Masked Language Prompting for Generative Data Augmentation in Few-shot Fashion Style Recognition

Yuki Hirakawa
ZOZO Research
yuki.hirakawa@zozo.com

Ryotaro Shimizu
ZOZO Research
ryotaro.shimizu@zozo.com

Abstract

Constructing datasets for fashion style recognition is challenging due to the subjectivity and ambiguity of style concepts. Recent text-to-image (T2I) models enable generative data augmentation, but existing methods conditioned on class names or captions often fail to balance visual diversity and fidelity. We propose Masked Language Prompting (MLP), which masks words in a reference caption and uses large language models to generate diverse yet coherent completions. These prompts guide T2I models to synthesize labeled images without fine-tuning, making MLP suitable for trend-sensitive concepts such as fashion styles. Experiments on FashionStyle14 show that MLP outperforms class-name and caption-based baselines under limited supervision. Quantitative and qualitative analyses demonstrate that MLP produces substantially more diverse images, validating its effectiveness as a data augmentation strategy.

(a) Original (b) RandAug. (c) AutoAug.
(d) Class (e) Caption (f) MLP (ours)

Figure 1. **Comparison of data augmentation methods.** (a) shows the reference real image. (b) and (c) correspond to conventional pixel-level augmentations. (d), (e), and (f) illustrate generative augmentations based on the class label, the caption of the reference image, and our proposed MLP.

1. Introduction

Fashion style recognition [1, 8, 16, 17] is an important computer vision task with applications in recommendation and trend analysis. However, it remains challenging due to the complex interplay of visual attributes such as color, shape, and fabric, as well as the subjectivity and cultural dependence of style interpretation [1]. These factors make large-scale annotation costly and difficult [1, 17], leaving practitioners to train models from limited labeled data.

To mitigate data scarcity, conventional image-based augmentation methods—such as geometric and photometric transformations—are widely used. However, these pixel-level operations provide limited semantic variation (Fig. 1b, Fig. 1c), restricting their ability to model style-related differences. Recent advances in text-to-image (T2I) models [14, 15] enable generative augmentation [4, 5, 9, 18, 21, 25], which synthesizes semantically rich training samples from textual prompts. While promising for object and scene recognition, its applicability to subjective concepts such as fashion styles remains underexplored. In this work, we investigate T2I-based generative augmentation to improve few-shot fashion style recognition by generating diverse yet style-consistent images from limited labeled data.

Existing generative augmentation approaches typically rely on class-name prompts (e.g., "A photo of a woman wearing a `[CLASS]` style outfit"), which often produce semantically misaligned results for culturally grounded and polysemous styles. For example, Japanese ethnic style is characterized by traditional patterns and loose silhouettes (Fig. 1a), whereas class-name prompts generate culturally divergent outputs such as saris or Middle Eastern garments (Fig. 1d). Using reference captions as prompts (e.g., "A photo of a woman wearing `[CAPTION]`") improves fidelity but yields limited diversity (Fig. 1e).

To balance fidelity and diversity, we propose Masked Language Prompting (MLP), which selectively masks words in reference captions and leverages large language models (LLMs) to generate diverse yet semantically coher-

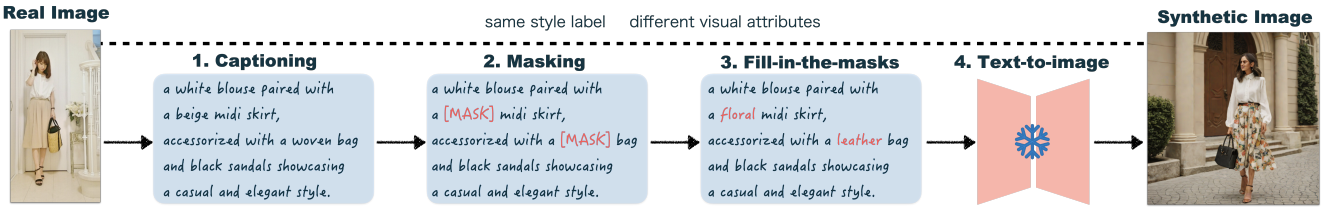

Figure 2. **Masked language prompting.** Our augmentation pipeline consists of four steps: (1) *captioning*, (2) *masking*, (3) *fill-in-the-masks*, and (4) *text-to-image generation*. This process enables effective synthesis of images that preserve the style semantics of the original reference while introducing attribute-level variations without fine-tuning T2I models.

ent prompt variants. These prompts guide T2I models to synthesize labeled images without fine-tuning. We evaluate MLP on the FashionStyle14 [17] dataset by training linear classifiers on CLIP [13] embeddings using a small set of real images augmented with synthetic data. Experimental results show that MLP consistently outperforms class-name and caption-based baselines, demonstrating its effectiveness for data augmentation in conceptually ambiguous tasks such as fashion style recognition.

