FashionFAE: Fine-grained Attributes Enhanced Fashion Vision-Language Pre-training

Jiale Huang[1*], Dehong Gao[3,4*], Jinxia Zhang[1,2†], Zechao Zhan[1], Yang Hu[1], Xin Wang[5]

[1]*Key Laboratory of Measurement and Control of CSE, School of Automation, Southeast University, China*
[2]*Advanced Ocean Institute of Southeast University, Nantong, China* [3]*Northwestern Polytechnical University, China*
[4]*Binjiang Institute of Artificial Intelligence, ZJUT, Hangzhou, China* [5]*Alibaba Group, China*

Abstract—**Large-scale Vision-Language Pre-training (VLP) has demonstrated remarkable success in the general domain. However, in the fashion domain, items are distinguished by fine-grained attributes such as texture and material, which are crucial for tasks such as retrieval. Existing models often fail to take advantage of these fine-grained attributes from both text and image modalities. To address the above issue, we propose a novel approach for the fashion domain, Fine-grained Attributes Enhanced VLP (FashionFAE), which focuses on the detailed characteristics of the fashion data. An attribute-emphasized text prediction task is proposed to predict fine-grained attributes of the items. This forces the model to focus on the salient attributes from the text modality. In addition, a novel attribute-promoted image reconstruction task is proposed, which further enhances the fine-grained ability of the model by leveraging the representative attributes from the image modality. Extensive experiments show that FashionFAE outperforms State-Of-The-Art (SOTA) methods, achieving 2.9% and 5.2% improvements in retrieval on sub-test set and full test set, respectively, and an average improvement of 1.6% in recognition tasks.**

Index Terms—**Vision-Language Pre-training, Fashion Domain, Cross-modal Retrieval**

I. INTRODUCTION

Vision-Language Pre-training (VLP) has experienced remarkable progress [1]–[4] in recent years, becoming a fundamental cornerstone for various multimodal tasks. VLP models learn general knowledge from numerous image-text pairs, enabling them to handle various downstream tasks effectively.

Numerous studies have attempted to apply VLP models to fashion-related tasks [5]–[7]. However, directly using general pre-trained models does not yield optimal results, as they lack specific adaptability to the nuances of fashion data. Compared to the general domain, the fashion domain relies more on **fine-grained information**, such as materials and textures, to distinguish items, rather than simply focusing on coarse-grained information, such as category and location. To enhance the suitability of VLP models for the fashion domain, attention to fine-grained information is essential. Efforts are already
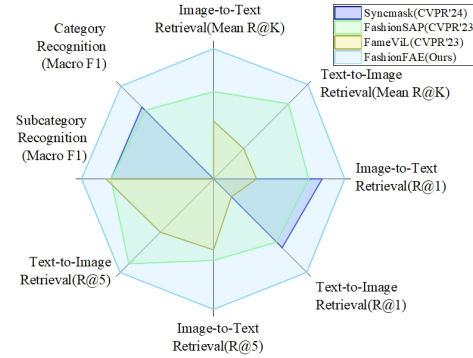
Fig. 1. FashionFAE achieves SOTA performance in various metrics for cross-modal retrieval and (sub)category recognition in the fashion domain.

underway to focus on fine-grained information in the fashion domain [8], [9]. Nevertheless, these models still encounter the following challenges:

1) *In the fashion domain, besides primary descriptive information, there are additional attributes that contain unique features and significant information. Existing research has not explicitly modeled these crucial attributes.*

2)*Current models overlook detailed visual attributes by treating all image parts equally. Focusing only on refining fine-grained text representations is insufficient. A genuine and comprehensive understanding requires the acquisition of fine-grained information from both text and images.*

To solve these problems, we propose a novel approach for the **fashion** domain, called **F**ine-grained **A**ttributes **E**nhanced VLP (FashionFAE), which simultaneously extracts fine-grained information from both the textual and visual sides within the fashion domain. Specifically, our approach introduces fine-grained information into our FashionFAE model through the Attribute-Emphasized Text Prediction (AETP) task on the textual side and the Attribute-Promoted Image Reconstruction (APIR) task on both visual and textual sides. It achieves superior performance on a set of diverse fashion tasks, as in Fig. 1. The overview of the proposed AETP and APIR tasks is illustrated in Fig. 2 (d) and (e).

