FAME-ViL: Multi-Tasking Vision-Language Model for Heterogeneous Fashion Tasks
(Supplementary File)

Xiao Han[1,2] Xiatian Zhu[1,3] Licheng Yu Li Zhang[4] Yi-Zhe Song[1,2] 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 [4] Fudan University
{xiao.han, xiatian.zhu, y.song, t.xiang}@surrey.ac.uk
lichengyu24@gmail.com lizhangfd@fudan.edu.cn

Model architecture	Vision encoder (VE)	CLIP (ViT-B/16) [3]
	Language encoder (LE)	CLIP (ViT-B/16) [3]
	Bottleneck dim.	64
Data augmentation	Resize	(256, 256)
	RandomCrop	(224, 224)
	RandomHorizontalFlip	✓
Training setting	Number of iterations	90k
	Batch size	64
	Initial LR of VE/LE	1e-6
	Initial LR of Adapters	1e-4
	LR schedule	Multi-step
	LR steps	50k and 80k
	LR decrease ratio	0.1
	Warmup iterations	10k
	Warmup factor	0.25
	Optimizer	AdamW (0.9, 0.999)
	Weight decay	1e-5
Hardware	GPU	4 × RTX 3090
	Training duration	31.5h

Table 1. Details for multi-task training FAME-ViL.

XMR* (Image to Text)			XMR* (Text to Image)		
R@1	R@5	R@10	R@1	R@5	R@10
45.99 ± 0.25	73.25 ± 0.15	81.84 ± 0.12	53.12 ± 0.07	77.55 ± 0.35	86.02 ± 0.22
TGIR (Dress)		**TGIR (Shirt)**		**TGIR (Toptee)**	
R@10	**R@50**	**R@10**	**R@50**	**R@10**	**R@50**
42.16 ± 0.32	67.10 ± 0.22	47.19 ± 0.33	67.91 ± 0.66	50.79 ± 0.08	73.48 ± 0.38
SCR (Acc)	**SCR (F1)**	**FIC (B)**	**FIC (M)**	**FIC (R)**	**FIC (C)**
94.62 ± 0.08	87.80 ± 0.25	30.59 ± 0.15	24.22 ± 1.99	55.69 ± 0.13	148.9 ± 1.19

Table 2. Statistical significance quantification of our results. * XMR evaluation is under `full database` protocol

A. Implementation details

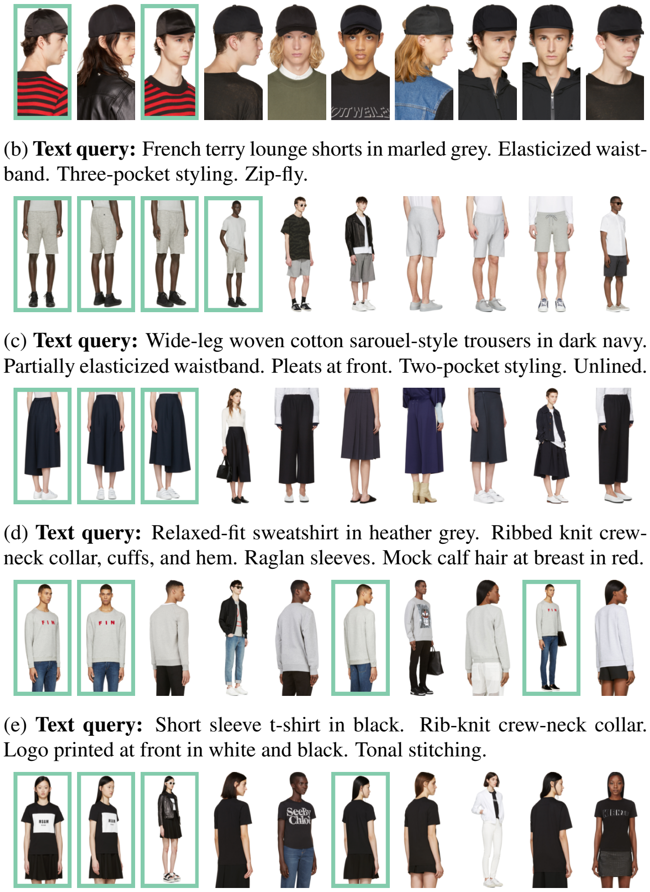
This section describes our implementation and multi-task training details for FAME-ViL.