2. Methodology

2.1. Generative Data Augmentation

Suppose we have a small labeled dataset $\mathcal{D}_r = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n_r}$, where $\boldsymbol{x}_i \in \mathbb{R}^d$ denotes a real image and $y_i \in \{1, \dots, C\}$ is its class label for C categories. To mitigate overfitting due to data scarcity, we introduce a set of synthetic labeled images $\mathcal{D}_s = \{(\boldsymbol{x}'_i, y'_i)\}_{i=1}^{n_s}$ generated via generative models, where $\boldsymbol{x}'_i \in \mathbb{R}^d$ denotes a synthetic image and $y'_i \in \{1, \dots, C\}$ is its class label. The extended dataset $\mathcal{D}_e = \mathcal{D}_r \cup \mathcal{D}_s$ is used to train a classifier F_Θ by minimizing the sum of cross-entropy losses $\mathcal{L}_{\text{ce}}$ over real and synthetic mini-batches:

$$\mathcal{L}(F_\Theta, \mathcal{B}_r \cup \mathcal{B}_s) = \mathcal{L}_{\text{ce}}(F_\Theta, \mathcal{B}_r) + \mathcal{L}_{\text{ce}}(F_\Theta, \mathcal{B}_s), \qquad (1)$$

where $\mathcal{B}_r \subset \mathcal{D}_r$ and $\mathcal{B}_s \subset \mathcal{D}_s$ denote mini-batches from real and synthetic data, respectively.

2.2. Masked Language Prompting

Effective generative augmentation demands synthetic images that preserve stylistic fidelity while exhibiting sufficient diversity. To address this issue, we propose Masked Language Prompting (MLP), a strategy that generates diverse yet semantically consistent textual prompts by masking parts of reference captions and completing them with LLMs. As illustrated in Fig. 2, given a labeled sample $(\boldsymbol{x}, y) \in \mathcal{D}_r$, our pipeline consists of four steps: (1) generating a detailed caption using LLMs that describes visual attributes such as color, garment type, and style; (2) identifying and masking key attribute words (*e.g.* color, material, category); (3) completing the masked caption with contextually appropriate alternatives via LLMs; and (4) synthesizing an image from the completed prompt using T2I models. The resulting synthetic image $\boldsymbol{x}'$ retains the style semantics y of the original image $\boldsymbol{x}$ while introducing controlled attribute-level variations, thereby enhancing data diversity without compromising label consistency. We design the prompt for the *captioning* step to elicit fine-grained descriptions of the clothing worn by the person in the image, covering attributes such as color, category, and design [20]. In the *fill-in-the-masks* step, we design the prompt to accept a partially masked caption as a placeholder and use the LLMs to complete it in a contextually appropriate manner, resulting in a natural variant of the original description. A key advantage of MLP is that it leverages the prior knowledge of LLMs to generate task-relevant augmentations without requiring fine-tuning of T2I models.

3. Experiments

Data Augmentation. We compare our MLP with standard augmentations, including RandAugment [3], AutoAugment [2], CutMix [22], Mixup [23], and AugMix [6]. For generative augmentation, we compare three prompt types—Class, Caption, and our MLP—following the formats: "*A photo of a woman wearing a* `[CLASS]` *style outfit.*," "*A photo of a woman wearing* `[CAPTION]`." and "*A photo of a woman wearing* G_{MLP}(`[CAPTION]`).", respectively. Here, G_{MLP} masks part of a caption and completes it with LLMs. Both *captioning* and *fill-in-the-masks* steps are performed using GPT-4o [12], with the temperature fixed at 0.0 to suppress hallucinations. In the *masking* step, 50% of nouns and adjectives are randomly replaced with `[MASK]` tokens. Finally, we synthesize 512 images per style with SDXL-Turbo [10].

Classification Models. We fine-tune a single linear layer on top of the frozen penultimate layer of CLIP-ViT-B/16 [13] using the AdamW [11] optimizer with a learning rate of 1×10^{-4} and weight decay of 1×10^{-2}. Batch sizes are set to $\min(32, n_r)$ for real images and 512 for synthetic images, respectively.