The AETP task is designed to fully leverage the wealth of additional attribute information available within the fashion dataset. These attributes encompass a variety of elements, including item title, category, subcategory, season, composition,
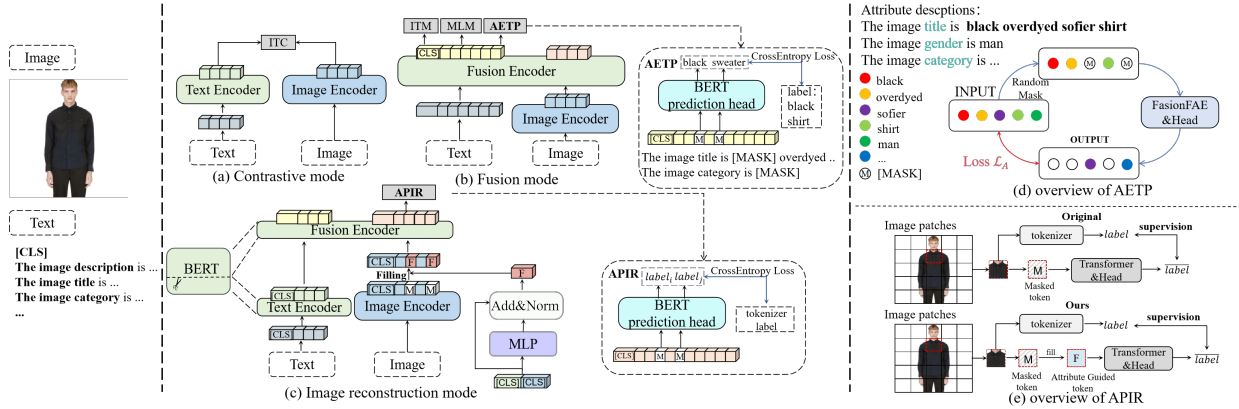
Fig. 2. Overview of our FashionFAE model architecture and proposed AETP and APIR tasks. To accommodate the different 5 pre-training tasks, the model has a total of 3 modes: (a) Contrastive mode; (b) Fusion mode; (c) Image reconstruction mode.

and gender. Firstly, attributes and their respective values are structured into complete attribute description statements and incorporated into the input text. Then, the attribute values are randomly masked and predicted. The AETP task not only maximizes the utilization of various attributes in the fashion dataset but also empowers the model to discern and extract intricate attribute details, contributing to a more comprehensive and insightful analysis of the data.

The APIR task is proposed for both visual and textual modalities. For an input image, it is divided into a grid of patches, which are then discretized into visual labels using latent codes from a Vector Quantized Variational Autoencoder (VQ-VAE). During the APIR task, certain image patches are masked, and their corresponding labels are predicted. Auxiliary information from the text assists in predicting attributes in missing image patches, enhancing its ability to capture fine-grained attribute details.

Our main contributions are summarized below: 1)An effective fashion vision-language pre-training model is proposed, which delves deep into the characteristics of fashion domain data; 2)We propose the AETP task, enhancing the model's ability to capture salient attribute information; 3)We propose the APIR task, enabling the model to more comprehensively learn fine-grained attributes within the images; 4)FashionFAE achieves a new SOTA with a significant improvement in image-to-text retrieval, text-to-image retrieval and (sub)category recognition.

II. METHOD

A. Model Overview

FashionFAE's architecture, shown in Fig. 2, includes an image encoder based on Vision Transformer [10] and a transformer module based on BERT [11].

