SUPPLEMENTARY MATERIAL FOR ”BIDIRECTIONAL SEMANTIC CONSTRUCTION NETWORK FOR COMPOSED QUERY-BASED IMAGE RETRIEVAL”

In this supplementary material, we provide more analyses of our proposed BSCN method, which are difficult to describe in the main paper due to the space limitation. The content of the additional material is shown below:

- We first show the effect of each module on the Shoes and Fashion200k datasets. Then We show a great amount of experimental analyses that complement the main paper, including 1) additional analysis on the learned semantics, 2) additional evaluation metrics.
- We perform more analyses of the comparison with the recent state-of-the-art methods and show the overall training procedure of BSCN.
- We conduct more qualitative analyses, including additional retrieved results and comparison with ARTEMIS on the learned semantics.

1. THE EFFECT OF EACH MODULE AND FURTHER ANALYSIS

1.1. Effect of Each Module on Shoes and Fashion200K

Table 1: Ablation results for each module on the Shoes dataset.

MFE	AM	VSCM	TSCM	R@10	R@50	mR
				52.64	76.78	64.71
✔				53.19	78.21	65.71
✔	✔			53.57	78.74	66.16
✔	✔	✔		54.38	79.46	66.92
✔	✔		✔	54.62	79.28	66.95
✔	✔	✔	✔	55.84	80.55	68.20

Table 2: Ablation results for each module on the Fashion200K dataset.

MFE	AM	VSCM	TSCM	R@10	R@50	mR
				43.8	67.3	55.6
✔				53.6	71.5	62.6
✔	✔			53.8	71.8	62.8
✔	✔	✔		54.8	73.9	64.4
✔	✔		✔	55.1	74.1	64.6
✔	✔	✔	✔	56.2	74.9	65.6

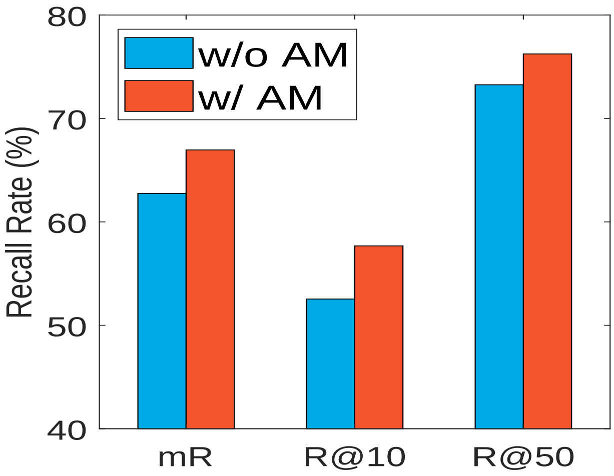

Fig. 1: Effect of aggregating gate on Fashion200k.

To furtherly investigate the importance of each module in our model, we complement ablation studies on the Shoes and Fashion200k datasets. The experimental results of ablation studies are shown in Table 1 and 2. From the results, we obtained the following conclusions: 1) Only using the MFE performs better than the baseline since the baseline fails to adequately leverage the fine-grained and coarse-grained features, which makes the model overlook the relation between the semantic information of different layers. And as shown in Figure 1, the aggregating gate enhances retrieval accuracy by obtaining comprehensive semantic information. However, it is worse than the complete model, demonstrating that using the VSCM and TSCM can obtain more precise semantic information. 2) The performance of Full BSCN shows that guiding the model focus on the correct image parts through performing bidirectional semantic construction, is necessary for the model to achieve proper semantic information. 3) In terms of small-scale and large-scale datasets, our proposed BSCN model always shows great performance. Such results demonstrate that our model can effectively obtain refined semantic information, which is crucial to enhance retrieval accuracy.

1.2. The Further Qualitative Comparison

1.2.1. Additional Analysis on the Learned Semantics Quality.

To further demonstrate the effectiveness of our BSCN method, we further study the learned semantic quality of our BSCN model on the Shoes dataset. According to Figure 2, we can observe that: 1) In terms of the small-scale dataset, compared to the ARTEMIS, the cosine similarity between the textual feature and text-guided modification is higher, while our method achieves more accurate retrieval results. 2) Moreover, the two trends demonstrate our motivation that the learned semantics quality is important to focus on proper parts is reasonable, for both small and large datasets.

