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

Xiao Han\textsuperscript{1,2} Xiatian Zhu\textsuperscript{1,3} Licheng Yu Li Zhang\textsuperscript{4} Yi-Zhe Song\textsuperscript{1,2} Tao Xiang\textsuperscript{1,2}
\textsuperscript{1} CVSSP, University of Surrey \hspace{1cm} \textsuperscript{2} iFlyTek-Surrey Joint Research Centre on Artificial Intelligence
\textsuperscript{3} Surrey Institute for People-Centred Artificial Intelligence \hspace{1cm} \textsuperscript{4} Fudan University

\{xiao.han, xiatian.zhu, y.song, t.xiang\}@surrey.ac.uk
lichengyu24@gmail.com lizhangfd@fudan.edu.cn

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|}
\hline
Model architecture & Vision encoder (VE) & CLIP (ViT-B/16) \cite{CLIP} \\
& Language encoder (LE) & CLIP (ViT-B/16) \cite{CLIP} \\
& Bottleneck dim. & 64 \\
\hline
Data augmentation & Resize & (256, 256) \\
& RandomCrop & (224, 224) \\
& RandomHorizontalFlip & ✓ \\
\hline
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 \\
& Warmup iterations & 10k \\
& Warmup factor & 0.25 \\
& Optimizer & AdamW (0.9, 0.999) \\
& Weight decay & 1e-5 \\
\hline
Hardware & GPU & 4 \times RTX 3090 \\
& Training duration & 31.5h \\
\hline
\end{tabular}
\caption{Details for multi-task training FAME-ViL.}
\end{table}

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 \cite{CLIP}. We utilize the ViT-B/16 version and get the pre-trained weights from HuggingFace Transformers \cite{HuggingFace}. Specifically, as described in the original paper \cite{CLIP}, the language encoder is a 12-layer 512-wide Transformer \cite{Transformer} with 8 attention heads, while the vision encoder is a base-size Vision Transformer (ViT) \cite{ViT} 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 \cite{CLIP}.

Training details. We list all hyper-parameters used for multi-task training in Tab. 1, including data augmentation methods, optimizer setting, scheduler setting, and \textit{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 \cite{FashionGen}, 6k for tasks on FashionIQ \cite{FashionIQ}).

B. Additional quantitative results

We followed the same protocol used by previous works \cite{PreviousWorks} 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.
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
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.

### References


