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

Abstract

With the advent of large pre-trained vision-language models such as CLIP, prompt learning methods aim to enhance the transferability of the CLIP model. They learn the prompt given few samples from the downstream task given the specific class names as prior knowledge, which we term as semantic-aware classification. However, in many realistic scenarios, we only have access to few samples and no knowledge of the class names (e.g., when considering instances of classes). This challenging scenario represents the semantic-agnostic discriminative case. Text-to-Image (T2I) personalization methods aim to adapt T2I models to unseen concepts by learning new tokens and endowing these tokens with the capability of generating the learned concepts. These methods do not require knowledge of class names as a semantic-aware prior. Therefore, in this paper, we first explore Textual Inversion and reveal that the new concept tokens possess both generation and classification capabilities by regarding each category as a single concept. However, learning classifiers from single-concept textual inversion is limited since the learned tokens are suboptimal for the discriminative tasks. To mitigate this issue, we propose Multi-Class textual inversion, which includes a discriminative regularization term for the token updating process. Using this technique, our method MC-TI achieves stronger Semantic-Agnostic Classification while preserving the generation capability of these modifier tokens given only few samples per category. In the experiments, we extensively evaluate MC-TI on 12 datasets covering various scenarios, which demonstrates that MC-TI achieves superior results in terms of both classification and generation outcomes.

1. Introduction

Leveraging extensive datasets of image-text pairs, trained visual-language models (VLMs) [1, 21, 35] encapsulate critical general knowledge. Different from traditional representation learning based on discretized labels, the alignment between image and text features endows the VLMs with superior generalization capabilities for downstream tasks. While VLMs are effective in extracting both visual and textual descriptions, their training requires large-scale, high-quality datasets. To circumvent this issue, prompt learning methods [43, 44] adapt a pre-trained VLM (e.g., CLIP [35]) to downstream tasks. They have demonstrated impressive performance across a variety of few-shot and zero-shot visual recognition cases. Despite their effectiveness, these prompting methods typically require knowledge of *class names* for context optimization, which we term *semantic-aware* classification. This requirement potentially limits the applicability of these methods in realistic scenarios where class names are not determined.

To learn token representations for new concepts, recent T2I personalization methods [11, 24, 39] propose to adapt a given T2I diffusion model with user-provided images and associating the new concept with a unique token as their own *"names"*. Consequently, the adapted model can generate various renditions of the new concept guided by text prompts. However, these learned tokens only address generative paradigms. As mentioned in the seminal paper [20], generative approaches are also crucial for discrimination tasks. Therefore, it should be possible to *unify* both discriminative and generative paradigms in the T2I model personalization case, such a problem setup is shown in Fig. 1. A recent work, named Diffusion Classifier (*DiC*) [25], first examines how T2I diffusion models compare against discriminative models. Nonetheless, the *DiC* method requires knowledge of category names as a *semantic-aware* prior, similar to the prompt learning methods. Consequently, these methods are not suitable for discrimination tasks involving unknown concepts that are difficult to name before training, which we refer to as the *semantic-agnostic* classification. Moreover, each unknown concept typically has only *few-shot* samples available, which further complicates the task.

To achieve *semantic-agnostic* classification with few samples, we first explore a naive scenario where we directly employ tokens learned with Textual Inversion (*TI*) for classi-