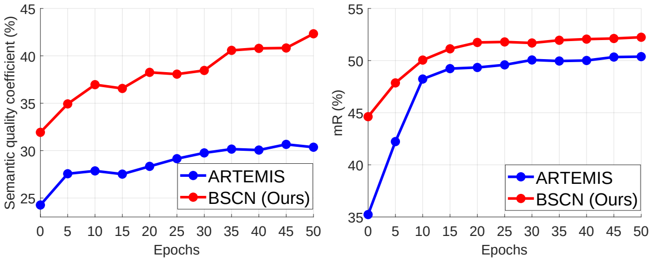

Fig. 2: Effect of the learned semantics quality on retrieval accuracy of Shoes.

1.2.2. Additional Evaluation Metrics

We complement the additional experimental results shown in Figure 3, Figure 4 and Figure 5, where we use the extra metrics including median rank and mean rank (the rank of the ground truth), to reflect the superiority of our BSCN method. From the results of three benchmark datasets, our proposed BSCN method shows more precise retrieval results over the current state-of-the-art method ARTEMIS on all datasets, which suggests that our model has a consistent improvement over different metrics. Specifically, our BSCN method has a better performance on the FashionIQ dataset, we speculate the reason is that images and captions of FashionIQ are more complete and contain more complex semantics, which suggests that our model can properly tackle complex semantic information. Such results verify that our proposed VSCM and TSCM have more advantages in utilizing fine-grained and coarse-grained features and learning proper semantic information.

2. MORE ANALYSES OF COMPARISON WITH THE STATE-OF-THE-ART METHODS

In the main paper, we show the significant improvement of our BSCN method compared to the other approaches. Note that following [1], we use the official validation set of FashionIQ for the fair comparison, due to the validation set proposed by [2] contains only the reference and target images of the different evaluated queries. Here, we perform detailed comparisons with recent state-of-the-art methods. In terms of the methods using the different backbones, for example, we still have a huge improvement over DCNet [3], which encodes both images and text as the composition of multiple embeddings represented by particular experts that are specialized for different spatial locations. Furthermore, as for the methods that learned the unified image-text representation via feature fusion, our BSCN method also has better performance on the Shoes and Fashion200k datasets compared to the current state-of-the-art method CLVC-Net [4].

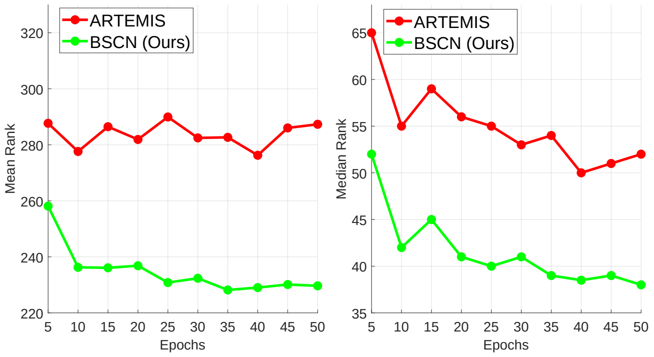

Fig. 3: The comparison of our method (BSCN) with ARTEMIS on Median Rank and Mean Rank of FashionIQ.

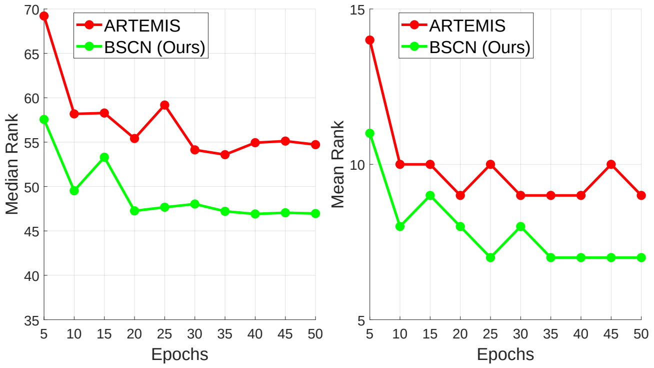

Fig. 4: The comparison of our method (BSCN) with ARTEMIS on Median Rank and Mean Rank of Shoes.

