Highlight

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

TUE-AM-254

FAME-ViL: Multi-Tasking Vision-Language Model for Heterogeneous Fashion Tasks

Xiao Han[1,2], Xiatian Zhu[1,3], Licheng Yu, Li Zhang[4], Yi-Zhe Song[1,2] and Tao Xiang[1,2]

1 CVSSP, University of Surrey

2 iFlyTek-Surrey Joint Research Centre on Artificial Intelligence

3 Surrey Institute for People-Centred Artificial Intelligence, University of Surrey

4 School of Data Science, Fudan University

UNIVERSITY OF SURREY

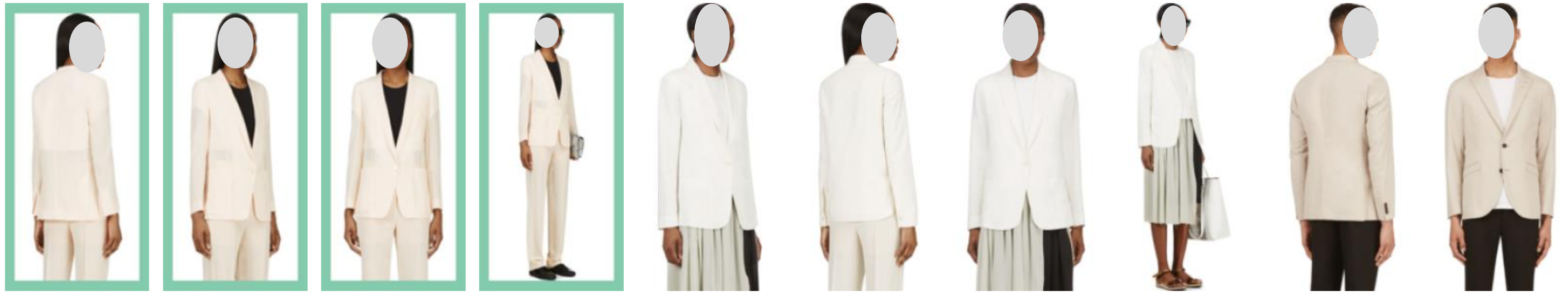
iFLYTEK

復旦大學
FUDAN UNIVERSITY

Introduction: Fashion Tasks

Cross-Modal Retrieval (XMR)

Text Query: Long sleeve relaxed-fit silk blazer in light peach. Shawl collar. Single-button closure and patch pockets at front. Breast pocket. Slits at sleeve cuffs. Vented at back.



Sub-Category Recognition (SCR)

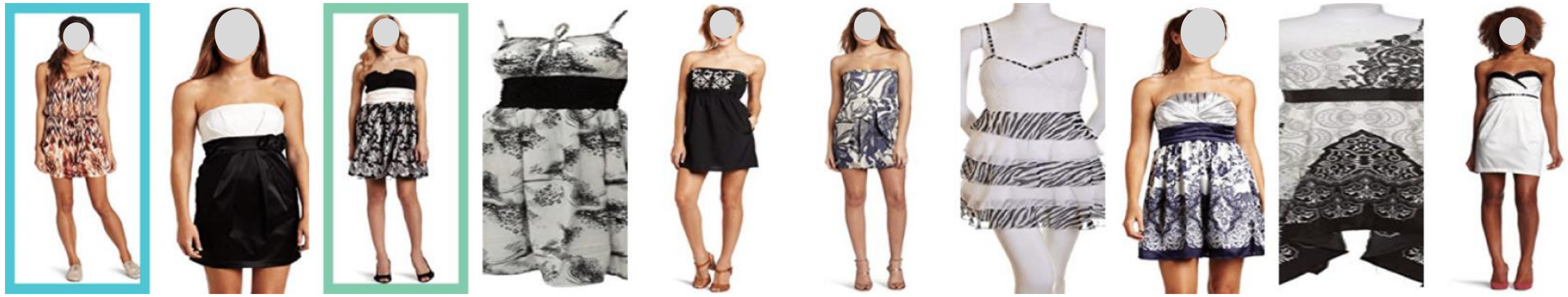
Slouchy lamb nubuck patrol hat in black. Wrinkling and light distressing throughout. Fully lined.

Predicted Class: [FLAT CAPS]

Text-Guided Image Retrieval (TGIR)

Reference Image

Modifying Text: is a black and white dress, is strapless



Fashion Image Captioning (FIC)

Generated Caption: Grey & brown camo print tank top. Relaxed-fit tank top in tones of grey, brown, and black. Signature snake graphic print throughout. Ribbed crewneck collar. Tonal stitching.

