Text2Outfit: Controllable Outfit Generation with Multimodal Language Models Supplementary Material

1. Text-to-Outfit Retrieval On Disjoint Set

In the Table 1 of the main text, we report the performance on the non-disjoint set. Here, we report the performance on the dis-joint testing set (cf. Table 1). We observe similar behavior as the non-disjoint set. Our approach achieves better feature matching and outfit compatibility than GILL and single item retrieval approaches.

2. Seed-to-outfit retrieval on other categories and using other LLMs for evaluation

In the Table 2 of the main text, we report the performance using Claude-3. Here, we also conduct experiments of using other LLMs (e.g., Pixtral Large from Mistral AI), which show similar trend as Claude-3 where our results significantly outperform the baseline.

Method	Composition	Comp. score	CP. score
Seed text + prompt (baseline)	All categories	2.20	2.13
Seed text + prompt (ours)	Predicted	3.94	3.79
Seed image + prompt (baseline)	All categories	2.19	2.13
Seed image + prompt (ours)	Predicted	3.86	3.74

Table 3. Seed-to-outfit retrieval performance using Pixtral Large.

Besides reporting performance using the top category as the seed item, we also report the performance across all categories for the seed item (cf. Table 2).

3. Visualization for Attention

We visualize the attention values for top token from different approaches in Figure 1. The result shows that the mask loss facilitates the image tokens to attend to the right text descriptions.

4. Qualitative Results

We show more qualitative results of text-to-outfit generation (cf. Figure 2) and seed-to-outfit generation (cf. Figure 3).

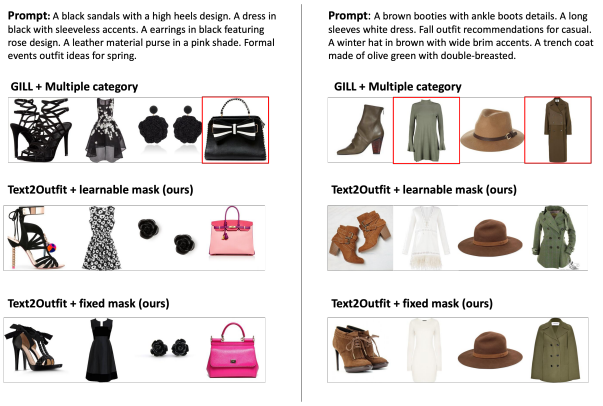

Figure 2. Results for text-to-outfit retrieval. Non-matched items are marked with the red bounding boxes.

Figure 3. Results for seed-to-outfit retrieval. The predicted composition are shown below the outfit items.

References

[1] Patrick John Chia, Giuseppe Attanasio, Federico Bianchi, Silvia Terragni, Ana Rita Magalhães, Diogo Goncalves, Ciro Greco, and Jacopo Tagliabue. Contrastive language and vision learning of general fashion concepts. *Scientific Reports*, 12(1):18958, 2022. 2

[2] Jing Yu Koh, Daniel Fried, and Russ R Salakhutdinov. Generating images with multimodal language models. *NeurIPS*, 36, 2023. 2

[3] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with

Method	Modality	Color (P@10)	Product feature (P@10)	Season (P@10)	Occasion (P@10)	Avg (P@10)	CP. score
Text-to-single-item retrieval							
CLIP [4]	Text only	95.58	69.46	90.02	76.60	82.91	64.10
Fashion-CLIP [1]	Text only	96.42	78.78	89.48	80.61	86.32	64.69
BLIP-2 [3]	Text only	92.76	66.26	85.51	83.98	82.13	66.63
CLIP [4]	Text to image	83.02	60.75	85.24	66.33	73.84	51.33
Fashion-CLIP [1]	Text to image	86.99	71.21	85.49	68.33	78.00	54.29
BLIP-2 [3]	Text to image	86.05	65.74	85.40	67.14	76.08	52.46
Text-to-outfit retrieval							
GILL [2] + multiple category	Text to image	47.80	48.08	86.59	69.09	62.89	74.64
Text2Outfit + learnable mask (ours)	Text to image	78.21	75.32	85.89	69.42	77.21	66.16
Text2Outfit + learnable mask (ours)	Text to image/text	86.05	83.65	95.44	92.90	89.51	83.38
Text2Outfit + fixed mask (ours)	Text to image/text	96.03	84.81	94.46	91.59	91.72	81.91

Table 1. Text-to-outfit retrieval performance on the dis-joint testing set.