3. QUALITATIVE ANALYSES

3.1. Additional Retrieved Results

We provide more qualitative results of our BSCN method on the three benchmarks datasets (FashionIQ, Shoes, and Fashion200k), which are shown in Figure 6, Figure 7 and Figure 8, where correct targets are tagged with green boxes. The composed queries are shown on the left of the figures, and the top-6 retrieved results from different methods reflect that our model is able to retrieve the correct target images. We can

Algorithm 1 Overall training procedure of BSCN.

```
Phase 1: Freeze the base encoder.
Input: $\mathcal{X} = \{(\mathbf{V}_r^i, \mathbf{M}^i, \mathbf{V}_t^i)\}_{i=1}^B$, batch size $B$, learnable temperature factor $\mu$, hyper-parameters $\lambda$, $\gamma_1^i$ and $\gamma_2^i$, learning rate $\zeta$, $\alpha_1$, $\alpha_2$, $\alpha_3$.
Output: Model Parameters $\theta_t$
Build the model architecture (initialization parameters $\theta_t$).
Freeze the parameters of ResNet50 and GloVe.
repeat
    Sample a batch of composed queries and targets.
    Compute the objective $\mathcal{L}$.
    Update $\theta_t$ using AdamW optimizer.
    $\theta_t \leftarrow \theta_t - \zeta \nabla_{\theta_t} \mathcal{L}$.
until Reach maximum iterations.
Take $\theta_t$ as the input of phase 2.
Phase 2: Training end-to-end.
Input: $\mathcal{X} = \{(\mathbf{V}_r^i, \mathbf{M}^i, \mathbf{V}_t^i)\}_{i=1}^B$, batch size $B$, learnable temperature factor $\mu$, hyper-parameters $\lambda$, $\gamma_1^i$ and $\gamma_2^i$, learning rate $\zeta$, $\alpha_1$, $\alpha_2$, $\alpha_3$, the saved model parameters $\theta_t$.
Output: Model parameters $\theta_t$.
Load the saved model parameters $\theta_t$.
repeat
    Sample a batch of composed queries and targets.
    Compute the objective $\mathcal{L}$.
    Update $\theta_t$ using AdamW optimizer.
    $\theta_t \leftarrow \theta_t - \zeta \nabla_{\theta_t} \mathcal{L}$.
until Reach maximum iterations.
Take trained BSCN to conduct CQBIR.
```

observe that: 1) With the proposed MFE, our BSCN method can satisfy the different levels of modification requirements, *i.e.*, the fine-grained modification *"solid color"* or the coarse-grained modification *"long sleeves"*. 2) Our proposed BSCN method can utilize the unmentioned concepts to correctly infer the resemblance with reference images. Moreover, our model refines the learned semantic information to force the model to focus on the improper image parts, which furtherly improves the performance of retrieving the target image. 3) We notice that the model is also able to retrieve the correct target images from the galleries of different sizes, which demonstrates the sufficient generalization ability of our model.

3.2. Additional Comparison with ARTEMIS on Learned Semantics

As shown in Figure 9, Figure 11, and Figure 10, we perform more comparisons on the learned semantic information between our method BSCN and ARTEMIS [1], to better verify the superiority of our model. From the attention heatmap of the FashionIQ and Shoes, we can observe that: 1) As for different kinds of reference images (dress, shirt and toptee), our proposed BSCN method still more precisely learns the semantic information compared to the ARTEMIS. In detail, given the modification texts and reference images, the ARTEMIS learns improper semantics, which is not relevant to properties mentioned in the text. However, our proposed BSCN model can focus on appropriate target image parts by effectively learning the correct semantics, which verifies that the proposed VSCM and TSCM can obviously enhance semantic quality. 2) Moreover, our model not only has a better performance on the resemblance with reference but also learned the precise text-guided modification.

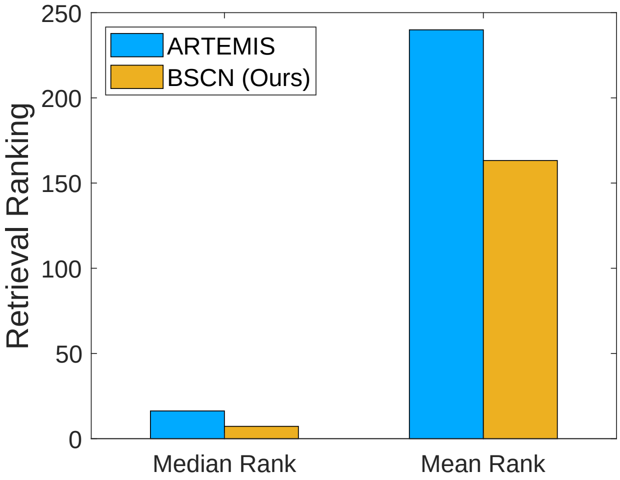

Fig. 5: The comparison of our method (BSCN) with ARTEMIS on Median Rank and Mean Rank of Fashion200k.