*Corresponding author: feiyang@nankai.edu.cn

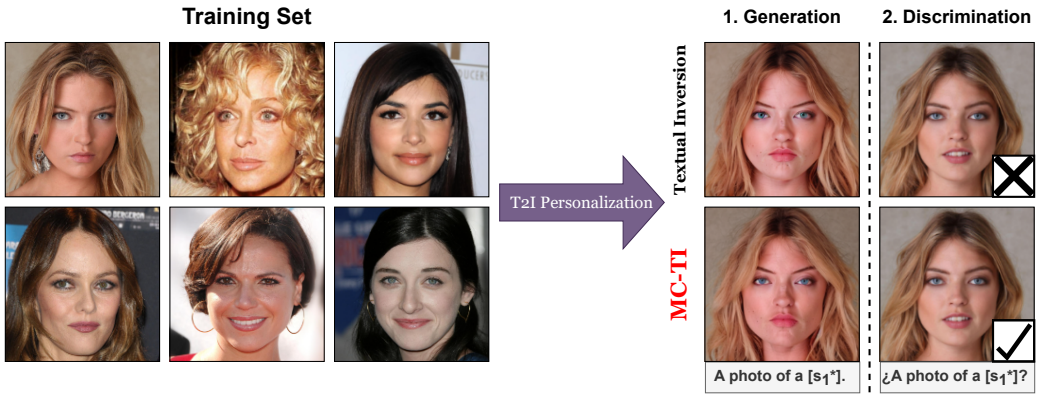

Figure 1. While both existing Single-Concept Textual Inversion and our proposed Multi-Class Textual Inversion (*MC-TI*) can generate satisfactory results with *few samples* per person, the single-concept *TI* lacks the ability to ensure discrimination performance. This is because it does not constrain token updates in a discriminative manner. (CelebA faces [29].)

fication (as shown in Fig. 2 for 5-shot classification), which exhibits much lower accuracies compared with *DiC* but still demonstrates the potential applicability in the classification task. Furthermore, observing that Custom Diffusion [24] behaves poorly in the classification task, we hypothesize that this discrepancy is due to interference caused by backbone fine-tuning. In this paper, we build on the frozen backbone method *TI* as our primary baseline.

On the basis of the above observations, we conclude that the current personalization algorithms primarily emphasize the generation ability of new tokens. More specifically, they update the tokens solely based on the noise reconstruction loss. While these approaches provide more freedom in updating the tokens, it may compromise their discriminative capabilities (as shown in Fig. 4). To enhance the training process and guide these tokens to be applicable to the *semantic-agnostic* classification, we propose *Multi-Class textual inversion* (instead of the existing single-concept textual inversion). In our method, we propose constraining the updating paths to follow a *discrimination-regularized* direction. In other words, the tokens are required to be updated in a manner that leads to a better distribution to achieve discriminative textual prompt representations. More specifically, in each training step, we randomly sample image features for each category (concept). These features contribute to composing a temporary classifier for the current textual prompt, which already includes the learnable token. Subsequently, we compute the classification probability between the textual prompt and these image features based on their cosine similarities. To ensure discrimination, a cross-entropy loss is applied to the probability scores, serving as a *discriminative regularization*. This loss is added to the original ϵ-prediction loss from the personalization approaches.

Our method, designated as Multi-Class Textual Inversion (*MC-TI*) is evaluated across twelve recognition datasets covering various recognition scenarios. It demonstrates remarkable performance in *Semantic-Agnostic Classification* tasks with few-shot samples. Notably, even compared to the strong *semantic-aware* zero-shot method, namely CLIP (ViT-L/14) [35], which rely on class names trained over millions of image-text pairs, *MC-TI* outperforms them with as few as 1-shot (Flowers, EuroSAT) and up to 8-shot (Stanford-Cars, Aircrafts, DTD). As a summarization, our method has the following main contributions:

- We are the *first* paper to explore *semantic-agnostic classification* of T2I personalized tokens. By this means, we unify the generative and discriminative paradigms upon T2I diffusion adaptation.
- We propose *Multi-Class textual inversion* that includes a simple but effective discriminative regularization term, which is defined as a cosine cross-entropy loss, to augment the personalization method for discrimination.
- We conducted experiments over twelve benchmarks covering diverse recognition tasks and various N-shot settings. These experimental evaluations demonstrate the efficiency of our proposed *MC-TI*.

