# Multi-Class Textual-Inversion Secretly Yields a Semantic-Agnostic Classifier

Kai Wang[1], Fei Yang[3,4*], Bogdan Raducanu[1,2], Joost van de Weijer[1,2]
[1]Computer Vision Center [2] Universitat Autònoma de Barcelona, Spain
[3] VCIP, College of Computer Science, Nankai University, China
[4] Nankai International Advanced Research Institute (SHENZHEN· FUTIAN), China

## Supplementary Material

## A. Boarder Impacts

The utilization of personalized text-to-image (T2I) models holds promise for a diverse array of applications in various domains. Our model seeks to enhance the dual functionality of the updated tokens in these models. However, it is imperative to acknowledge potential risks, including the dissemination of misinformation, potential misuse, and the introduction of biases. Ethical considerations and broader impacts require a thorough examination to ensure responsible utilization of these models and their capabilities.

## B. Limitations

One limitation of our approach is its dependence on the Textual Inversion technique, neglecting the thorough exploration into methods that fine-tune diffusion backbones. Moreover, our method necessitates the features of all few-shot samples during training, which might not be available in scenarios where such information is not provided in advance. Lastly, *MC-TI* may introduce increased time complexity in datasets featuring thousands of categories.

## C. Experiments

**Dataset details.** The detailed statistics of all the datasets, are shown in Table S1. The initialization tokens for *TI* [6] and *MC-TI* are also detailed in the table.

**Mean and Standard deviations.** The experimental mean and standard deviation values are provided in Table S2. These results further confirm the robustness of *MC-TI* to randomness.

**Generative performance.** The full table for generation quality comparison with CLIP-Similarity is shown in Table S3, which further demonstrates that *MC-TI* is not damaging the generation capability.

**Additional textual feature visualization.** Additionally, we visualize the textual feature changes from *TI* to *MC-TI* on four datasets, as depicted in Fig. S2. We extract the textual features of 27 prompt templates from 5-shot learning schemes. As observed, *MC-TI* enhances the clustering of textual characteristics by enforcing discriminative regularization terms, resulting in improved classification performances.

**Extended comparison with prompt tuning methods.** In Fig.S1, we provide a comparison of our method, *MC-TI*, with two classical prompt tuning methods, namely CoOp [16] and CLIP-Adapter [7]. It's important to note that these prompt tuning methods primarily operate on the CLIP (ResNet-50) backbones. This comparison serves to further demonstrate that our method is comparable to current prompt tuning approaches.

## References

[1] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European Conference on Computer Vision*, pages 446–461. Springer, 2014. 2

[2] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014. 2

[3] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 215–223. JMLR Workshop and Conference Proceedings, 2011. 2

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2

[5] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pages 178–178. IEEE, 2004. 2

[6] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image gen-

[*] Corresponding author: feiyang@nankai.edu.cn

Table S1. The detailed statistics of datasets used in experiments.

| Dataset | classes | Train size | Test size | Task | Initialization token |
|---|---|---|---|---|---|
| Oxford-Pets [14] | 37 | 2944 | 3669 | Fine-grained pet recognition | "pet" |
| Flowers [13] | 102 | 4093 | 2463 | Fine-grained flowers recognition | "flower" |
| Food101 [1] | 101 | 50500 | 30300 | Fine-grained food recognition | "food" |
| Aircrafts [12] | 100 | 3334 | 3333 | Fine-grained aircraft recognition | "aircraft" |
| Stanford-Cars [9] | 196 | 6509 | 8041 | Fine-grained car recognition | "car" |
| CIFAR10 [10] | 10 | 50000 | 10000 | Generic object recognition | "object" |
| STL10 [3] | 10 | 1000 | 8000 | Generic object recognition | "object" |
| Caltech101 [5] | 102 | 4128 | 2465 | Generic object recognition | "object" |
| DTD [2] | 47 | 2820 | 1692 | texture recognition | "texture" |
| EuroSAT [8] | 10 | 13500 | 8100 | Satellite image recognition | "object" |
| UCF101 [15] | 101 | 7639 | 3783 | Action recognition | "action" |
| ImageNet-1000 [4] | 1000 | 1.28M | 50000 | Generic object recognition | "object" |

Table S2. We present the mean ± standard deviation of the evaluation accuracies (%) achieved by *MC-TI* on four fine-grained datasets. This analysis demonstrates that randomization does not significantly influence the performance.