4. REFERENCES

[1] Ginger Delmas, Rafael Sampaio de Rezende, Gabriela Csurka, and Diane Larlus, "Artemis: Attention-based retrieval with text-explicit matching and implicit similarity," *CoRR*, vol. abs/2203.08101, 2022.

[2] Yanbei Chen, Shaogang Gong, and Loris Bazzani, "Image search with text feedback by visiolinguistic attention learning," in *CVPR*, 2020.

[3] Jongseok Kim, Youngjae Yu, Hoeseong Kim, and Gunhee Kim, "Dual compositional learning in interactive image retrieval," in *AAAI*, 2021, vol. 35.

[4] Haokun Wen, Xuemeng Song, Xin Yang, Yibing Zhan, and Liqiang Nie, "Comprehensive linguistic-visual composition network for image retrieval," in *SIGIR*, 2021.

[5] 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.

[6] Tamara L Berg, Alexander C Berg, and Jonathan Shih, "Automatic attribute discovery and characterization from noisy web data," in *ECCV*, 2010.

[7] Xiaoxiao Guo, Hui Wu, Yu Cheng, Steven Rennie, Gerald Tesauro, and Rogerio Feris, "Dialog-based interactive image retrieval," in *NIPS*, 2018.

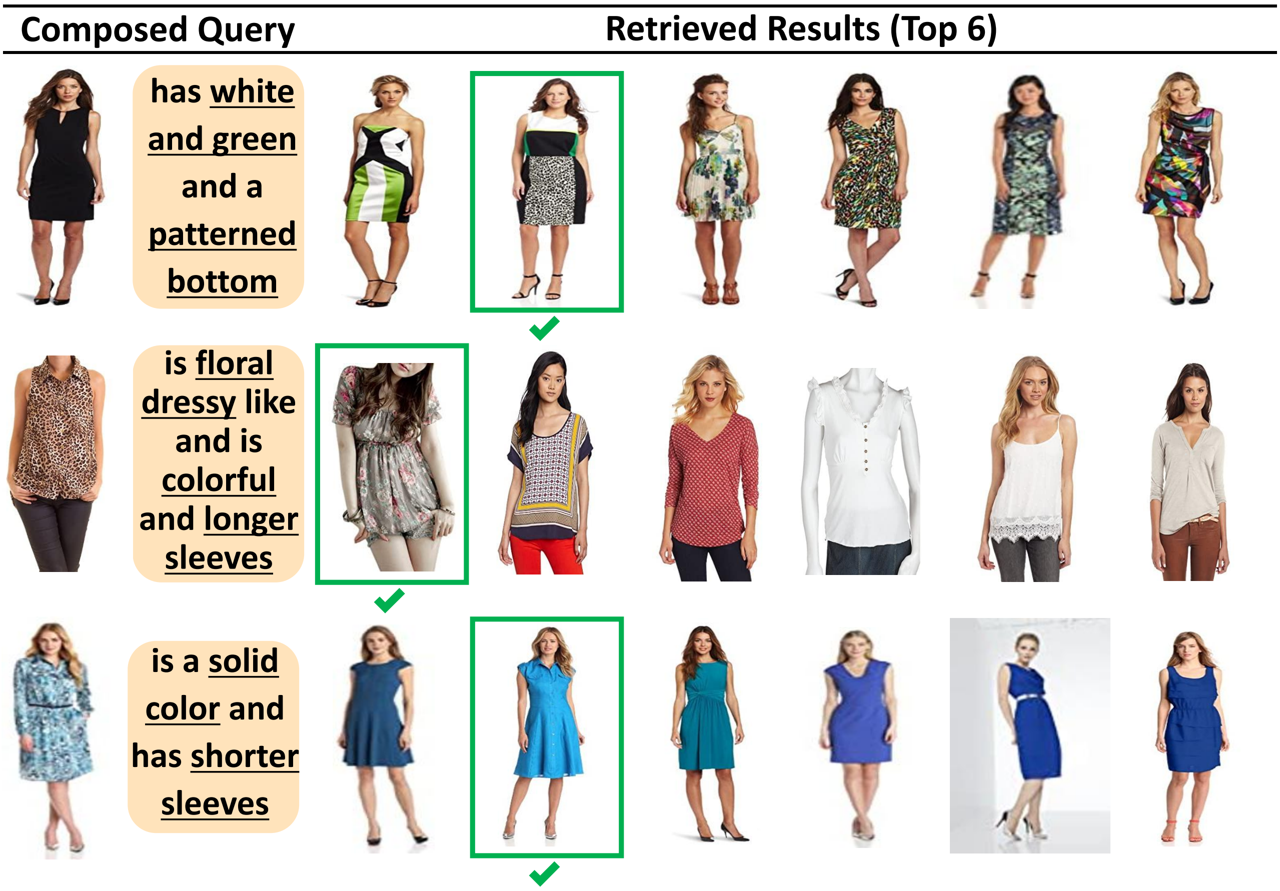
Composed Query
Retrieved Results (Top 6)
has white and green and a patterned bottom
is floral dressy like and is colorful and longer sleeves
is a solid color and has shorter sleeves