2. Related Work

Generative models for classification works [20,32,36] have emphasized the importance of modeling data distributions to facilitate discriminative learning. Additionally, in the realm of generative modeling, efforts have been made to acquire efficient representations for classification tasks, as evidenced by works like MAE [17], BERT [9], etc. However, these endeavors typically involve joint training for discriminative and generative modeling or fine-tuning generative representations for downstream tasks. A recent contribution, known as the *DiC* [25], takes a different approach by directly using the Stable Diffusion [37] for discrimination tasks. This method computes text-conditional likelihoods by estimating noise reconstruction losses, demonstrating promising performance compared to discriminative approaches in zero-shot scenarios. However, the *DiC* incurs significantly longer diffusion

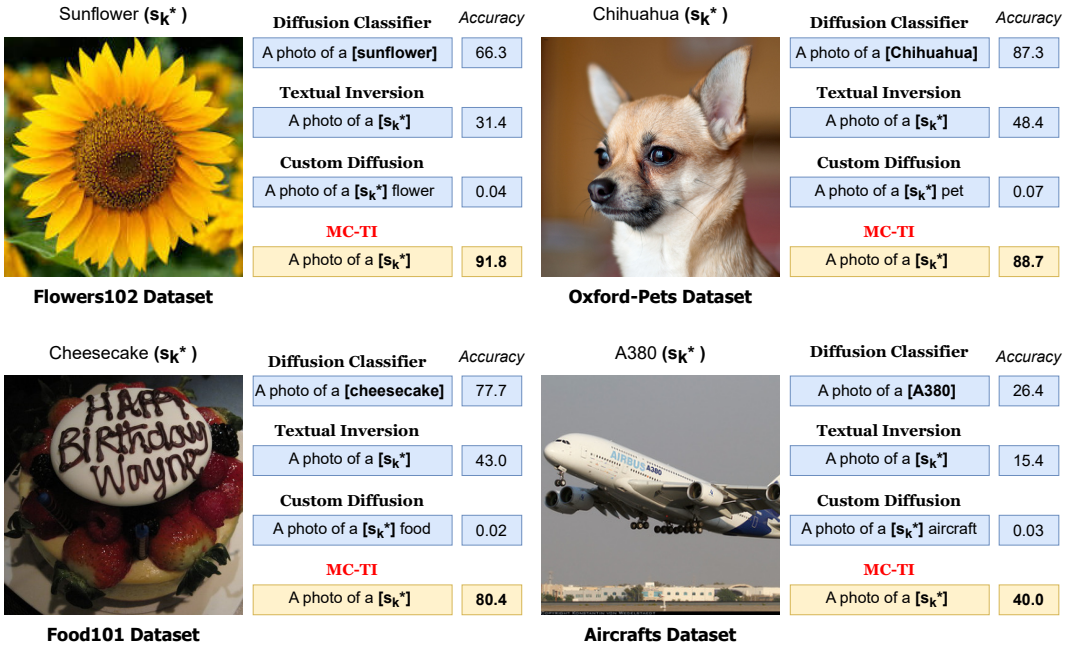

Figure 2. Diffusion Classifier [25] approach classifies samples by computing the text-conditional likelihoods and requires knowledge of the *category names*. By comparison, Textual Inversion (*TI*) [11] and *MC-TI* only need to learn the concept tokens with few-shot samples (*5-shot* in the figure examples). *MC-TI* further strengthen the *TI* by augmenting with a discriminative regularization term and significantly improve the performance. Custom Diffusion [24] is one of the best personalization methods by fine-tuning the UNet, it works poorly in classification.

sampling times for each input image, rendering it impractical for real-time classification with large numbers of samples. Furthermore, it also imposes the prerequisite of knowing the class names as a *semantic-aware* prior.