To achieve a fashion pre-training model with high performance, the architecture comprises three modes that can flexibly handle our proposed pre-training tasks, AETP and APIR, as well as three common pre-training tasks. For different modes, the transformer module based on BERT can function as either a text encoder or a fusion encoder, or it can be divided into two separate components, with one acting as the text encoder and the other as the fusion encoder. Importantly, all these modules share identical parameters.

For the input text, a primary description as well as additional attributes are employed to fully leverage the fine-grained information of the item. From the FashionGen dataset, a total of six crucial attributes are identified: *title*, *category*, *subcategory*, *gender*, *composition*, and *season*. For each attribute and its corresponding value, a comprehensive attribute statement is constructed in the following format: "*The image* [*attribute*] *is* [*value*]." (e.g., "*The image category is shirts.*"). All attribute statements are appended following the initial description, collectively constituting the textual input provided to the model.

B. Pre-training Tasks

1) Attribute-Emphasized Text Prediction (AETP): In attributes like title and category, rich fine-grained information is embedded. These attributes elevate the uniqueness of items, contributing to improved retrieval and classification tasks. The introduction of the AETP task guides the model in concentrating on noteworthy features in the text, thereby augmenting its capacity to comprehend fine-grained details and construct more discriminative representations of items. In this task, an item's title consists of adjectives or nouns describing it (e.g., *title: Black Overdyed Sofier Denim Shirt*), with each word as a subattribute. Other attributes like category and subcategory each align with a single attribute value (e.g., *category: shirts*). We randomly sample N subattributes from the title along with one from the remaining five attributes. The values of the chosen attributes are replaced with a special token [MASK].

The masked text input and the image embedding are fed into the fusion encoder, where the prediction head predicts the attributes that are masked with the remaining text embedding and image embedding. With the image-text pair (v, w) sampled from the dataset D, the AETP task is optimized by minimizing the negative log-likelihood:

$$\mathcal{L}_{AETP} = -\mathbb{E}_{(v,w)\sim D} \log P_\theta(a_m | w_{\backslash m}, v) \quad (1)$$

where P represents the predicted probability, θ represents the learnable parameters, a_m represents masked attributes, $w_{\backslash m}$

represents the embedding without masked attributes m and v represents the image embedding.
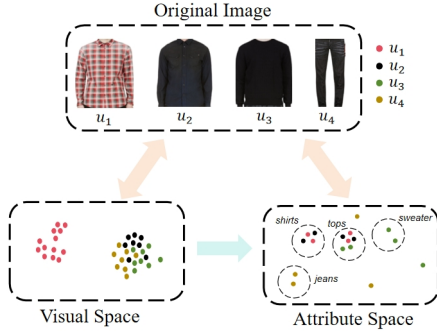

Fig. 3. The visual representations of black shirts, black sweaters, and black pants are indistinguishable by showing similarities in the visual space but can be well differentiated in more advanced attribute space.

2) Attribute-Promoted Image Reconstruction (APIR): In the fashion domain, images often resemble each other in patches due to shared characteristics such as color, texture, and material. As illustrated in Fig. 3, relying solely on pixel-level semantics proves to be insufficient in capturing the entirety of an item's characteristics. Therefore, it is necessary to use a more sophisticated attribute space, particularly at the attribute level, to effectively convey an item's defining attributes. To better abstract embeddings into the semantic space at the attribute level, reconstructing the attributes of these patches is more appropriate. Incorporating characteristics from the textual domain also improves image reconstruction. Thus, the APIR task is proposed to enhance the image modality's capacity to capture detailed information by incorporating prior knowledge from the text modality. The goal of this task is to predict the patch label provided by an offline image tokenizer pre-trained on the fashion dataset.