Introduction: Problems

1. **Heterogeneity in Fashion:**
    1. Different input and output formats
    2. Different dataset sizes as the annotation difficulty of each task differ
2. **Problems of previous pre-training then fine-tuning pipeline:**
    1. Low parameter efficiency (redundant storage and computation)
    2. Lack of utilization of inter-task relatedness

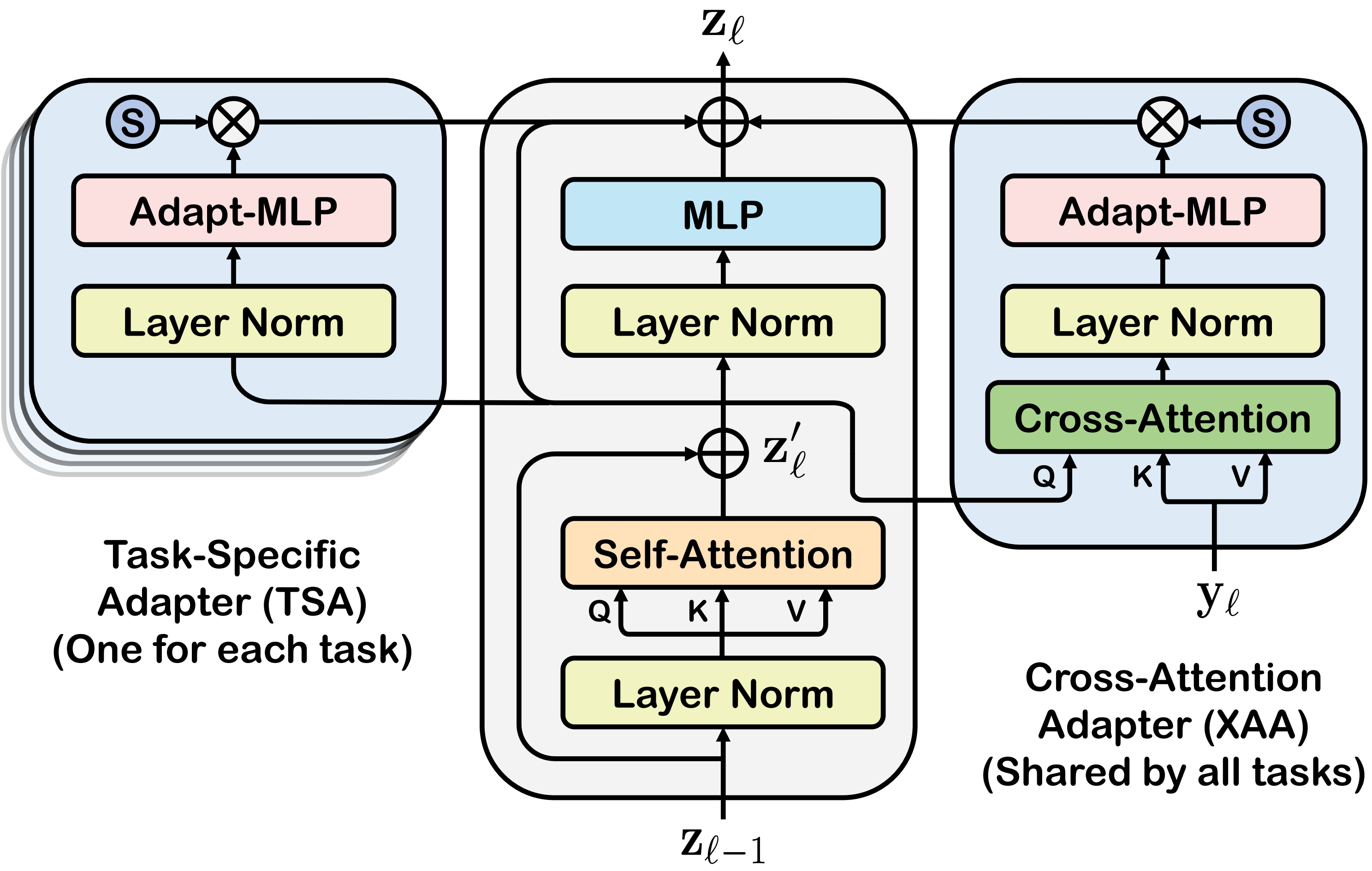
Contributions

1. **A task-versatile architecture** on top of CLIP with two novel lightweight adapters

2. **An efficient and effective multi-task training strategy** supporting heterogeneous task modes in one unified model

3. **SOTA performance** across 5 fashion downstream tasks **with 61.5% parameter saving**