Text-to-Image models personalization aims at adapting a given model to a *new concept* by giving users images and bind the new concept with a unique token. As a result, the adaptation model can generate various renditions of the new concept guided by text prompts. Depending on whether the adaptation method is fine-tuning the T2I model, they are categorized into two main streams: The freezing stream focuses on learning new concept tokens instead of fine-tuning the generative models. Textual Inversion [11] is a pioneering work focusing on finding new pseudo-words by personalizing the text embedding space. Recent methods [14, 15, 26, 27, 41] also belong to this technique stream. The most representative methods of fine-tuning stream include DreamBooth [39] and Custom Diffusion [24], where the pretrained T2I model such that it learns to bind a unique identifier with that specific subject given few images. Following research [3, 12, 16, 28] further extend this pipeline and improve the generation quality. Although existing T2I model adaptation methods have been successful in learning new concepts from a set of relevant images, they have overlooked that personalized concept tokens have gained semantic information from the relevant images. These inherent semantic information from these tokens is also applicable to image classification tasks, as shown in Fig. 2. To avoid misalignment of text and image pairs, we adopt the branch of the T2I model freezing, that is, the Textual Inversion [11] method as the backbone.

Prompt Learning is to adapt the pretrained Visual-Language Models (VLMs) to the downstream tasks, prompt learning always applies task-related textual tokens to infer the task-specific textual knowledge. For example, the hand-crafted template "a photo of a [CLASS]" in CLIP [35] is used to model the textual embedding for zero-shot prediction with knowing the "[CLASS]" as the *semantic-aware* prior. However, the hand-crafted prompts have fewer ability to describe the downstream task because they do not consider the specific knowledge of the current task. To address the above problem, Context Optimization (CoOp) [43] replaces hand-crafted prompts with a soft learning prompt inferred by the labeled few-shot samples. PLOT [4] proposes to apply optimal transport to match the vision and text modalities. Following these pioneering works, recent methods [13, 42, 44] continue to improve the prompt learning performance.

However, these prompt-tuning methods have the limitation that they require *semantic-aware* prior knowledge of the concept names before training. Furthermore, prompt learning involves the update of the network on *all samples* collected in each training time, and does not support *parallel* training as our method *MC-TI*. More importantly, the learned context tokens from these methods cannot be further applied in *image generation*.

3. Method

3.1. Preliminaries

T2I Diffusion Models. In this paper, we utilize Stable Diffusion [37] as our backbone model, which is a latent diffusion model (LDM). The model comprises two primary components: an autoencoder and a diffusion model applied in the latent space. The encoder $\mathcal{E}$ within the autoencoder segment of the LDMs maps an input image $\mathcal{I}$ to a latent code $z_0 = \mathcal{E}(\mathcal{I})$, while the decoder reverses this process, reconstructing the original image as $\mathcal{D}(\mathcal{E}(\mathcal{I})) \approx \mathcal{I}$. The diffusion model can be conditioned on various factors such as class labels, segmentation masks, or textual input. Let $\tau(y)$ denote the conditioning mechanism, which maps a condition y to a corresponding conditional vector for LDMs. The LDM model is updated using the noise reconstruction loss, also known as the ϵ-prediction loss:

$$L_{LDM} = \mathbb{E}_{z_0 \sim \mathcal{E}(x), y, \epsilon \sim \mathcal{N}(0,1)} \underbrace{\|\epsilon - \epsilon_\theta(z_t, t, \tau(y))\|_2^2}_{\mathcal{L}_{mse}}. \quad (1)$$

The neural network backbone ϵ_θ typically adopts a conditional UNet architecture [38], responsible for predicting the added noise. In text-guided diffusion models, the objective is to generate an image based on a combination of random noise z_T and a conditional input prompt $\mathcal{P}$. To distinguish from the general conditional notation in LDMs, we represent the textual condition as $\mathcal{C} = \tau_\phi(\mathcal{P})$. τ_ϕ refers to a CLIP text encoder pretrained on millions of text-image pairs.

T2I model adaptation. Given a pretrained T2I diffusion model, adaptation methods [11, 24, 39] integrate a new concept into the model using few images and their associated text descriptions. Typically, this involves learning new tokens through text encoding.

