FashionViL: Fashion-Focused Vision-and-Language Representation Learning *(Supplementary File)*

Xiao Han[1,2], Licheng Yu[3], Xiatian Zhu[1,4], Li Zhang[5]
Yi-Zhe Song[1,2], and Tao Xiang[1,2]

[1] Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey
[2] iFlyTek-Surrey Joint Research Centre on Artificial Intelligence
[3] Meta AI
[4] Surrey Institute for People-Centred Artificial Intelligence, University of Surrey
[5] School of Data Science, Fudan University
{xiao.han,xiatian.zhu,y.song,t.xiang}@surrey.ac.uk
lichengyu@fb.com lizhangfd@fudan.edu.cn

This supplementary material includes three sections. Sec. A describes our implementation details for the pre-training pipeline and each downstream task. Sec. B shows more experiments to demonstrate the effectiveness of FashionViL. Sec. C provides the additional visualization examples.

A Implementation details

A.1 Pre-training

Image tokenizer. As discussed in the main paper, we adopt the Masked Patch Feature Classification (MPFC) as one of our pre-training tasks. An image tokenizer is used to convert the raw pixel values into discrete labels. While previous works like BEiT [1] applied the off-the-shelf image tokenizer pre-trained on the large-scale generic image data [12], we train the image tokenizer by ourselves on the four available fashion datasets [13,19,17,7] as focusing more on the fashion domain.

Specifically, we implement a vector-quantized VAE (VQVAE) [16] with similar Encoder and Decoder architectures as VQGAN [4]. The model details are listed in Table 1. We apply the perceptual loss [8] to learn the codebook, but disregard the adversarial loss which was used in VQGAN [4] as it has been shown to be trivial for the representation learning [3]. We adopt the same training objective as PeCo [3] to learn our VQVAE with the hyper-parameters listed in Table 2. Some reconstruction samples can be found in Fig. 1.

Pre-training. FashionViL is end-to-end pre-trained on 6 tasks as mentioned in the main paper. Previous fashion V+L works, *i.e.* FashionBERT [5] and KaleidoBERT [20], perform all the pre-training tasks in one iteration, which is memory demanding. In this work, we follow UNITER [2] to sample one task per iteration and train it with one objective.

We implement FashionViL pre-training with MMF [15] on 4 RTX 3090 GPUs. All hyper-parameters are listed in Table 3.

Table 1. High-level architecture of the encoder and decoder of our VQVAE

Encoder
$x \in \mathbb{R}^{224\times224\times3}$
Conv2D $\rightarrow \mathbb{R}^{224\times224\times128}$
6× {Res Block, Res Block, Downsample Block} $\rightarrow \mathbb{R}^{7\times7\times512}$
2× {Non-local Block, Res Block} $\rightarrow \mathbb{R}^{7\times7\times512}$
GroupNorm, Swish, Conv2D $\rightarrow \mathbb{R}^{7\times7\times256}$

Decoder
$z_{\mathbf{q}} \in \mathbb{R}^{7\times7\times256}$
Conv2D $\rightarrow \mathbb{R}^{7\times7\times512}$
2× {Res Block, Non-local Block} $\rightarrow \mathbb{R}^{7\times7\times512}$
6× {Res Block, Res Block, Upsample Block} $\rightarrow \mathbb{R}^{7\times7\times128}$
GroupNorm, Swish, Conv2D $\rightarrow \mathbb{R}^{224\times224\times3}$

Table 2. Hyper-parameters for training our VQVAE

Data augmentation	RandomResizedCrop	(224, 224)
Model configuration	Codebook size	1024
	Latent feature dimension	256
	EMA decay	0.99
Training setting	Number of iterations	500,000
	Batch size	32
	Initial LR	1.44e-4
	Optimizer	Adam (0.5, 0.9)
Hardware	GPU	4 x RTX 3090
	Training duration	96h

Table 3. Hyper-parameters for pre-training FashionViL

Image encoder		ResNet50
Text/Fusion encoder		BERT-base-uncased
Text tokenizer	Sequence length	75
	Mask probability	15%
	Whole word mask	✓
Image tokenizer	Min masked patches	4
	Max masked patches	8
	Aspect ratio of mask	(1/3, 3)
Data augmentation	Resize	(256, 256)
	RandomCrop	(224, 224)
	RandomHorizontalFlip	✓
Training setting	Number of iterations	120,000
	Batch size	256
	Initial LR of TE/FE	1e-5
	Initial LR of IE	2e-4
	LR schedule	Multi-step
	LR steps	45,000 and 90,000
	LR decrease ratio	0.1
	Warmup iterations	15,000
	Warmup factor	0.25
	Optimizer	AdamW (0.9, 0.999)
	Weight decay	1e-4
Hardware	GPU	4 × RTX 3090
	Training duration	28.5h