Fig. 6: Qualitative results for composed queries in FashionIQ dataset [5]. We show the top 6 retrieved results for each query. The ground truths are marked by the green box.

[8] Xintong Han, Zuxuan Wu, Phoenix X Huang, Xiao Zhang, Menglong Zhu, Yuan Li, Yang Zhao, and Larry S Davis, "Automatic spatially-aware fashion concept discovery," in *ICCV*, 2017.

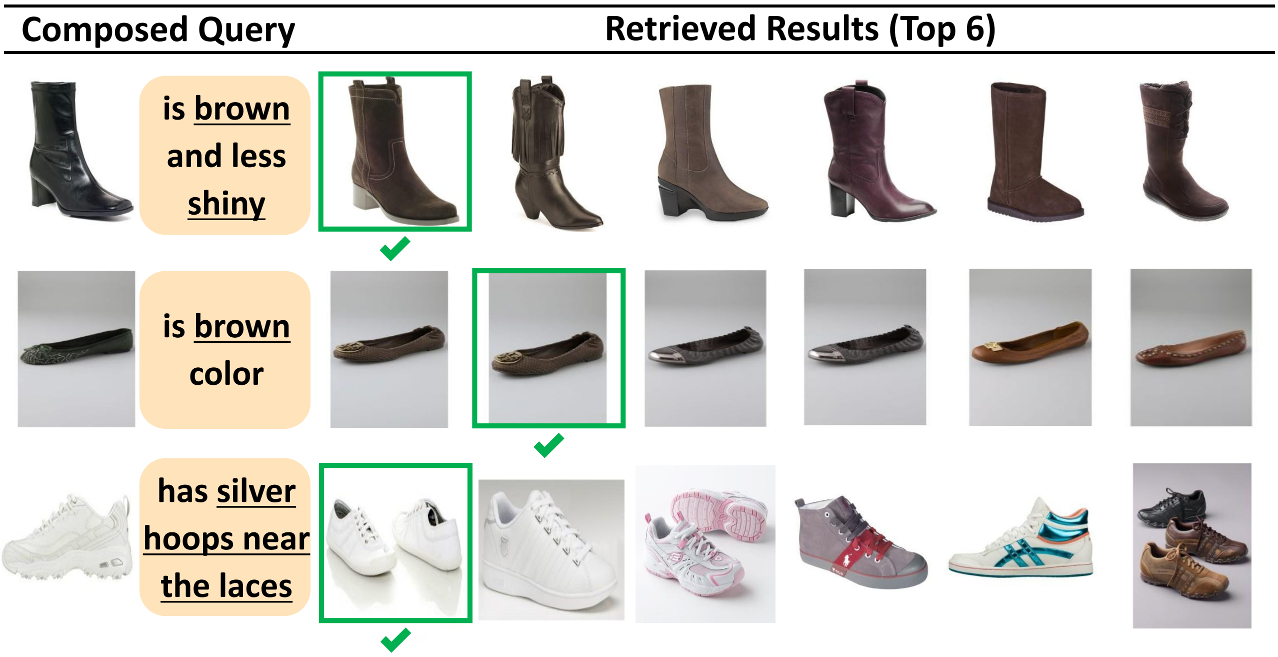

Fig. 7: Qualitative results for composed queries in Shoes dataset [6, 7]. We show the top 6 retrieved results for each query. The ground truths are marked by the green box.