Single-Concept Textual Inversion. Given the target concept with few images, a text caption is required. For personalization purpose where the target concept is a unique instance of a general category, we introduce a new modifier token $\mathcal{V}^*$ associated with the pseudo-word S^* for the concept. During training, $\mathcal{V}^*$ is initialized with the embeddings of a single-word coarse descriptor of the object and injected to a prompt template of the form "A photo of a S^*", "A nice photo of the small S^*", etc. Then the optimization goal with the LDM loss $L_{TI} = L_{LDM}$ is:

$$\mathcal{V}^* = \arg\min_{\mathcal{V}} \mathbb{E}_{z_0 \sim \mathcal{E}(x), y, \epsilon \sim \mathcal{N}(0,1)} \mathcal{L}_{mse}. \quad (2)$$

During this optimization, the network parameters are frozen and only the $\mathcal{V}^*$ tokens are learned.

3.2. *MC-TI*: Multi-Class Textual Inversion

Semantic-Agnostic Classification. Suppose we already have a dataset with *multiple concepts* as K classes, each class is given with only N-shot ($N = 1, 2, 4, 5, 8, 16$) samples for training purposes, which are denoted as $\mathcal{I}_k^n, n \in [1, N], k \in [1, K]$. Importantly, we lack *semantic* information about these classes, which means that we do not know their class names in advance. This *semantic-agnostic* scenario occurs when the classes we aim to learn do not have semantic labels (or their semantic labels are unknown to the language model, e.g. person names). This is the case with instances of human faces, very specialized fields such as fine-grained class names, etc. Then we aim to learn a specific token $\mathcal{V}_k^*$ for each class pseudo-word S_k^* separately, and these tokens $\mathcal{V}_k^*$ can help us to discriminate images from the same dataset distributions while maintaining their generation capability. Training solely with the noise ϵ-prediction loss in *single-concept* Textual Inversion (*TI*) can only ensure that the learned tokens are able to generate the desired concepts. However, as the tokens learned from the pretrained diffusion model already implicitly include the semantic information of the given images, they can already achieve rough discrimination tasks (as seen in Fig. 2). Nonetheless, the *TI* approach does not constrain the optimization direction of these tokens with discriminative guidance. This results in a lower classification performance than for diffusion-based classifiers, like *DiC* [25]. The PCA visualization depicted in Fig. 4 illustrates the textual features learned by the *TI*. In particular, these features often overlap among categories, indicating a lack of distinctiveness. To address this issue and improve classification accuracy, we propose to apply a *discriminative regularization* to achieve *Multi-Class textual inversion* (*MC-TI*).

Discriminative Regularization. For convenience, we denote the CLIP image encoder as π_ψ, which is trained paired as the CLIP text encoder τ_ϕ. During the token learning for concept $\mathcal{V}_k^*$, we first inject $\mathcal{V}_k^*$ into a prompt template and obtain its corresponding textual feature as $g_k = \tau_\phi(\mathcal{P}^{\mathcal{V}_k^*})$ we randomly sample *one image* for each class out of the *multiple concepts* in the few-shot training set $\mathcal{I}_j^n, j \in [1, K]$ and pass them to the image encoder for feature extraction as $f_j = \pi_\psi(\mathcal{I}_j^n)$. In this way, we have a group of features $f_1, f_2, ..., f_j, ..., f_K$ representing all K classes. Regarding these features as temporal prototypes, we propose to compute the prediction probability for the textual feature as follows:

$$P(\hat{y} = j | g_k) = \frac{s \cdot exp(cos(g_k, f_j))}{\sum_{j=1}^{K} s \cdot exp(cos(g_k, f_j))}, \quad (3)$$

where s is the scale factor as the reciprocal of the softmax temperature and $\hat{y}$ is the probability prediction for the textual feature with token $\mathcal{V}_k^*$. Then the cross-entropy loss is:

$$\mathcal{L}_{reg} = -\sum_{j=1}^{K} y_j \cdot logP(\hat{y} = j | g_k). \quad (4)$$