We randomly replace the 25% of the image patches with the fusion token F, which is generated based on the token w_{cls} from the text encoder and v_{cls} from the image encoder.

$$F = LN(MLP(v_{cls} + w_{cls}) + v_{cls} + w_{cls}) \qquad (2)$$

The model can be trained to predict the tokenized label of each masked patch by minimizing the negative log-likelihood.

$$\mathcal{L}_{APIR} = -\mathbb{E}_{(v,w)\sim D} \log P_\theta(v_m^t|v_{\backslash m}, F, w) \qquad (3)$$

where v_m^t represents the target label for the masked patch, $v_{\backslash m}$ represents the image embedding of unmasked image patches, w represents the text embedding.

3) Image-Text Contrastive Learning (ITC): The ITC task is used to promote close representations of the image and the corresponding text in latent space. To assess the similarity between the given text embedding and the image embedding(v, w), the dot product of their average pooling feature is utilized as a metric. To enhance the similarity, a symmetrical contrastive loss function is employed:

$$\mathcal{L}_{ITC} = \frac{1}{2}[\mathcal{L}_{InfoNCE}(v, w) + \mathcal{L}_{InfoNCE}(w, v)] \qquad (4)$$

$$\mathcal{L}_{InfoNCE} = -\mathbb{E}_{(x,y)\sim B} \log \frac{exp(s(x,y))}{\sum_{\tilde{y}\in B} exp(s(x,\tilde{y}))} \qquad (5)$$

where $s(\cdot, \cdot)$ represents similarity calculation, (x, y) represents the matched pair, $(x, \tilde{y})$ represents the unmatched pair within a batch B.

4) Masked Language Modeling (MLM): Conventional Masked Language Modeling (MLM) is used to model and learn from the description. We use a prediction head with parameters shared with the AETP task. MLM is optimized by minimizing the negative log-likelihood:

$$\mathcal{L}_{MLM} = -\mathbb{E}_{(v,w)\sim D} \log P_\theta(w_m^t|w_{\backslash m}, v) \qquad (6)$$

where w_m^t represents the target label for masked subwords, $w_{\backslash m}$ represents the embedding of unmasked subwords.

5) Image-Text Matching (ITM): In the ITM task, a binary label L is predicted by employing the [CLS] token outputted from the fusion mode, which indicates if each input image-text pair is a match. We apply the cross-entropy loss for the ITM task:

$$\mathcal{L}_{ITM} = -\mathbb{E}_{(v,w)\sim T} \log P_\theta(L|(v,w)) \qquad (7)$$

The complete pre-training objective is the combination of the mentioned tasks above, we randomly sample one task per iteration with the sampling probability P_{task}.

$$\mathcal{L} = \mathcal{L}_{task}$$

$$task \in \{AETP, APIR, ITC, MLM, ITM\}_{P_{task}} \qquad (8)$$

III. EXPERIMENTS

A. Dataset

The FashionGen dataset [12] is used for pre-training and downstream tasks. FashionGen includes 292k text-image pairs and 60k unique fashion items (260k for training and 32k for evaluation). A detailed description of the item is provided in the text annotation as well as a variety of attributes. Downstream tasks, i.e., text-to-image retrieval, image-to-text retrieval, category recognition, and subcategory recognition, are evaluated based on the test set of the FashionGen dataset.

B. Implementation Details

The proposed FashionFAE is implemented using Multi-Modal Framework (MMF) [13] and PyTorch [14]. The hyper-parameter N (the number of masked subattributes in the title) of the proposed AETP task is set to 2. An AdamW optimizer with a learning rate $1e$-5 is adopted. The batch size is 128. During the pre-training stage, the MLM task and the AETP task are optimized simultaneously.

C. Downstream Tasks and Results

1) Cross-modal Retrieval: We address both image-to-text retrieval and text-to-image retrieval on FashionGen dataset using two protocols employed by previous methods: (1) Random 100 Protocol (sub-test set): For each query, 100 candidates from the same category are randomly sampled to form a retrieval database [15], [16]. (2) Full Database Protocol (full

test set): We also embrace a more challenging and practical protocol that involves retrieval across the entire product set. Evaluation metrics include $Rank@K, \{K = 1, 5, 10\}$ and their mean value.