Architecture details. As mentioned in the main paper, we build our FAME-ViL upon off-the-shelf CLIP model [3]. We utilize the ViT-B/16 version and get the pre-trained weights from HuggingFace Transformers [6]. Specifically, as described in the original paper [3], the language encoder is a 12-layer 512-wide Transformer [5] with 8 attention heads, while the vision encoder is a base-size Vision Transformer (ViT) [1] with patch size as 16. Masked self-attention was used in the language encoder. For computational efficiency, the max text sequence length was capped at 76. The text sequence is bracketed with `[SOS]` and `[EOS]` tokens and the activations of the highest layer of the language encoder at the `[EOS]` token are treated as the text feature representation. Please find more details about CLIP and its pre-training in the original paper [3].

Training details. We list all hyper-parameters used for multi-task training in Tab. 1, including data augmentation methods, optimizer setting, scheduler setting, and *etc*. This results in about 31.5 hours of training time on four RTX 3090 GPUs (24GB memory for each). For single-task training (used by single-task teachers training and ablation study), we adopted the same hyper-parameters except for shorter training iterations (30k for tasks on FashionGen [4], 6k for tasks on FashionIQ [7]).

B. Additional quantitative results

We followed the same protocol used by previous works [2] and used the same random seed for training, to ensure a direct comparison to these main competitors. We also trained our model two more times with different random seeds to measure our method's stability. The statistical performance (w/ mean and std) over three trials in Tab. 2 shows that our model is stable.



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

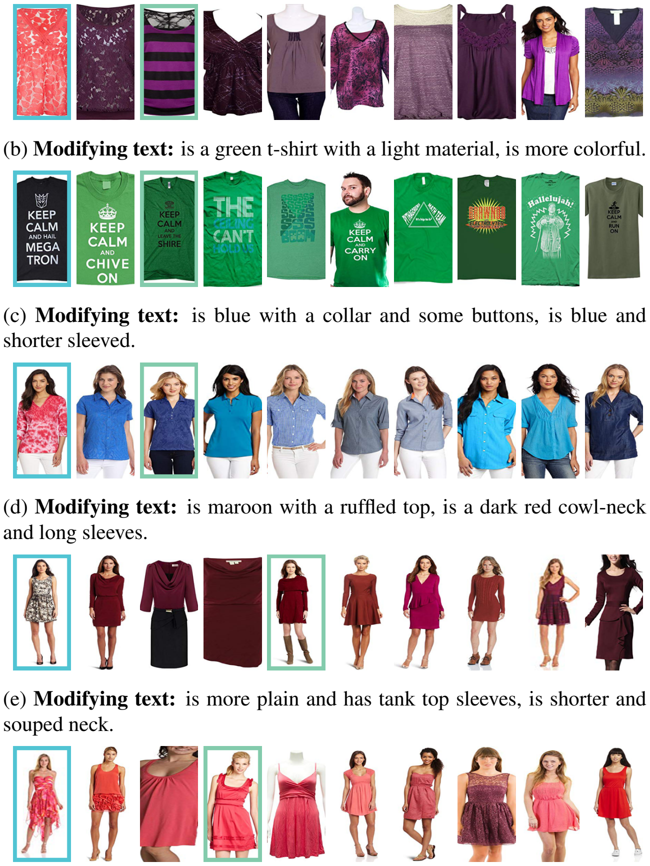
(b) **Text query:** French terry lounge shorts in marled grey. Elasticized waistband. Three-pocket styling. Zip-fly.

(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.

Figure 1. XMR examples. Green box indicates the ground truth.



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

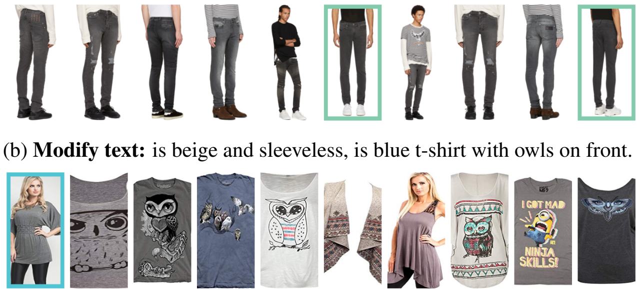
(b) **Modifying text:** is a green t-shirt with a light material, is more colorful.

(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.

Figure 2. TGIR examples. Blue box indicates the reference image. Green box indicates the ground truth.



(a) **Text query:** Skinny-fit stretch denim jeans in 'fade to' grey. Fading and whiskering throughout. Mid-rise. Five-pocket styling.