Method	Composition	Season (P@10)	Occasion (P@10)	Comp. score (c3)	CP. score	CP. score (c3)
Seed text + prompt (ours)	Predicted	98.89	95.46	4.27	90.23	3.93
Seed image + prompt (ours)	Predicted	98.84	93.64	4.30	88.18	4.07

Table 2. Seed-to-outfit retrieval performance across all categories for the seed item.

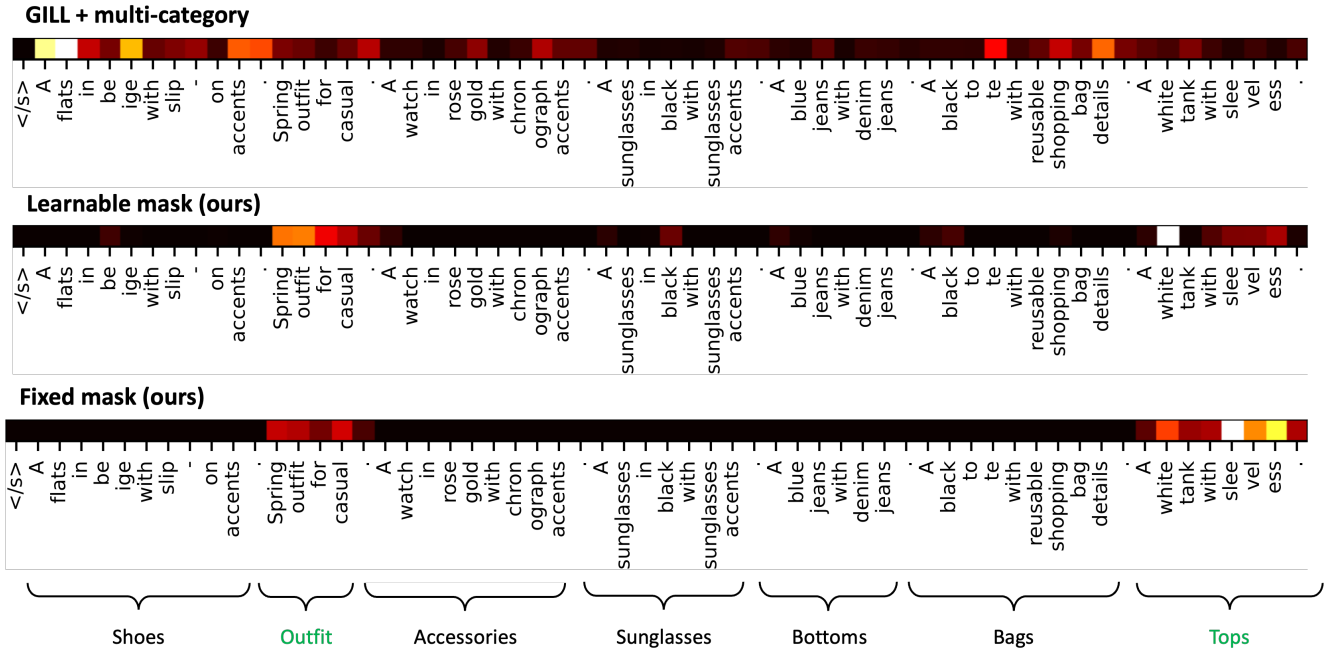

Figure 1. Visualization of the attention values for the top image token on a testing sample. Our learnable mask approach attends to the right text sections (outfit and top text sections), while GILL [2] wrongly attends to the text descriptions from other categories.

frozen image encoders and large language models. In *ICML*, pages 19730–19742. PMLR, 2023. 2

[4] 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*, pages 8748–8763. PMLR, 2021. 2