Methods	Pub./Year	I2T				T2I				Mean
		R@1	R@5	R@10	Mean	R@1	R@5	R@10	Mean	
FashionBERT [15]	SIGIR_20	23.96	46.31	52.12	40.80	26.75	46.48	55.74	42.99	41.89
KaleidoBERT [16]	CVPR_22	27.99	60.09	68.37	52.15	33.88	60.60	68.59	54.36	53.25
CommerceMM [17]	SIGKDD_22	41.60	64.00	72.80	59.47	39.60	61.50	72.70	57.93	58.70
EI-CLIP [18]	CVPR_22	38.70	72.20	84.25	65.05	40.06	71.99	82.90	64.98	65.02
MLVT [19]	MIR_23	33.10	77.20	91.10	67.13	34.60	78.00	89.50	67.37	67.25
FashionKLIP [20]	ACL_23	60.79	85.67	91.95	79.47	54.00	78.49	86.28	72.92	76.20
FaD-VLP [21]	EMNLP_22	64.30	86.75	93.48	81.51	58.66	84.92	91.58	78.39	79.95
FashionViL [8]	ECCV_22	65.54	91.34	96.30	84.39	61.88	87.32	93.22	80.81	82.60
Fame-ViL [22]	CVPR_23	65.94	91.92	97.22	85.03	62.86	87.38	93.52	81.25	83.14
FashionSAP [9]	CVPR_23	73.14	92.80	96.87	87.60	70.12	91.76	96.39	86.09	86.85
SyncMask [23]	CVPR_24	75.00	-	-	-	71.00	-	-	-	-
Ours	-	**78.12**	**96.88**	**98.98**	**91.33**	**74.95**	**92.96**	**96.60**	**88.17**	**89.75**

Methods	Pub./Year	I2T				T2I				Mean
		R@1	R@5	R@10	Mean	R@1	R@5	R@10	Mean	
EI-CLIP [18]	CVPR_22	25.70	54.50	66.80	49.00	28.41	57.10	69.40	51.64	50.32
FashionKLIP [20]	ACL_23	37.01	59.78	67.39	54.73	43.70	63.74	72.67	60.04	57.38
FashionViL [8]	ECCV_22	42.88	71.57	80.55	65.00	51.34	75.42	84.57	70.44	67.72
FashionSAP [9]	CVPR_23	54.43	77.30	83.15	71.63	62.82	83.96	90.16	78.98	75.31
SyncMask [23]	CVPR_24	55.39	-	-	-	64.06	-	-	-	-
Ours	-	**57.51**	**83.81**	**89.75**	**77.02**	**68.34**	**89.10**	**94.51**	**83.98**	**80.50**

The results in Tab. I show that our method significantly outperforms the latest SOTA approach in all metrics for I2T and T2I tasks on the sub-test set, with an average improvement of 3.5% in the R@1 metric and an overall performance gain of 2.9%. On the more challenging full test set, as shown in Tab. II, our model also excels, achieving a 5.2% improvement in mean retrieval performance and a 3.2% increase in R@1 accuracy compared to previous SOTA models.

2) Category/Subcategory Recognition (CR&SCR): Following the previous works [8], [16], the CR&SCR tasks are evaluated on the FashionGen dataset. Evaluation metrics include Accuracy (Acc) and Macro-Averaged F1 Score (Macro-F1). The results are shown in Tab. III. FashionFAE achieves the highest Acc and also significantly improves the Macro-F1 score (2.6% on average) in both CR&SCR tasks.