Here y denote the one-hot groundtruth. Finally, the loss for

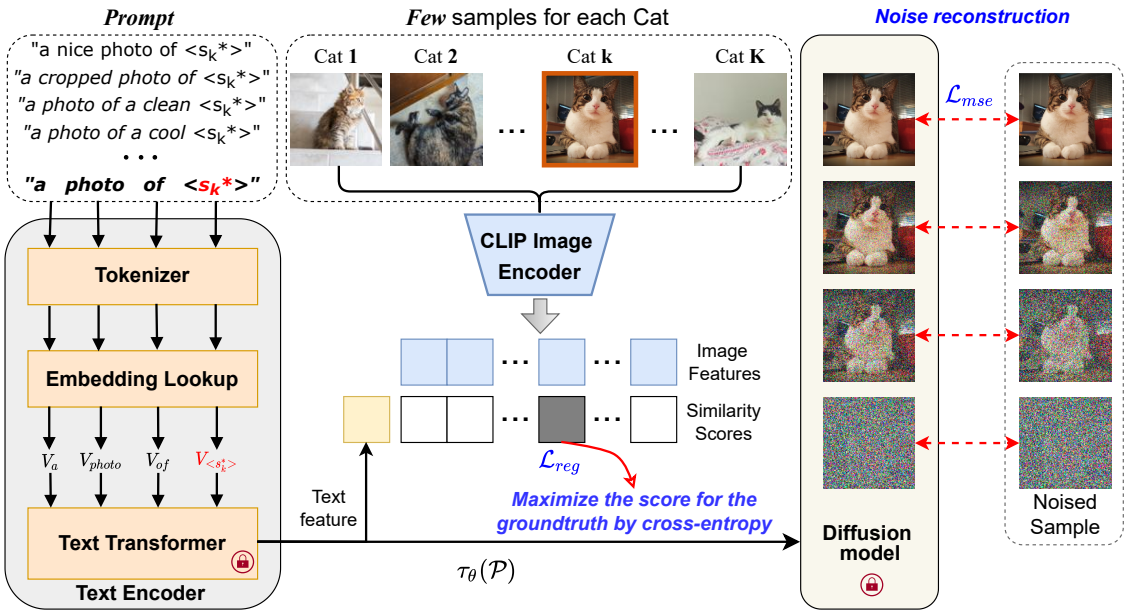

Figure 3. Illustration of our *MC-TI* approach. During the training of token $<s_k^*>$ for concept k (cat k for example), we add a discriminative regularization term $\mathcal{L}_{reg}$, which is defined as the cosine cross-entropy from the current training text feature to the image features.

