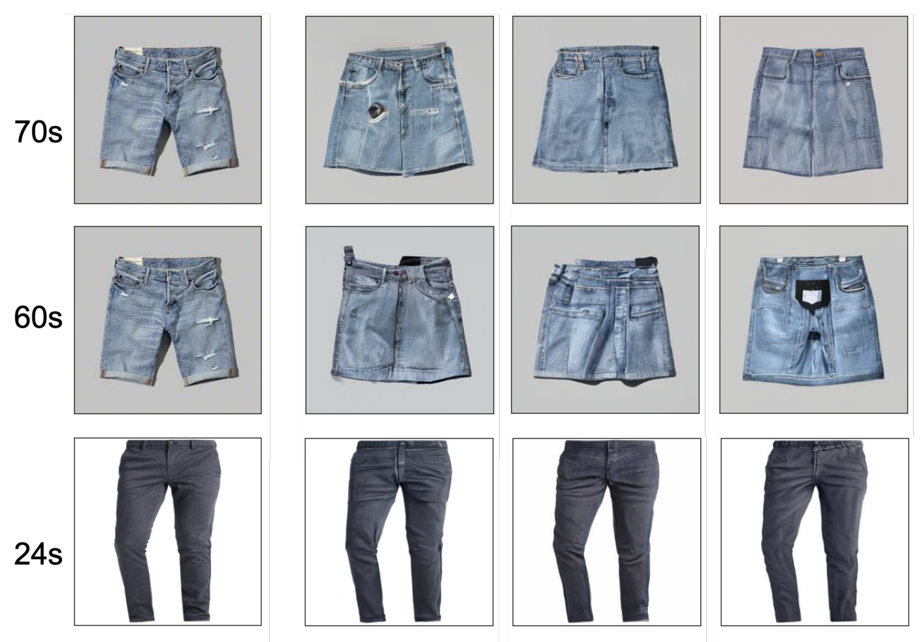
Ablations



Loss Weights



Classifier Performance



Inference Time (compute)

- Better classifier provides not only superior semantic guidance but also more accurate attention map.
- Simple operations (e.g., texture) are more robust to the number of diffusion steps compared to more complex modifications (e.g., shape).