Table 4. Hyper-parameters for fine-tuning FashionViL on cross-modal retrieval

Image encoder		ResNet50
Text/Fusion encoder		BERT-base-uncased
Text tokenizer	Sequence length	75
Data augmentation	Resize	(256, 256)
	RandomCrop	(224, 224)
	RandomHorizontalFlip	✓
Training setting	Number of iterations	75,120
	Batch size	64
	Initial LR of TE	1e-5
	Initial LR of IE	2e-4
	LR schedule	Multi-step
	LR steps	28,170 and 56,340
	LR decrease ratio	0.1
	Warmup iterations	9,390
	Warmup factor	0.25
	Optimizer	AdamW (0.9, 0.999)
	Weight decay	1e-4
Hardware	GPU	1 x RTX 3090
	Training duration	9h

A.2 Fine-tuning

Cross-modal retrieval (ITR & TIR). As ITR and TIR have the same objective as image-text contrastive learning (ITC), we directly fine-tune FashionViL with $\mathcal{L}_{\text{ITC}}$ on the FashionGen dataset [13], where the learnable temperature τ is initialized as 0.625. All hyper-parameters are listed in Table 4.

Text-guided image retrieval (TGIR). Previous works [10,14] found TGIR is a sensitive task (or dataset). Even a small change in the training setting can result in a quite different model performance. For a fair and stable comparison, we keep the same experimental setting for all the experiments in Table 4 in the main paper. Specifically, we removed tricks like ensemble learning and only keep the composition module implementation. For methods with lightweight text encoders (C5, C6, C7), we use CLIP embeddings [11] as the initialization of the word embeddings, which is shown to be effective in [6]. We apply batch-based classification (BBC) loss [18] for TGIR. All experiments are conducted using the hyper-parameters in Table 5.

Category / Subcategory recognition (CR / SCR). For CR and SCR, we directly follow the setting of KaleidoBERT [20] with the cross entropy (CE) as the loss function. All the hyper-parameters are listed in Table 6.

Outfit complementary item retrieval (OCIR). We follow CSA-Net [9] for the task of OCIR. We tried hard but cannot get the proposed data splits and reproduction code in CSA-Net [9]. We thus reorganize Polyvore Outfits [17] and reproduce CSA-Net by ourselves according to the paper. As a result, our results differ from the original paper, but we will release our splits and reproduction code for the convenience of future research. All the experiments implemented by us follow the same hyper-parameters listed in Table 7. Contrastive loss is applied as the training objective.

Table 5. Hyper-parameters for fine-tuning FashionViL on TGIR

Image encoder		ResNet50
Text/Fusion encoder		BERT-base-uncased
Text tokenizer	Sequence length	75
Data augmentation	Resize	(256, 256)
	RandomCrop	(224, 224)
	RandomHorizontalFlip	✓
Training setting	Number of iterations	44,960
	Batch size	32
	Initial LR of FE	1e-5
	Initial LR of IE	2e-4
	LR schedule	Multi-step
	LR steps	16,860 and 28,100
	LR decrease ratio	0.1
	Warmup iterations	2,810
	Warmup factor	0.25
	Optimizer	AdamW (0.9, 0.999)
	Weight decay	1e-4
Hardware	GPU	1 x RTX 3090
	Training duration	5.5h

Table 6. Hyper-parameters for fine-tuning FashionViL on (S)CR

Image encoder		ResNet50
Text/Fusion encoder		BERT-base-uncased
Text tokenizer	Sequence length	75
Data augmentation	Resize	(256, 256)
	RandomCrop	(224, 224)
	RandomHorizontalFlip	✓
Training setting	Number of iterations	37,580
	Batch size	32
	Initial LR of FE	1e-5
	Initial LR of IE	2e-4
	Optimizer	AdamW (0.9, 0.999)
	Weight decay	1e-4
Hardware	GPU	1 x RTX 3090
	Training duration	2.5h

Table 7. Hyper-parameters for fine-tuning FashionViL on OCIR

Image encoder		ResNet50
Data augmentation	Resize	(256, 256)
	RandomCrop	(224, 224)
	RandomHorizontalFlip	✓
Training setting	Number of iterations	8,000
	Batch size	64
	Initial LR of IE	1e-4
	LR schedule	Multi-step
	LR steps	1,500 and 5,000
	LR decrease ratio	0.1
	Warmup iterations	1,000
	Warmup factor	0.25
	Optimizer	AdamW (0.9, 0.999)
	Weight decay	1e-4
Hardware	GPU	1 x RTX 3090
	Training duration	1.5h