Shot	Method	N/A	RandAug.	AutoAug.	CutMix	Mixup	AugMix	Avg.
1	N/A	0.347	0.362	0.351	0.346	0.320	0.368	0.349
	Class	0.366	0.393	0.376	0.394	0.370	0.389	0.381
	Caption	0.405	0.425	0.423	0.415	0.390	0.422	0.413
	MLP (ours)	**0.424**	**0.447**	**0.445**	**0.432**	**0.410**	**0.447**	**0.434**
2	N/A	0.471	0.510	0.485	0.493	0.472	0.490	0.487
	Class	0.472	0.513	0.492	0.511	0.484	0.500	0.495
	Caption	0.537	0.559	0.545	0.542	0.533	0.551	0.545
	MLP (ours)	**0.551**	**0.568**	**0.559**	**0.555**	**0.546**	**0.560**	**0.557**
4	N/A	0.588	0.607	0.590	0.592	0.577	0.594	0.591
	Class	0.565	0.596	0.569	0.586	0.571	0.581	0.578
	Caption	0.627	0.645	0.640	0.629	0.620	0.638	0.633
	MLP (ours)	**0.630**	**0.651**	**0.647**	**0.640**	**0.628**	**0.640**	**0.639**
8	N/A	0.677	0.698	0.679	0.689	0.672	0.681	0.682
	Class	0.636	0.655	0.637	0.655	0.631	0.651	0.644
	Caption	0.698	0.706	0.700	0.703	0.693	0.704	0.701
	MLP (ours)	**0.703**	**0.714**	**0.706**	**0.712**	**0.696**	**0.709**	**0.707**
16	N/A	0.739	**0.760**	0.745	0.752	0.744	0.753	0.749
	Class	0.696	0.702	0.687	0.712	0.678	0.701	0.696
	Caption	0.751	0.753	0.752	0.752	0.743	0.750	0.750
	MLP (ours)	**0.755**	0.759	**0.758**	**0.756**	**0.745**	**0.759**	**0.755**

Table 1. **Mean classification accuracy of linear classifiers trained with different data augmentation methods for each shot.** Rows correspond to generative augmentation settings, where "N/A" denotes the use of real images only, while columns indicate standard augmentation settings, with "N/A" denotes no standard augmentation applied. Underlined values indicate (1) augmentation-free baselines trained on real data without augmentation. **Bold** highlights (2) the best result within each shot and prompting methods. Red marks (4) the best overall result for each shot. Blue denotes (3) performance lower than the augmentation-free baseline.

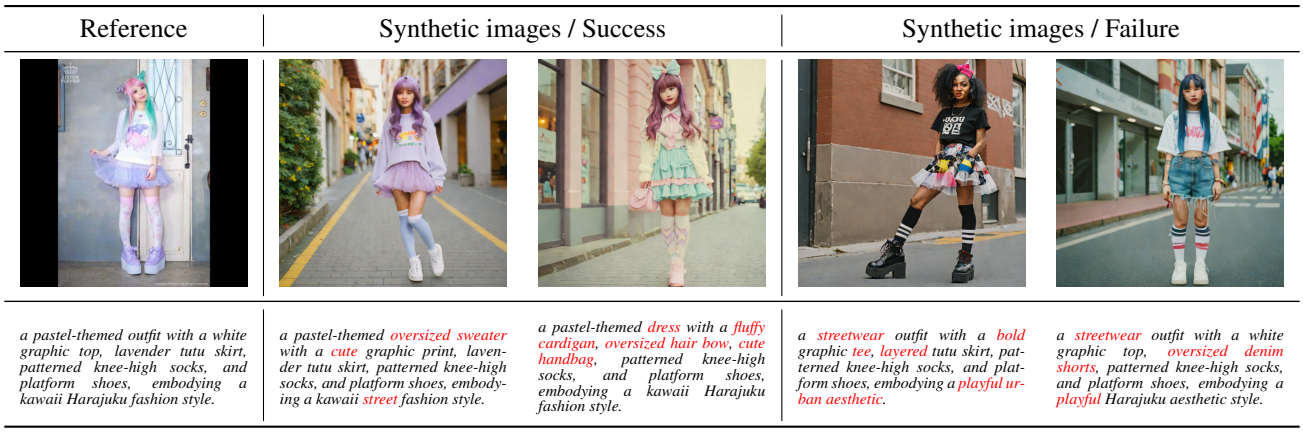

Figure 3. **Visualization of image-caption pairs generated using MLP, including both successful and failure cases.** The leftmost column shows a reference *fairy* style image with a GPT-generated caption. Columns 2-5 display image-caption pairs generated by MLP conditioned on the reference, where columns 2-3 successfully capture key characteristics of the *fairy* style—*e.g.*, aesthetics of cuteness, delicacy, and soft color tones—while columns 4-5 illustrate failure cases that instead exhibit traits more aligned with the *street* style, such as a rough appearance and a casual silhouette. Words highlighted in red indicate those completed during the *fill-in-the-masks* step.