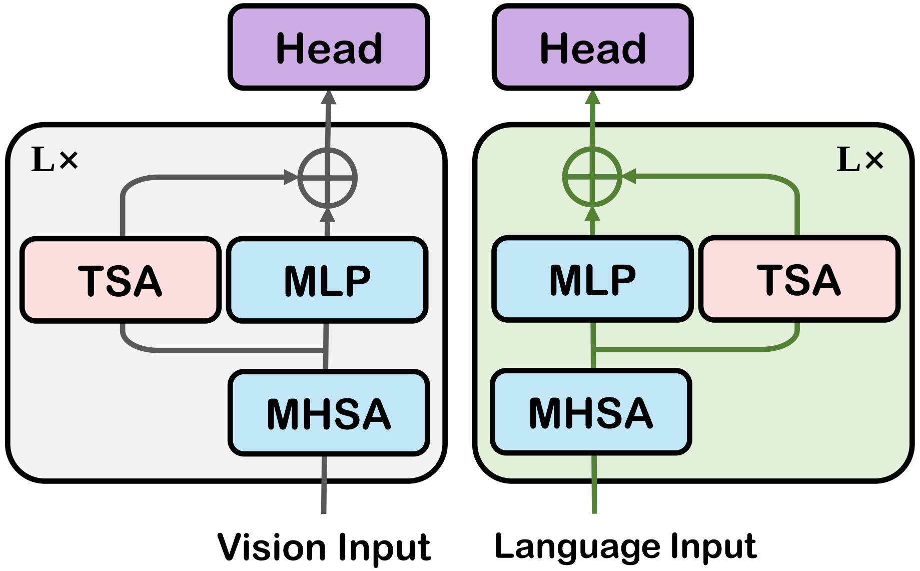
Task-versatile Transformer Layer



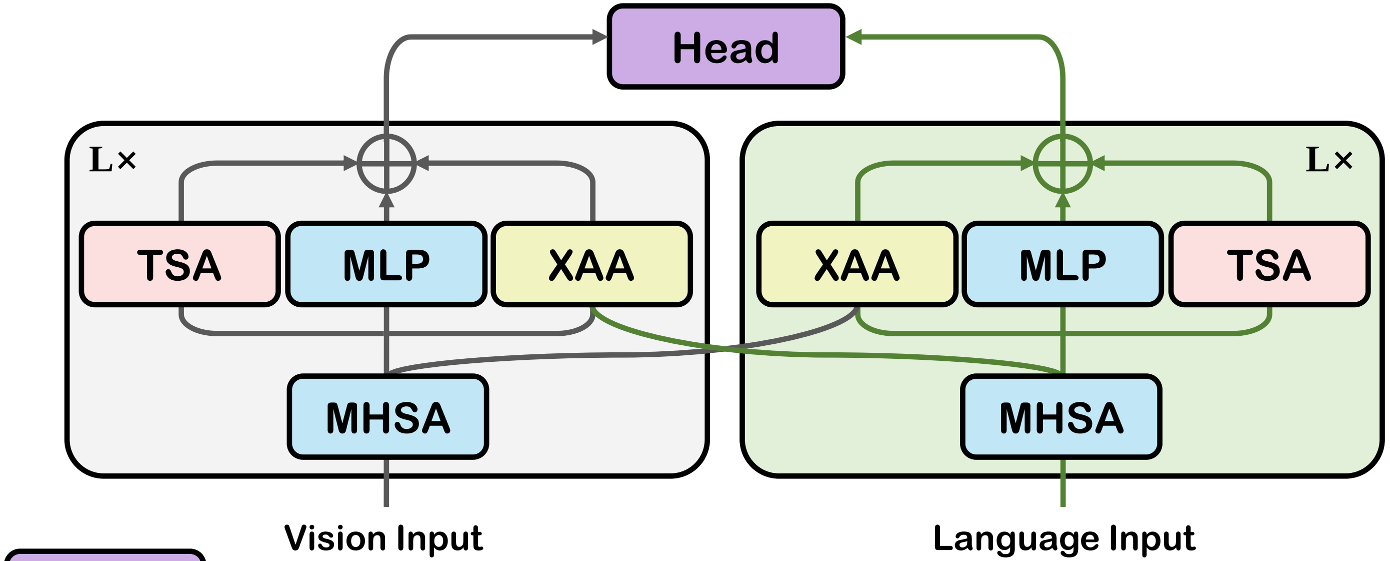
Task-Specific
Adapter (TSA)
(One for each task)

Cross-Attention
Adapter (XAA)
(Shared by all tasks)