Methods	Pub./Year	CR		SCR		Mean
		Acc	Macro-F1	Acc	Macro-F1	
FashionViL [8]	ECCV_22	97.48	88.60	89.23	83.02	89.58
Fame-ViL [22]	CVPR_23	-	88.21	94.67	88.21	91.44
FashionSAP [9]	CVPR_23	98.34	89.84	94.33	87.67	92.55
SyncMask [23]	CVPR_24	98.41	90.31	94.21	87.83	92.69
Ours	-	**98.79**	**93.32**	**95.25**	**90.07**	**94.36**

D. Ablation Study

The effectiveness of the pre-training tasks adopted is analyzed in this section. The average metrics on the sub-test set and full test set for both I2T and T2I tasks, as well as the average value of Acc and Macro-F1 for CR & SCR, are listed. The results of the ablation study are shown in Tab. IV.

The results of ablation experiments for ITC, ITM, and MLM, widely used in various VLP models, are depicted in the first three rows.

ITC	ITM	MLM	AETP	APIR	sub-test set			full test set			CR	SCR
					I2T	T2I	Mean	I2T	T2I	Mean	Mean	Mean
✓					82.82	78.04	80.43	63.18	69.68	66.43	91.17	84.78
✓	✓				82.71	78.52	80.61	63.48	69.80	66.64	92.46	86.04
✓	✓	✓			83.23	79.10	81.17	63.70	70.45	67.07	92.21	85.65
✓	✓	✓	✓		90.84	87.45	89.14	76.03	83.44	79.74	95.73	91.68
✓	✓	✓	✓	✓	**91.33**	**88.17**	**89.75**	**77.02**	**83.98**	**80.50**	**96.05**	**94.69**

The last two rows of the table present ablation experiments for AETP and APIR. Compared to more widely adopted methods, AETP and APIR are more focused on exploring the features of fashion domain data, enabling the model to pay a thorough attention to the representative characteristics of the data. The ablation experiments demonstrate that the AETP task has highlighted the crucial attributes inherent to items, substantially enhancing their discriminative qualities. This enhancement has led to significant performance improvements across various downstream tasks. The APIR task further exploits the attribute information present in the image domain and synergizes with the textual information, resulting in additional enhancements beyond the already commendable performance achieved.

To validate the effectiveness of individual attributes (*title*(TI), *category* (CA), *subcategory* (S-CA), *gender* (GEN), *composition* (COM), *season* (SEA)) in the AETP task, we conducted ablation experiments. As shown in Tab. V, the first row represents the scenario without additional attributes. The 2nd to 7th rows progressively add one attribute at a time, showing the effectiveness of each attribute.

TI	CA	S-CA	GEN	SEA	COM	sub-test set		full test set	
						I2T	T2I	I2T	T2I
						83.23	79.11	63.70	70.45
✓						83.75	79.67	64.64	70.73
✓	✓					83.73	80.10	65.20	71.38
✓	✓	✓				83.59	79.39	65.44	71.30
✓	✓	✓	✓			85.68	81.53	67.84	74.02
✓	✓	✓	✓	✓		91.25	87.69	76.75	83.47
✓	✓	✓	✓	✓	✓	**91.33**	**88.17**	**77.02**	**83.98**

IV. CONCLUSION

In this paper, we propose FashionFAE, a fine-grained attribute enhanced vision-language pre-training model for the fashion domain. The model includes attribute-emphasized text prediction and attribute-promoted image reconstruction tasks to extract fine-grained information from both text and images. Comparisons and ablation studies show that FashionFAE significantly outperforms SOTA models, with a 2.9% improvement in average retrieval on the sub-test set, a 5.2% improvement on the full test set, and a 1.8% improvement in category and subcategory recognition.

REFERENCES

[1] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, "Visualbert: A simple and performant baseline for vision and language," *arXiv preprint arXiv:1908.03557*, 2019.

[2] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, and J. Dai, "Vl-bert: Pre-training of generic visual-linguistic representations," *International Conference on Learning Representations*, 2020.

[3] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei *et al.*, "Oscar: Object-semantics aligned pre-training for vision-language tasks," in *European Conference on Computer Vision*. Springer, 2020, pp. 121–137.

[4] W. Kim, B. Son, and I. Kim, "Vilt: Vision-and-language transformer without convolution or region supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 5583–5594.