3.1. Few-shot Classification Results

Tab. 1 presents classification accuracy for each shot.

Vs. standard augmentations. MLP without non-generative augmentation outperforms the real-only baselines with standard augmentation in all shot settings except

(a) *ethnic* (b) *fairy*

Figure 4. **Wordclouds of LLM-generated completions for masked captions.** Words are aggregated per class from LLMs outputs under the $n_{\text{shot}} = 16$ setting, showing how `[MASK]` tokens are replaced during *fill-in-the-masks* step.

Method	$\bar{D}_{\text{SSIM}}$ (↑)	$\bar{D}_{\text{LPIPS}}$ (↑)	CMMD(↓)		
			1	4	16
Class	**0.424**	0.436	3.516		
Caption	0.398	0.418	3.203	2.146	**1.937**
Ours	0.414	**0.439**	**2.596**	**2.060**	1.963

Table 2. **Quantitative evaluation of synthetic images.** We assess diversity using SSIM and LPIPS, and quality using CMMD. As Class does not use reference images, its CMMD is constant; the corresponding cells are highlighted.

for $n_{\text{shot}} = 16$. This result highlights the advantage of MLP, especially under extremely limited supervision.

Vs. Class & Caption. Within the column labeled "N/A", MLP attains the highest accuracy across all shots, underscoring the importance of balancing fidelity and diversity in generative data augmentation. Class often underperforms relative to the Real Only baseline, indicating the importance of fidelity of synthetic images.

Combination with standard augmentations. Combining MLP with RandAug. or AugMix further improves performance, suggesting that MLP offers complementary diversity.

3.2. Analysis on Synthetic Images

Quantitative Evaluation. We design *reference-conditioned diversity*, $D^{(r)}_{\text{SSIM}}$ and $D^{(r)}_{\text{LPIPS}}$, to assess the diversity among images generated from the same reference image.

$$D^{(r)}_{\text{SSIM}} = 1 - \frac{2}{N(N-1)} \sum_{1 \le i < j \le N} \text{SSIM}(x^{(r)}_i, x^{(r)}_j), \quad (2)$$

$$D^{(r)}_{\text{LPIPS}} = \frac{2}{N(N-1)} \sum_{1 \le i < j \le N} \text{LPIPS}(x^{(r)}_i, x^{(r)}_j), \quad (3)$$

where N denotes the number of generated samples per reference, and $x^{(r)}_i$ is the i-th image synthesized from the r-th reference. For each reference, we compute the Structural Similarity Index Measure (SSIM) [19] and the Learned Perceptual Image Patch Similarity (LPIPS) [24] over all intra-set pairs of generated samples, which quantify pixel-level structural diversity and feature-level semantic diversity, respectively. The final diversity scores, denoted as $\bar{D}_{\text{SSIM}}$ and $\bar{D}_{\text{LPIPS}}$, are obtained by averaging over all R reference images. In our experiments, we prepare $R = 624$ reference images evenly sampled from 13 fashion styles from FashionStyle14 dataset. For each reference image, we generate $N = 32$ synthesized images and conduct the evaluation. To evaluate the fidelity of synthetic images, we employ the CLIP Maximum Mean Discrepancy (CMMD) [7], which measures distributional divergence between real and generated images in the CLIP embedding space. For each fashion style, CMMD is computed between the real test samples from the FashionStyle14 dataset and synthetic samples generated under $n_{\text{shot}} = 1, 4, 16$. The final CMMD score is obtained by averaging across all styles. Tab. 2 reports the evaluation results of diversity and fidelity of synthetic images. From the evaluation results of $\bar{D}_{\text{SSIM}}$ and $\bar{D}_{\text{LPIPS}}$, we observe that MLP achieves higher diversity than Caption and comparable diversity to Class. In terms of fidelity, MLP produces distributions closer to real images than those generated by Class, as measured by CMMD. These results highlight the effectiveness of our framework in producing diverse yet style-consistent images, and demonstrate its positive impact on classification performance.