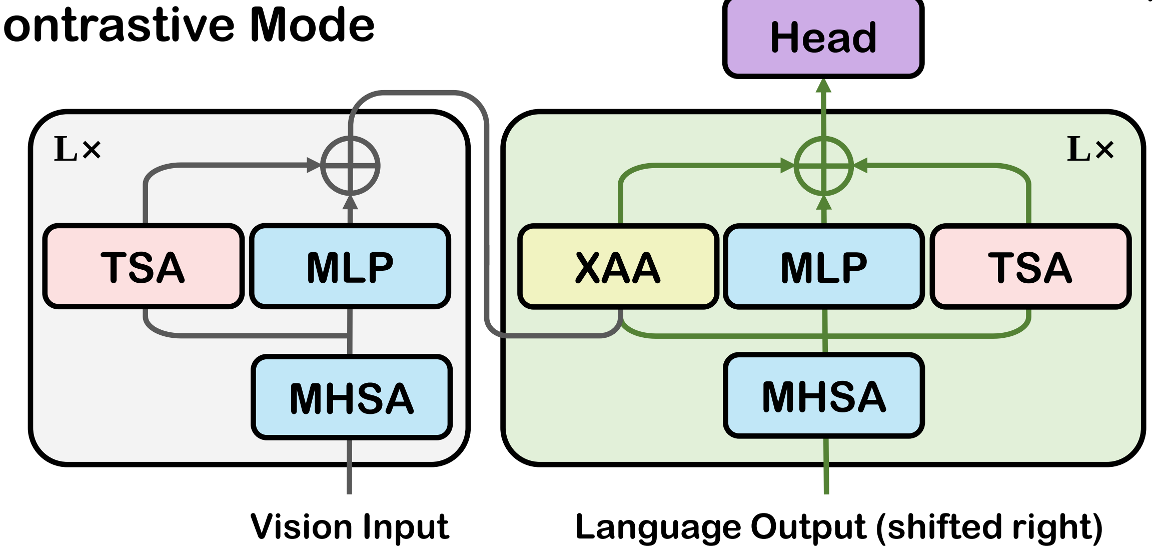
Three Operational Modes



(a) Contrastive Mode



(b) Fusion Mode



(c) Generative Mode

Multi-teacher Distillation

$$\mathcal{L}_{\text{XMR}} = \frac{1}{2}\left[\mathcal{L}_{\text{InfoNCE}}(\mathbf{T},\mathbf{I}) + \mathcal{L}_{\text{InfoNCE}}(\mathbf{I},\mathbf{T})\right]$$

$$\mathcal{L}^{\text{D}}_{\text{XMR}} = \frac{1}{2B}\sum_{b}^{B}\left(\text{KL}\left(\mathbf{s}_{b,\cdot} \,\|\, \tilde{\mathbf{s}}_{b,\cdot}\right) + \text{KL}\left(\mathbf{s}_{\cdot,b} \,\|\, \tilde{\mathbf{s}}_{\cdot,b}\right)\right)$$

$$\mathcal{L}_{\text{SCR}} = -\mathbb{E}_{(I,T)\sim D}\log P\left(f_{\theta}^{[f]}(I,T)\right), \mathcal{L}^{\text{D}}_{\text{SCR}} = \text{KL}\left(f_{\theta}^{[f]}(I,T) \,\|\, g_{\text{scr}}(I,T)\right)$$