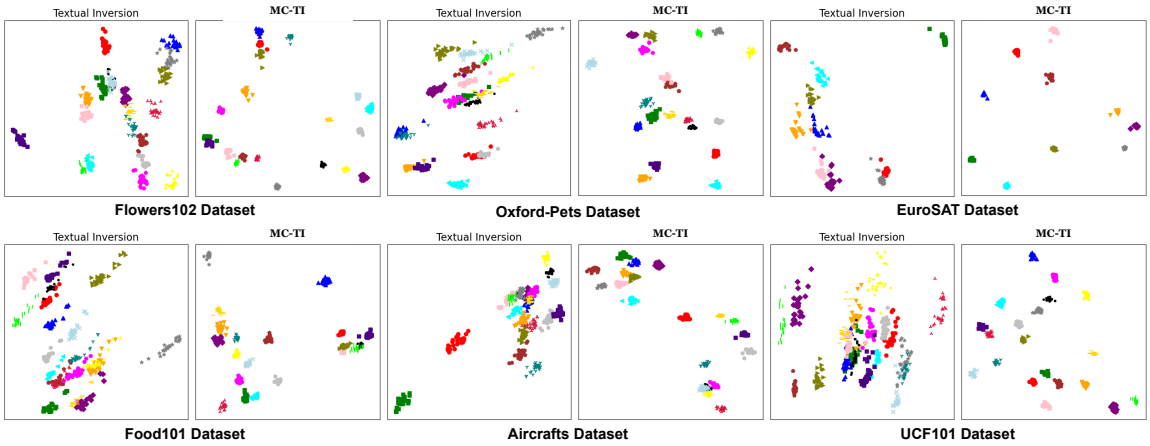

Figure 4. To visualize the textual prompts features, we took the 5-shot conceptual tokens learned by Textual Inversion and *MC-TI*, respectively. By applying 27 types of various prompt templates, we visualize the PCA components in 2-D maps for 20 categories out of these six datasets. *MC-TI* improves the clustering of textual characteristics by enforcing discriminative regularization terms.

training the token $\mathcal{V}_k^*$ is defined as:

$$\mathcal{V}_k^* = \arg\min_{\mathcal{V}_k} \mathbb{E}_{z_0 \sim \mathcal{E}(x), y, \epsilon \sim \mathcal{N}(0,1)} \left[\alpha \mathcal{L}_{mse} + \beta \cdot \mathcal{L}_{reg} \right]. \tag{5}$$

The α, β work as the trade-off parameters to balance these two losses. Note that each token $\mathcal{V}_k^*$ is updated separately for each concept S_k^*. Therefore, we can obtain a set of $\mathcal{V}_k^*, k \in [1, K]$ in *parallel*. After applying our robust regularization terms, the textual features are better distributed for future discrimination tasks, as shown in Fig. 4.

Inference. During inference time, each sample $\mathcal{I}$ is classified by predicting its probability of belonging to all the learned *multiple concepts* (tokens). To achieve this, we first inject all category tokens $\mathcal{V}_k^*$ into the prompt template $T =$ "a photo of a S_k^*" and then extract the textual features as $g_k = \tau_\phi(\mathcal{P}^{\mathcal{V}_k^*}), k \in [1, K]$. The input image with feature as $f_\mathcal{I} = \pi_\psi(\mathcal{I})$ is categorized as:

$$P(\bar{y} = k | f_\mathcal{I}) = \frac{s \cdot exp(cos(g_k, f_\mathcal{I}))}{\sum_{l=1}^{K} s \cdot exp(cos(g_l, f_\mathcal{I}))}. \tag{6}$$

Where the $\bar{y}$ is the prediction probabilities for the input $\mathcal{I}$. Moreover, given that *MC-TI* is orthogonal to prompt learning methods, we can enhance our classification performance by integrating these approaches. In this study, we specifically employ the CoOp [43] to learn a unified context.

4. Experiments

4.1. Experimental setups

Implementation Details. We utilize the Stable Diffusion (SD) v1.5 [37] as our backbone. Within the SD model, we employ the text branch of CLIP (ViT-L/14) [35] as the text encoder τ_ϕ, and we additionally utilize the image encoder π_ψ to extract features for the training samples. We also include SD v1.4 and v2.0 in the ablation study. For *single-concept* Textual Inversion (*TI*), we adhere to the same training scheme as outlined in the original paper, conducting training for 3000 steps. Subsequently, *MC-TI* initializes with the learned tokens from Textual Inversion as a warm-up, followed by an additional 100 steps of continuous updating using both the loss of noise reconstruction and the loss of regularization. We configure the learning rate as 5e-4 and utilize the AdamW optimizer [30]. Default hyperparameters are set as $\alpha = 1.0$, $\beta = 1.0$, and $s = 10.0$. Additionally, to expedite training, we extract image features beforehand, ensuring that *MC-TI* does not introduce higher time complexity compared to the *TI* approach. During inference, we adhere to the standard CLIP practice, employing the textual template T_1 as "A photo of a $S_k^*, k \in [1, K]$". We also investigate the impact of different prompt templates during inference. To enhance *MC-TI* with CoOp [43], we learn a unified context for each dataset with $M = 16$ context tokens. All the experiments are conducted on A40 GPUs.