Table 8. Results of multi-image subcategory recognition on FashionGen [13]

SCR w/o pt		SCR w/ pt		M-SCR w/o pt		M-SCR w/ pt	
91.45	78.13	92.33	83.02	90.33	72.16	93.39	84.30

B Additional quantitative results

B.1 Performance on multi-image subcategory recognition

Our model can be easily extended to support multi-image input by concatenating all image tokens together. However, there is no existing downstream task taking multiple images for direct comparison with published results, thus such experiments are omitted. We have now simulated a new one – multi-image subcategory recognition (M-SCR), which takes multiple images as input. Table 8 shows that our pre-training (pt) can yield even larger gain (Acc & Macro$\mathcal{F}$). More interestingly, SCR outperforms M-SCR w/o pre-training, but the comparison is reversed after pre-training, indicating (a) the fusion of multiple images and text is not trivial, and (b) our FashionViL is effective in the fusion task.

C Additional qualitative results

We provide more visualization results in this section to better understand the performance of our FashionViL in a qualitative way.

C.1 VQVAE reconstruction

We show some reconstruction results generated by our VQVAE in Figure 1. The overall quality of the reconstructed images is satisfactory with those basic semantic information (*e.g.*, the outline and color of the object) well preserved.

C.2 Additional t-sne visualization

We provide more t-sne visualizations for FashionViL's joint representations on the fine-grained categories in Figure 2. In each column, we visualize all t-sne embeddings belonging to the same category (*e.g.*, `TOPS`) and color them according to their subcategory labels (*e.g.*, `BLOUSES` and `T-SHIRTS`). With the help of our pre-training tasks, the multimodal representations are better clustered in the latent space at both category-level and subcategory-level, which further proved the effectiveness of our pre-training.

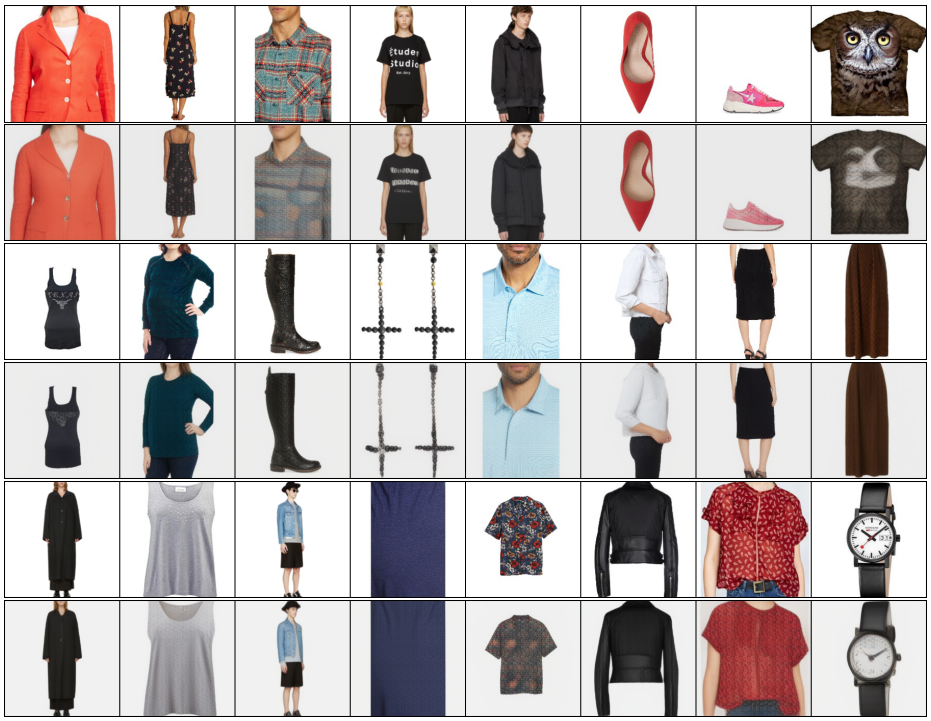

Fig. 1. Some reconstruction results generated by our VQVAE. Odd rows are the original images, and even rows are the reconstructed images from the previous row