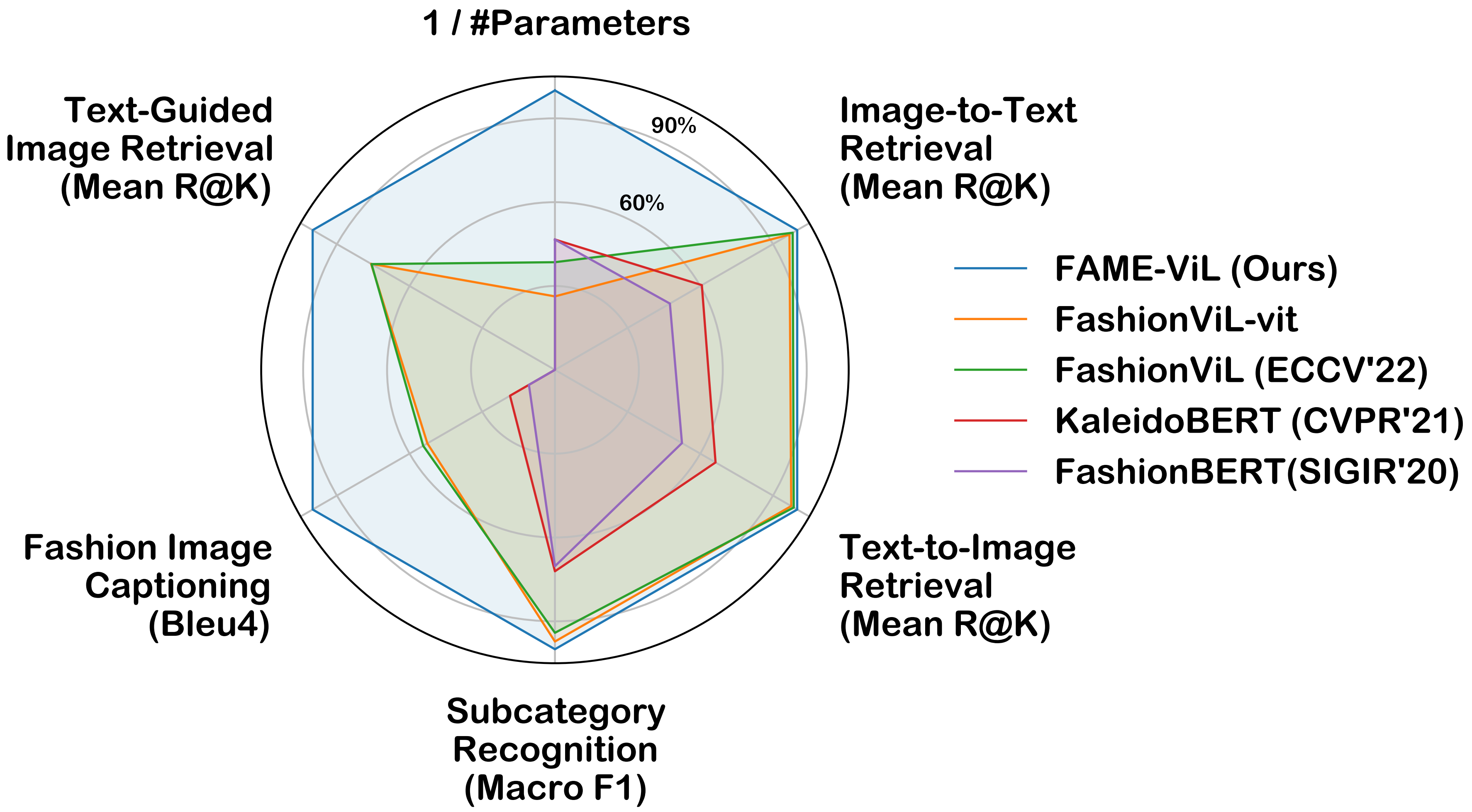
$$\mathcal{L}_{\text{TGIR}} = \mathcal{L}_{\text{InfoNCE}}\left((\mathbf{I}^{r},\mathbf{T}),\mathbf{I}^{t}\right), \mathcal{L}^{\text{D}}_{\text{TGIR}} = \frac{1}{B}\sum_{b}^{B}\text{KL}\left(\mathbf{s}_{(b,b),\cdot} \,\|\, \tilde{\mathbf{s}}_{(b,b),\cdot}\right)$$

$$\mathcal{L}_{\text{FIC}} = -\mathbb{E}_{(I,T)\sim D}\sum_{a=1}^{A}\log P\left(T_{a} \middle| f_{\theta}^{[g]}(I;T_{<a})\right)$$

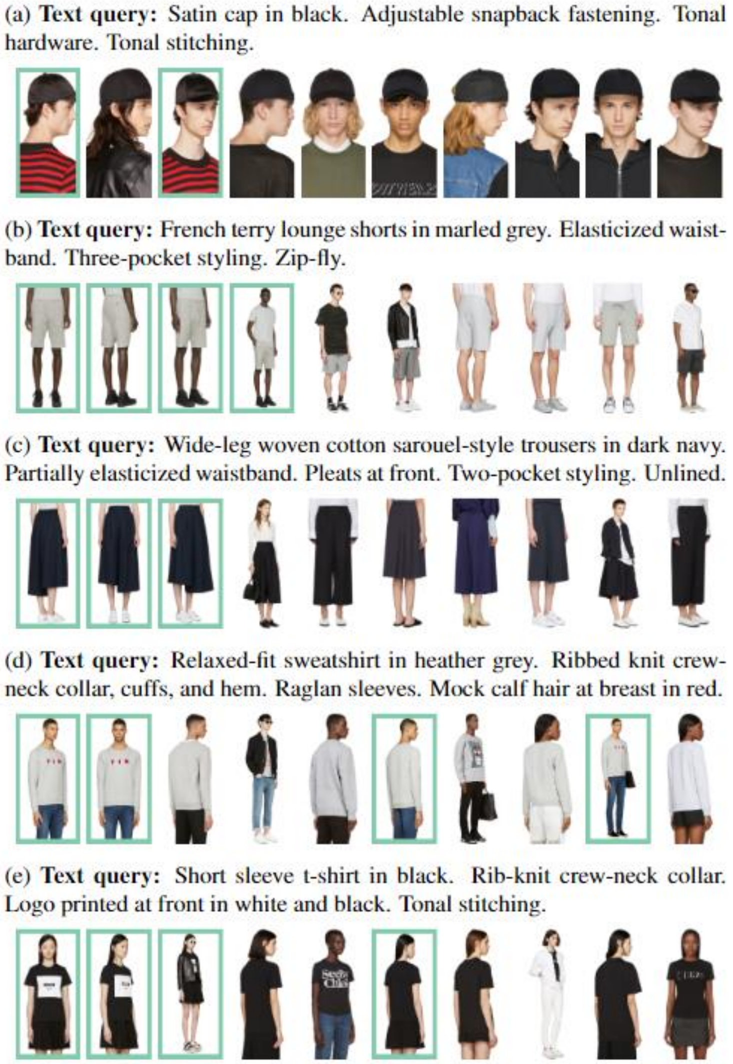
$$\mathcal{L}^{\text{D}}_{\text{FIC}} = \sum_{a=1}^{A}\text{KL}\left(f_{\theta}^{[g]}(I;T_{<a})_{a} \,\|\, g_{\text{fic}}(I;T_{<a})_{a}\right)$$