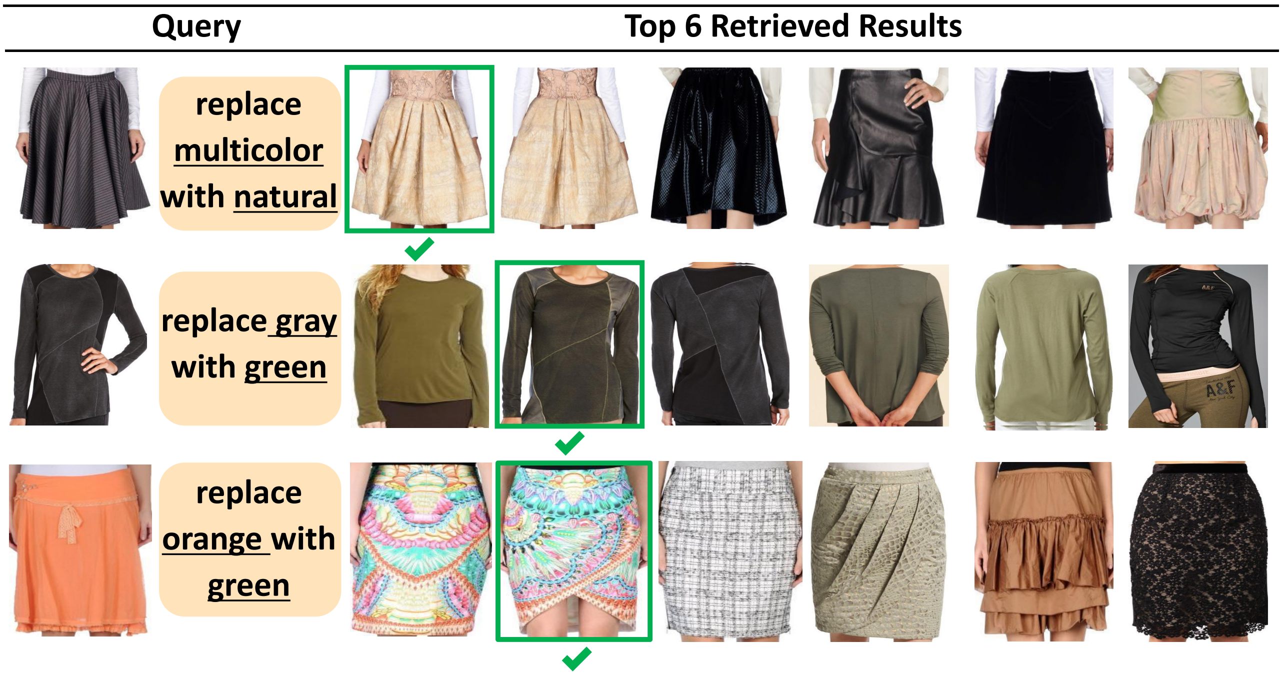

Fig. 8: Qualitative results for composed queries in Fashion200k dataset [8]. We show the top 6 retrieved results for each query. The ground truths are marked by the green box.