| Method | *MC-TI* (Ours) | | | | | |
|---|---|---|---|---|---|---|
| $N$-shot | 1 | 2 | 4 | 5 | 8 | 16 |
| Oxford-Pets | 65.2±0.4 | 77.8±0.5 | 84.6±1.5 | 88.7±0.7 | 89.8±1.8 | 91.7±0.5 |
| Flowers | 80.3±0.7 | 87.3±0.4 | 91.8±0.3 | 93.1±0.9 | 94.8±0.8 | 95.9±0.1 |
| Food101 | 53.6±0.9 | 68.6±1.3 | 77.6±1.4 | 80.4±0.7 | 82.2±1.9 | 86.0±1.7 |
| Aircrafts | 24.9±1.9 | 32.2±2.1 | 39.0±1.8 | 40.0±1.0 | 45.5±0.8 | 49.2±2.2 |

eration using textual inversion. *International Conference on Learning Representations*, 2023. 1, 3

[7] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2):581–595, 2024. 1, 3

[8] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019. 2

[9] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013. 2

[10] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 2

[11] Alexander C Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. *Proceedings of the International Conference on Computer Vision*, 2023. 3

[12] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 2

[13] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pages 722–729. IEEE, 2008. 2

[14] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3498–3505. IEEE, 2012. 2

[15] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 2

[16] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 1, 3

Table S3. Comparison between *MC-TI* and *TI* in image generation across eleven datasets by computing the CLIP similarity (%) between the training few-shot samples and the generated images of both methods. Superior scores are highlighted in **bold**.

|  | CLIP-Similarity (*MC-TI*/ *TI*) |  |  |  |  |  |
|---|---|---|---|---|---|---|
| $N$-shot | 1 | 2 | 4 | 5 | 8 | 16 |
| Oxford Pets | **82.3** / 81.8 | **81.3** / 80.6 | **81.2** / 80.0 | **81.0** / 80.6 | **81.5** / 80.3 | **81.1** / 80.5 |
| Flowers | **81.1** / 81.0 | 81.8 / **81.9** | 81.7 / **81.8** | 81.9 / **82.3** | 82.7 / **83.0** | 82.8 / **83.2** |
| Food101 | **77.3** / 75.8 | **77.9** / 77.3 | 77.7 / **78.6** | 77.5 / **78.3** | 77.4 / **78.7** | 77.3 / **78.8** |
| Aircrafts | **77.0** / 76.4 | **76.1** / 76.0 | 76.2 / **76.5** | 76.4 / **76.5** | 76.3 / **76.8** | 76.2 / **76.8** |
| Stanford Cars | **78.9** / 77.9 | **78.7** / 78.5 | 78.6 / **78.8** | **78.7** / 78.3 | **78.5** / 78.4 | **78.7** / 78.5 |
| CIFAR10 | **65.4** / 62.2 | **65.8** / 63.2 | **67.2** / 64.1 | **65.7** / 64.5 | **65.4** / 64.7 | **66.3** / 64.8 |
| STL10 | **75.0** / 73.2 | **74.7** / 72.5 | **70.1** / 65.6 | **70.6** / 68.2 | **71.1** / 68.4 | **69.3** / 67.8 |
| Caltech101 | **75.8** / 73.7 | **76.4** / 73.9 | **75.2** / 74.6 | **75.5** / 74.2 | **75.1** / 73.9 | **74.7** / 73.7 |
| DTD | **78.5** / 77.7 | **73.8** / 73.2 | **72.8** / 72.5 | **72.6** / 72.2 | **73.3** / 72.8 | **72.8** / 72.0 |
| EuroSAT | 57.6 / **58.7** | **60.7** / 60.1 | 57.9 / **58.8** | **60.9** / 60.7 | 59.1 / **59.6** | 58.6 / **58.7** |
| UCF101 | 61.3 / **61.8** | 62.1 / **62.9** | **63.3** / 63.2 | 62.8 / **62.9** | 62.0 / **62.7** | **63.2** / 62.4 |