$$\mathcal{L} = \mathcal{L}_{[\text{task}]} + \mathcal{L}^{\text{D}}_{[\text{task}]}, \quad [\text{task}] \overset{P}{\sim} \{\text{XMR},\text{TGIR},\text{SCR},\text{FIC}\}$$

Quantitative Results



Qualitative Results



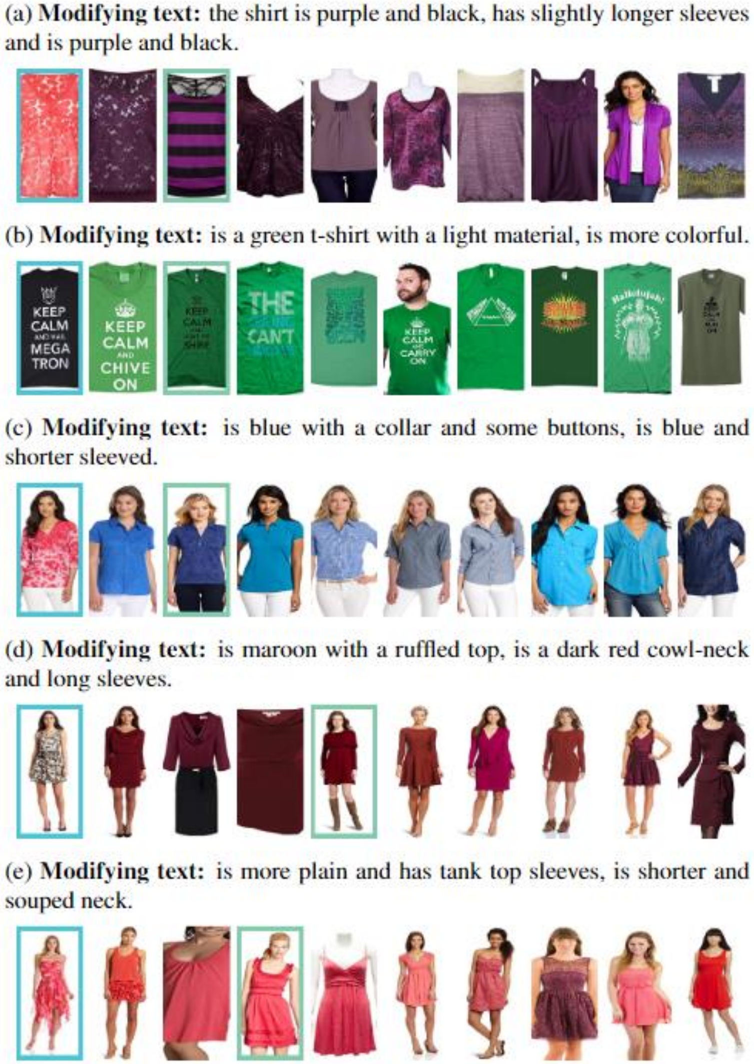
(a) **Text query:** Satin cap in black. Adjustable snapback fastening. Tonal hardware. Tonal stitching.

(b) **Text query:** French terry lounge shorts in marled grey. Elasticized waistband. Three-pocket styling. Zip-fly.

(c) **Text query:** Wide-leg woven cotton sarouel-style trousers in dark navy. Partially elasticized waistband. Pleats at front. Two-pocket styling. Unlined.

(d) **Text query:** Relaxed-fit sweatshirt in heather grey. Ribbed knit crew-neck collar, cuffs, and hem. Raglan sleeves. Mock calf hair at breast in red.

(e) **Text query:** Short sleeve t-shirt in black. Rib-knit crew-neck collar. Logo printed at front in white and black. Tonal stitching.