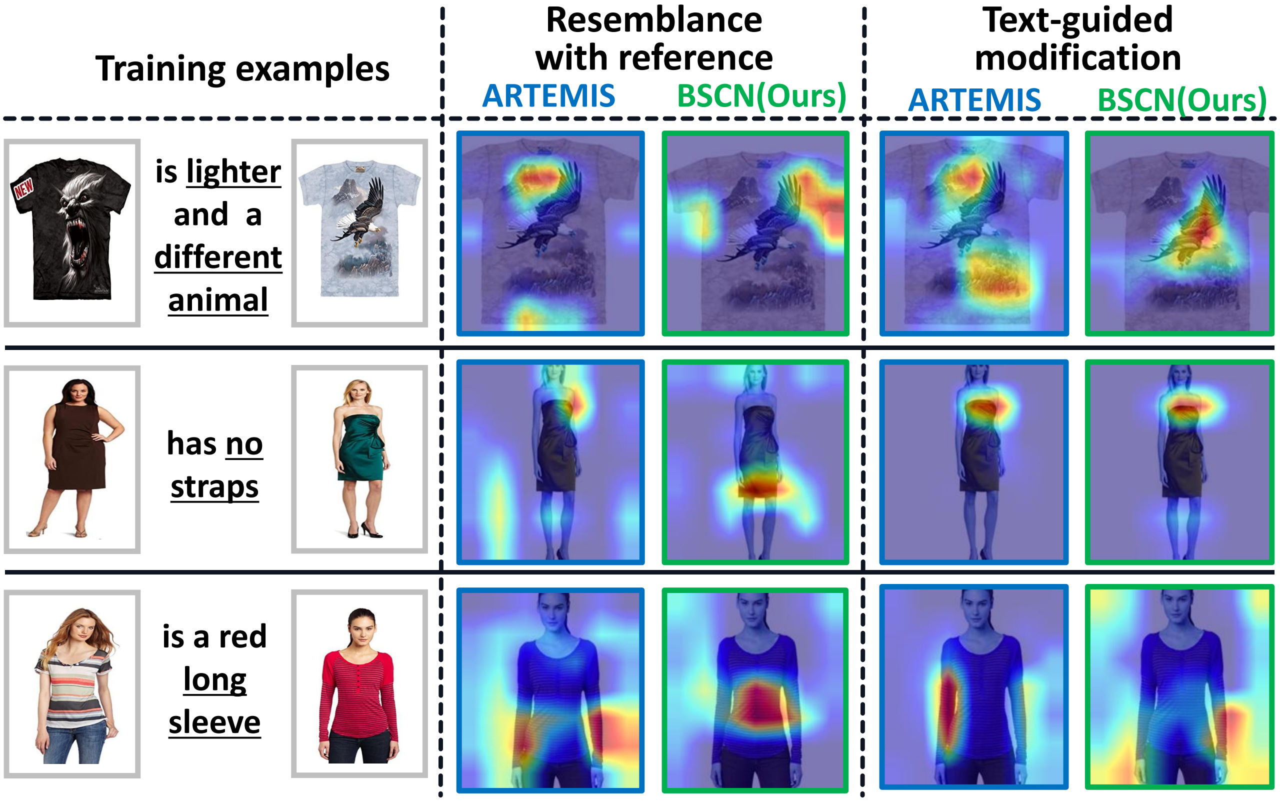

Fig. 9: Comparisons of the learned semantics with ARTEMIS.

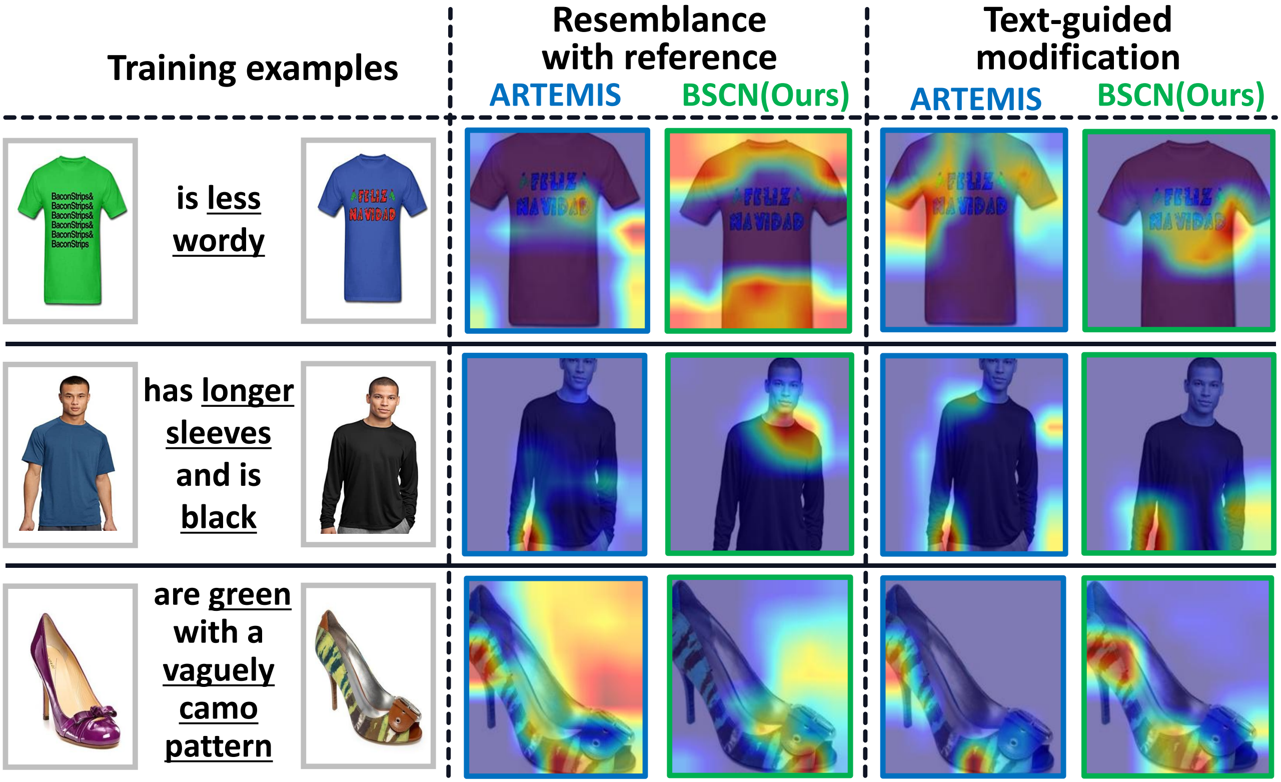
Training examples
Resemblance with reference
Text-guided modification
ARTEMIS
BSCN(Ours)
ARTEMIS
BSCN(Ours)
is less wordy
has longer sleeves and is black
are green with a vaguely camo pattern