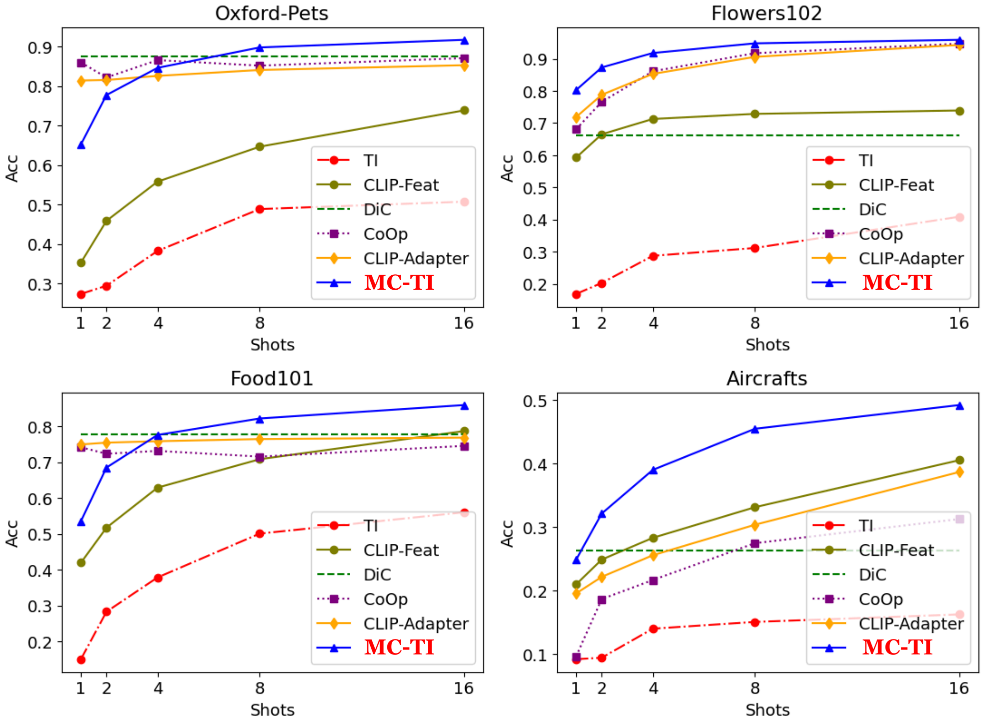

Figure S1. *MC-TI* is compared with the Textual Inversion (*TI*) [6], the CLIP-feat baseline, Diffusion Classifier (*DiC*) [11] and two prompt tuning methods (CoOp [16] and CLIP-Adapter [7]) over four fine-grained datasets by computing classification accuracies. We vary the $N$-shot ($N = 1, 2, 4, 8, 16$) numbers to draw the trend plots.

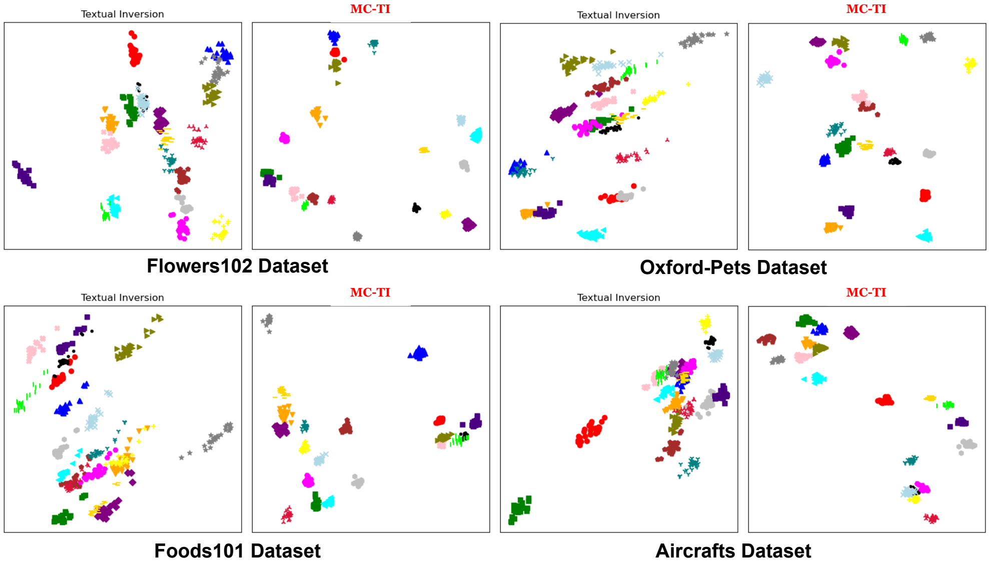

Figure S2. To visualize the textual prompts features, we took the 5-shot conceptual tokens learned by Textual Inversion and *MC-TI*, respectively. By applying 27 types of various prompt templates, we visualize the PCA components in 2-D maps.