(a) **Modifying text:** the shirt is purple and black, has slightly longer sleeves and is purple and black.

(b) **Modifying text:** is a green t-shirt with a light material, is more colorful.

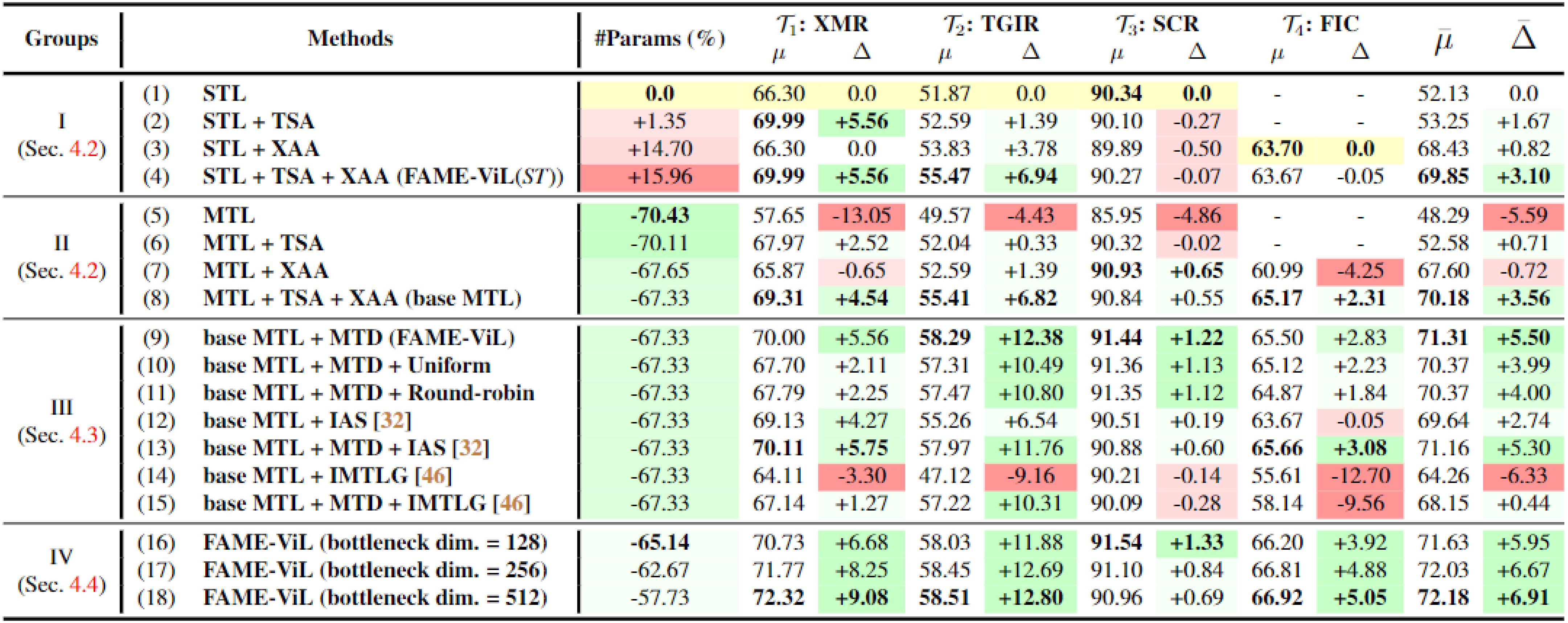
(c) **Modifying text:** is blue with a collar and some buttons, is blue and shorter sleeved.

(d) **Modifying text:** is maroon with a ruffled top, is a dark red cowl-neck and long sleeves.

(e) **Modifying text:** is more plain and has tank top sleeves, is shorter and souped neck.

	Images	Ground Truth Captions	Generated Captions
(a)		White logo tank top. Relaxed-fit tank top in white. Ribbed scoopneck collar and armscyes. Logo print at black. Tonal logo embroidered at back hem. Tonal stitching.	White logo tank top. Racer-back tank top in white. Scoopneck collar. Logo printed at front in black. Curved hem. Tonal stitching.
(b)		Black python print shirt. Short sleeve shirt in tones of grey and black. Detailed python scale print throughout with ombre effect at bottom portions. Spread collar. Button closure at front. Tonal stitching. Single-button barrel cuffs with buttoned sleeve placket.	Black paint splatter shirt. Long sleeve shirt in black. Graphic print throughout in white. Spread collar. Button closure at front. Tonal stitching. Single-button barrel cuffs with buttoned sleeve placket.
(c)		Black jersey leather trim lounge pants. Leather-trimmed stretch jersey lounge pants in black. Partially elasticized waistband with leather drawstring closure. Zip fly. Leather pocket trim. Elasticized grosgrain cuffs.	Black lounge pants. Lounge pants in black. Elasticised waistband with drawstring closure. Four-pocket styling. Elasticised ankle cuffs. Tonal stitching. Zip fly.
(d)		Navy pixel print atari edition polo. Short sleeve oversized polo in navy. Atari pixel print at front. Spread collar with two-button placket. Slits at side seams. Tonal stitching.	Navy embroidered patch polo. Short sleeve cotton piqu & eacute polo in navy. Ribbed spread collar and trim at sleeve opening. Five-button placket at front. Signature tri-color tab at back collar. Tennis tail hem. Tonal stitching.