Qualitative Evaluation. Fig. 4 shows word frequencies of completions as a word cloud. For the *fairy* style, terms such as "*cute*" and "*pastel*" appear frequently, reflecting its characteristic aesthetics of cuteness, delicacy, and soft color tones. For the *ethnic* style, words like "*bohemian*," "*floral*," and "*brimmed*" are common, capturing traditional and ethnic designs. These results suggest that GPT tailors completions to the target style. However, as shown in columns 4–5 of Fig. 3, incorrect completions (*e.g.*, replacing "*pastel-themed*" with "*streetwear*") can yield images misaligned with the intended style. We leave developing methods to estimate each token's contribution to style and avoid masking highly influential words as future work.

4. Conclusion

We propose Masked Language Prompting (MLP), a generative augmentation method for few-shot fashion style recognition. By masking words in reference captions and leveraging LLMs for contextual completion, MLP generates prompts that produce images with high diversity and style consistency. Experiments show that MLP outperforms class-name and caption baselines, and further improves performance when combined with traditional methods. Moreover, analysis on synthetic images shows that MLP enhances diversity over caption-based prompts and better aligns with real data than class-name prompts. In future work, we will construct a multi-label fashion style dataset for more realistic evaluation.

References

[1] Hyosun An, Kyo Young Lee, Yerim Choi, and Minjung Park. Conceptual framework of hybrid style in fashion image datasets for machine learning. *Fashion and Textiles*, 10 (1):18, 2023. 1

[2] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. pages 113–123, 2019. 2

[3] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. pages 702–703, 2020. 2

[4] Lisa Dunlap, Alyssa Umino, Han Zhang, Jiezhi Yang, Joseph E. Gonzalez, and Trevor Darrell. Diversify your vision datasets with automatic diffusion-based augmentation. 2023. 1

[5] Ruifei He, Shuyang Sun, Xin Yu, Chuhui Xue, Wenqing Zhang, Philip Torr, Song Bai, and Xiaojuan Qi. Is synthetic data from generative models ready for image recognition? 2022. 1

[6] Dan Hendrycks*, Norman Mu*, Ekin Dogus Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple method to improve robustness and uncertainty under data shift. 2020. 2

[7] Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking fid: Towards a better evaluation metric for image generation. pages 9307–9315, 2024. 4

[8] Mohammad Hadi Kiapour, Kota Yamaguchi, Alexander C. Berg, and Tamara L. Berg. Hipster wars: Discovering elements of fashion styles. pages 472–488, 2014. 1

[9] Jae Myung Kim, Jessica Bader, Stephan Alaniz, Cordelia Schmid, and Zeynep Akata. Datadream: Few-shot guided dataset generation. pages 252–268, 2024. 1

[10] Shanchuan Lin, Anran Wang, and Xiao Yang. Sdxl-lightning: Progressive adversarial diffusion distillation. *arXiv preprint arXiv:2402.13929*, 2024. 2

[11] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. 2019. 2

[12] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2

[13] 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. pages 8748–8763, 2021. 2

[14] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv*, abs/2204.06125, 2022. 1

[15] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. pages 10684–10695, 2022. 1

[16] Edgar Simo-Serra and Hiroshi Ishikawa. Fashion style in 128 floats: Joint ranking and classification using weak data for feature extraction. pages 298–307, 2016. 1

[17] Moeko Takagi, Edgar Simo-Serra, Satoshi Iizuka, and Hiroshi Ishikawa. What Makes a Style: Experimental Analysis of Fashion Prediction. 2017. 1, 2

[18] Shijian Wang, Linxin Song, Ryotaro Shimizu, Masayuki Goto, and Hanqian Wu. Attributed synthetic data generation for zero-shot image classification. In *Synthetic Data for Computer Vision Workshop @ CVPR 2024*, 2024. 1

[19] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. 13(4):600–612, 2004. 4

[20] Xuewen Yang, Heming Zhang, Di Jin, Yingru Liu, Chi-Hao Wu, Jianchao Tan, Dongliang Xie, Jue Wang, and Xin Wang. Fashion captioning: Towards generating accurate descriptions with semantic rewards. 2020. 2

[21] Zhou Yongchao, Sahak Hshmat, and Ba Jimmy. Using synthetic data for data augmentation to improve classification accuracy. 2023. 1

[22] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. pages 6023–6032, 2019. 2

[23] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. 2018. 2

[24] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. 2018. 4

[25] Yifan Zhang, Daquan Zhou, Bryan Hooi, Kai Wang, and Jiashi Feng. Expanding small-scale datasets with guided imagination. 2023. 1