(b) **Modify text:** is beige and sleeveless, is blue t-shirt with owls on front.

Figure 3. Failure cases. Blue box indicates the reference image. Green box indicates the ground truth.

C. Additional qualitative results

We provide more visualization results in this section to better understand the performance of our FAME-ViL in a qualitative way. Specifically, we show cross-modal retrieval (XMR) results in Fig. 1, text-guided image retrieval (TGIR) results in Fig. 2 and fashion image captioning (FIC) results in Tab. 3. We didn't show subcategory recognition (SCR) results here because of the lack of intuition when visualizing this classification task, but the visualized attention maps are given in Fig. 4.

For retrieval tasks (XMR and TGIR), we observe ambiguities (*i.e.*, the ground truth is not the only one matching the query) in the fashion datasets [4, 7]. Especially in FashionIQ, there are many false negatives that are neglected during the data-annotation stage. Even so, our FAME-ViL can offer us a reliable and human-understandable ranking list, demonstrating its superiority in fine-grained discrimination. Fig. 3 shows example failure cases from (a) XMR and (b) TGIR. In the text query example, we can see that even the human-annotated ground truth (indicated by green boxes) images do not fit the text query perfectly. In both failure cases, the top retrieved results, though wrong according to the "ground truth", are still largely aligned with the

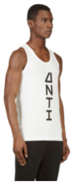
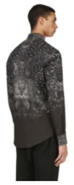
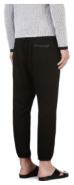
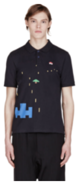
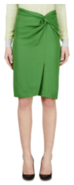
	Images	Ground Truth Captions	Generated Captions
(a)		White logo tank top. Relaxed-fit tank top in white. Ribbed scoopneck collar and armscyes. Logo print at black. Tonal logo embroidered at back hem. Tonal stitching.	White logo tank top. Racer-back tank top in white. Scoopneck collar. Logo printed at 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.
(e)		Green wrap pencil skirt. High-waisted wrap pencil skirt in green. Gathered knot detail at waist. Vent at front hem. Zip closure at back. Tonal stitching.	Green silk draping skirt. Silk skirt in green. Elasticized waistband with drawstring at interior. Vented at back waist. Seam pockets at sides. Tonal stitching.

Table 3. FIC examples. Green text indicates the matched phrases.

query/modify text.

Because of the fine-grained nature of the fashion domain, the ground truth captions in fashion contain much more fine-grained phrases than those in the generic domain [2]. Despite this challenge, our FAME-ViL can produce concrete and accurate phrases in the generated captions. Even if some of the generated phrases do not exist in the ground truth, they still conform to the content of the image and human intuition. This point proves the effectiveness of FAME-ViL in fine-grained generation.

To gain a more intuitive understanding of how attention is learned in our model, we visualize the *text-to-image attention maps* (the average over all heads) in the *last XAA* of the STL baseline (first row) and our MTL model (second row), as shown in Fig. 4. It is observed that the attention maps from our model are more accurate and meaningful.

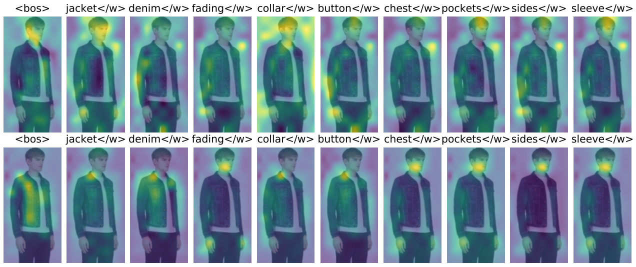


Figure 4. Visualized attention maps of SCR.

References

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020. 1

[2] Xiao Han, Licheng Yu, Xiatian Zhu, Li Zhang, Yi-Zhe Song, and Tao Xiang. Fashionvil: Fashion-focused vision-and-language representation learning. In *ECCV*, 2022. 1, 3

[3] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1

[4] Negar Rostamzadeh, Seyedarian Hosseini, Thomas Boquet, Wojciech Stokowiec, Ying Zhang, Christian Jauvin, and Chris Pal. Fashion-gen: The generative fashion dataset and challenge. *arXiv preprint arXiv:1806.08317*, 2018. 1, 2

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 1

[6] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *EMNLP*, 2020. 1

[7] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. Fashion iq: A new dataset towards retrieving images by natural language feedback. In *CVPR*, 2021. 1, 2