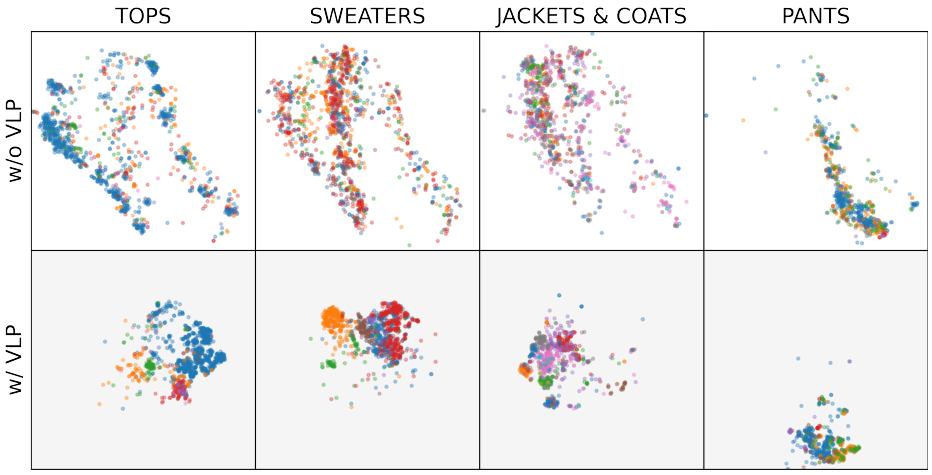

Fig. 2. T-sne of the multimodal representations from not pre-trained and pre-trained FashionViL. Different colors represent subcategories of the categories mentioned in each column header

References

1. Bao, H., Dong, L., Piao, S., Wei, F.: Beit: Bert pre-training of image transformers. In: ICLR (2022) 1
2. Chen, Y.C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: Uniter: Universal image-text representation learning. In: ECCV (2020) 1
3. Dong, X., Bao, J., Zhang, T., Chen, D., Zhang, W., Yuan, L., Chen, D., Wen, F., Yu, N.: Peco: Perceptual codebook for bert pre-training of vision transformers. arXiv preprint arXiv:2111.12710 (2021) 1
4. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: CVPR (2021) 1
5. Gao, D., Jin, L., Chen, B., Qiu, M., Li, P., Wei, Y., Hu, Y., Wang, H.: Fashionbert: Text and image matching with adaptive loss for cross-modal retrieval. In: SIGIR (2020) 1
6. Han, X., He, S., Zhang, L., Xiang, T.: Text-based person search with limited data. In: BMVC (2021) 3
7. Han, X., Wu, Z., Huang, P.X., Zhang, X., Zhu, M., Li, Y., Zhao, Y., Davis, L.S.: Automatic spatially-aware fashion concept discovery. In: ICCV (2017) 1
8. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: ECCV (2016) 1
9. Lin, Y.L., Tran, S., Davis, L.S.: Fashion outfit complementary item retrieval. In: CVPR (2020) 3
10. Liu, Z., Rodriguez-Opazo, C., Teney, D., Gould, S.: Image retrieval on real-life images with pre-trained vision-and-language models. In: ICCV (2021) 3
11. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021) 3
12. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: ICML (2021) 1
13. Rostamzadeh, N., Hosseini, S., Boquet, T., Stokowiec, W., Zhang, Y., Jauvin, C., Pal, C.: Fashion-gen: The generative fashion dataset and challenge. arXiv preprint arXiv:1806.08317 (2018) 1, 3, 5
14. Shin, M., Cho, Y., Ko, B., Gu, G.: Rtic: Residual learning for text and image composition using graph convolutional network. arXiv preprint arXiv:2104.03015 (2021) 3
15. Singh, A., Goswami, V., Natarajan, V., Jiang, Y., Chen, X., Shah, M., Rohrbach, M., Batra, D., Parikh, D.: Mmf: A multimodal framework for vision and language research. https://github.com/facebookresearch/mmf (2020) 1
16. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. In: NeurIPS (2017) 1
17. Vasileva, M.I., Plummer, B.A., Dusad, K., Rajpal, S., Kumar, R., Forsyth, D.: Learning type-aware embeddings for fashion compatibility. In: ECCV (2018) 1, 3
18. Vo, N., Jiang, L., Sun, C., Murphy, K., Li, L.J., Fei-Fei, L., Hays, J.: Composing Text and Image for Image Retrieval - an Empirical Odyssey. In: CVPR (2019) 3
19. Yang, X., Zhang, H., Jin, D., Liu, Y., Wu, C.H., Tan, J., Xie, D., Wang, J., Wang, X.: Fashion captioning: Towards generating accurate descriptions with semantic rewards. In: ECCV (2020) 1
20. Zhuge, M., Gao, D., Fan, D.P., Jin, L., Chen, B., Zhou, H., Qiu, M., Shao, L.: Kaleido-bert: Vision-language pre-training on fashion domain. In: CVPR (2021) 1, 3