Fig. 10: Comparisons of the learned semantics with ARTEMIS.

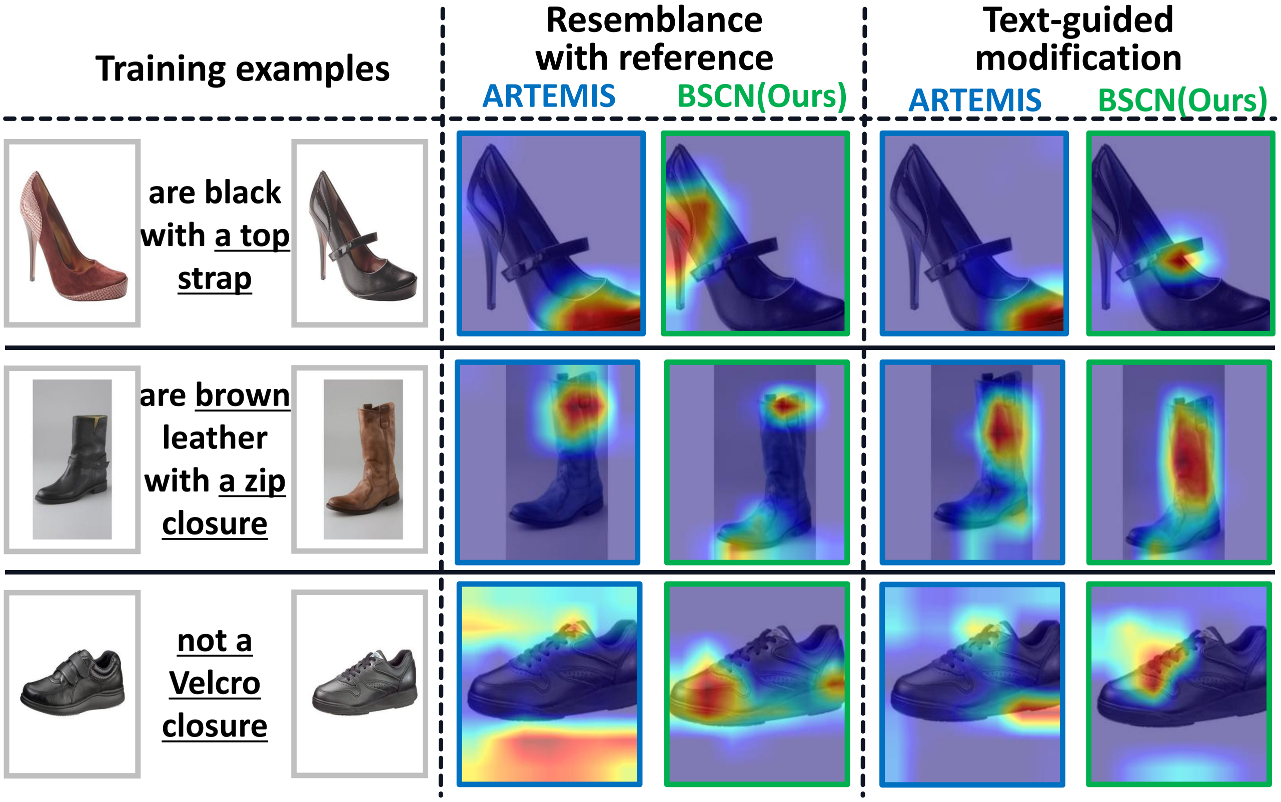

Fig. 11: Comparisons of the learned semantics with ARTEMIS