Further Analysis



Groups		Methods	#Params (%)	$\mathcal{T}_1$: XMR		$\mathcal{T}_2$: TGIR		$\mathcal{T}_3$: SCR		$\mathcal{T}_4$: FIC		$\bar{\mu}$	$\bar{\Delta}$
				μ	Δ	μ	Δ	μ	Δ	μ	Δ		
I (Sec. 4.2)	(1)	STL	**0.0**	66.30	0.0	51.87	0.0	**90.34**	**0.0**	-	-	52.13	0.0
	(2)	STL + TSA	+1.35	**69.99**	**+5.56**	52.59	+1.39	90.10	-0.27	-	-	53.25	+1.67
	(3)	STL + XAA	+14.70	66.30	0.0	53.83	+3.78	89.89	-0.50	**63.70**	**0.0**	68.43	+0.82
	(4)	STL + TSA + XAA (FAME-ViL($\mathcal{ST}$))	+15.96	**69.99**	**+5.56**	**55.47**	**+6.94**	90.27	-0.07	63.67	-0.05	**69.85**	**+3.10**
II (Sec. 4.2)	(5)	MTL	**-70.43**	57.65	-13.05	49.57	-4.43	85.95	-4.86	-	-	48.29	-5.59
	(6)	MTL + TSA	-70.11	67.97	+2.52	52.04	+0.33	90.32	-0.02	-	-	52.58	+0.71
	(7)	MTL + XAA	-67.65	65.87	-0.65	52.59	+1.39	**90.93**	**+0.65**	60.99	-4.25	67.60	-0.72
	(8)	MTL + TSA + XAA (base MTL)	-67.33	**69.31**	**+4.54**	**55.41**	**+6.82**	90.84	+0.55	**65.17**	**+2.31**	**70.18**	**+3.56**
III (Sec. 4.3)	(9)	base MTL + MTD (FAME-ViL)	-67.33	70.00	+5.56	**58.29**	**+12.38**	**91.44**	**+1.22**	65.50	+2.83	**71.31**	**+5.50**
	(10)	base MTL + MTD + Uniform	-67.33	67.70	+2.11	57.31	+10.49	91.36	+1.13	65.12	+2.23	70.37	+3.99
	(11)	base MTL + MTD + Round-robin	-67.33	67.79	+2.25	57.47	+10.80	91.35	+1.12	64.87	+1.84	70.37	+4.00
	(12)	base MTL + IAS [32]	-67.33	69.13	+4.27	55.26	+6.54	90.51	+0.19	63.67	-0.05	69.64	+2.74
	(13)	base MTL + MTD + IAS [32]	-67.33	**70.11**	**+5.75**	57.97	+11.76	90.88	+0.60	**65.66**	**+3.08**	71.16	+5.30
	(14)	base MTL + IMTLG [46]	-67.33	64.11	-3.30	47.12	-9.16	90.21	-0.14	55.61	-12.70	64.26	-6.33
	(15)	base MTL + MTD + IMTLG [46]	-67.33	67.14	+1.27	57.22	+10.31	90.09	-0.28	58.14	-9.56	68.15	+0.44
IV (Sec. 4.4)	(16)	FAME-ViL (bottleneck dim. = 128)	**-65.14**	70.73	+6.68	58.03	+11.88	**91.54**	**+1.33**	66.20	+3.92	71.63	+5.95
	(17)	FAME-ViL (bottleneck dim. = 256)	-62.67	71.77	+8.25	58.45	+12.69	91.10	+0.84	66.81	+4.88	72.03	+6.67
	(18)	FAME-ViL (bottleneck dim. = 512)	-57.73	**72.32**	**+9.08**	**58.51**	**+12.80**	90.96	+0.69	**66.92**	**+5.05**	**72.18**	**+6.91**

Highlight

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Thanks for your attention!

If you have any questions, please feel free to contact **Xiao Han**.

✉ xiao.han@surrey.ac.uk

https://github.com/BrandonHanx/FAME-ViL