[5] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[6] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi, "Align before fuse: Vision and language representation learning with momentum distillation," *Advances in neural Information Processing Systems*, vol. 34, pp. 9694–9705, 2021.

[7] Y. Wu, H. Li, F. Wang, Y. Zhang, and R. Liang, "Self-distilled dynamic fusion network for language-based fashion retrieval," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 3260–3264.

[8] X. Han, L. Yu, X. Zhu, L. Zhang, Y.-Z. Song, and T. Xiang, "Fashionvil: Fashion-focused vision-and-language representation learning," in *European Conference on Computer Vision*. Springer, 2022, pp. 634–651.

[9] Y. Han, L. Zhang, Q. Chen, Z. Chen, Z. Li, J. Yang, and Z. Cao, "Fashionsap: Symbols and attributes prompt for fine-grained fashion vision-language pre-training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 15 028–15 038.

[10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *International Conference on Learning Representations*, 2021.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *Proceedings of the Conference of North American Chapter of Association for Computational Linguistics*, 2019.

[12] N. Rostamzadeh, S. Hosseini, T. Boquet, W. Stokowiec, Y. Zhang, C. Jauvin, and C. Pal, "Fashion-gen: The generative fashion dataset and challenge," *arXiv preprint arXiv:1806.08317*, 2018.

[13] A. Singh, V. Goswami, V. Natarajan, Y. Jiang, X. Chen, M. Shah, M. Rohrbach, D. Batra, and D. Parikh, "Mmf: A multimodal framework for vision and language research," *MMF: A multimodal framework for vision and language research*, 2020.

[14] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[15] D. Gao, L. Jin, B. Chen, M. Qiu, P. Li, Y. Wei, Y. Hu, and H. Wang, "Fashionbert: Text and image matching with adaptive loss for cross-modal retrieval," in *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020, pp. 2251–2260.

[16] M. Zhuge, D. Gao, D.-P. Fan, L. Jin, B. Chen, H. Zhou, M. Qiu, and L. Shao, "Kaleido-bert: Vision-language pre-training on fashion domain," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 647–12 657.

[17] L. Yu, J. Chen, A. Sinha, M. Wang, Y. Chen, T. L. Berg, and N. Zhang, "Commercemm: Large-scale commerce multimodal representation learning with omni retrieval," in *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022, pp. 4433–4442.

[18] H. Ma, H. Zhao, Z. Lin, A. Kale, Z. Wang, T. Yu, J. Gu, S. Choudhary, and X. Xie, "Ei-clip: Entity-aware interventional contrastive learning for e-commerce cross-modal retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 051–18 061.

[19] G.-P. Ji, M. Zhuge, D. Gao, D.-P. Fan, C. Sakaridis, and L. V. Gool, "Masked vision-language transformer in fashion," *Machine Intelligence Research*, vol. 20, no. 3, pp. 421–434, 2023.

[20] X. Wang, C. Wang, L. Li, Z. Li, B. Chen, L. Jin, J. Huang, Y. Xiao, and M. Gao, "Fashionklip: Enhancing e-commerce image-text retrieval with fashion multi-modal conceptual knowledge graph," in *Proceedings of the Conference on Association for Computational Linguistics*, 2023, pp. 149–158.

[21] S. Mirchandani, L. Yu, M. Wang, A. Sinha, W. Jiang, T. Xiang, and N. Zhang, "Fad-vlp: Fashion vision-and-language pre-training towards unified retrieval and captioning," *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2022.

[22] X. Han, X. Zhu, L. Yu, L. Zhang, Y.-Z. Song, and T. Xiang, "Fame-vil: Multi-tasking vision-language model for heterogeneous fashion tasks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2669–2680.

[23] C. H. Song, T. Hwang, J. Yoon, S. Choi, and Y. H. Gu, "Syncmask: Synchronized attentional masking for fashion-centric vision-language pretraining," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.