Datasets. We evaluated the performance of the few-shot classification in 12 datasets: Oxford-Pets [34], Flowers [33], Food101 [2], Aircrafts [31], Stanford-Cars [22], CIFAR10 [23], STL10 [7], Caltech101 [10], DTD [5], EuroSAT [19], UCF101 [40] and ImageNet [8]. The datasets chosen for evaluation form a comprehensive benchmark, encompassing a wide array of vision tasks ranging from generic object and scene classification to fine-grained categorization, as well as specialized tasks such as texture recognition and satellite image analysis. Detailed statistics for each dataset are provided in the Appendix. During the experiments, we *randomly select* N-shot ($N = 1, 2, 4, 5, 8, 16$) samples from the training split of each dataset as the N-shot train set. The classification performance is then evaluated on the test split.

Comparison methods. To assess the effectiveness of *MC-TI*, we compare with several types of methods: (i) Four diffusion classification baselines from the Diffusion Classifier (*DiC*) [25]: Synthetic SD Data, SD features, *DiC* and DM-ZSC [6] approach. (ii) CLIP (ResNet-50) and CLIP (ViT-L/14) zero-shot performance as comparisons. (iii) prompt learning approaches, including CoOp [43], ProGrad [44] and PLOT [4] based on RN50 [18]. (iv) *single-concept* Textual Inversion (*TI*) [11] as a baseline for the T2I personalization method. (v) CLIP-feat is another baseline where we obtain prototype-based classifiers with image features from the CLIP (ViT-L/14) model. Note that the first three groups of comparison methods are *Semantic-Aware*, which means that the *class names* are required before training. These last two groups are *Semantic-Agnostic* as *MC-TI*.

Evaluation metrics. We assess the performance across various N-shot scenarios ($N = 1, 2, 4, 5, 8, 16$) using classification accuracy on the test set as the primary metric for discriminative performance evaluation. Additionally, we calculate the CLIP similarity, a widely used metric in T2I personalization methods [11, 12, 24, 28], which measures the distance between the generated images by *TI*/*MC-TI* and the few-shot training samples. Specifically, for both *TI* and *MC-TI*, we randomly generate ten examples with the same prompt "a photo of a S^*" for each category and compute the average CLIP similarity. This evaluation metric serves to verify the generation performance by assessing whether the learned tokens can successfully *reconstruct* the original concepts. We utilize CLIP (ViT-L/14) for this purpose, consistent with the Stable Diffusion v1.5 setup. All experimental results are averaged over three runs, with standard deviations provided in the Appendix.

4.2. Experimental Results

Discriminative performance with classification accuracies are shown in Table 1, where our method (*MC-TI* and *MC-TI* +CoOp) are compared with five groups of comparison methods. Across the datasets examined in *DiC*, our approach *MC-TI* consistently outperforms with no more than five samples. Remarkably, in nine datasets, our method achieves peak performances with 16-shot samples per class, exceeding even the powerful pretrained model CLIP (ViT-L/14). This trend is particularly pronounced in the Flowers and EuroSAT datasets, where we outperform the competition with CLIP (ViT-L) by 1.6% and 0.2% respectively, using only one-shot samples. In the remaining datasets, including Food101, STL10 and ImageNet, we also rank as the second-best, trailing only behind the robust model CLIP (ViT-L). Furthermore, *MC-TI* always surpasses *DiC* and DM-ZSC even with no more than 5-shot examples. Moreover, the CoOp [43] context optimization can further enhance the performance of *MC-TI* in most datasets.

The tendency curves evaluated over three datasets, depicting the performance with increasing N-shot, are illustrated in Fig. 5. Here we compare our method with semantic-agnostic baselines (*TI*, CLIP-feat) and prompting approaches. By comparison, our method *MC-TI* consistently outperforms them by no more than 4 shots, especially over the Aircraft dataset. And the CoOp prompting can always improve *MC-TI* under various N-shot setups.

Generative performance is shown in Table 2. In this paper, we aim to improve the discriminative capability of the newly learned tokens. However, here we verify that our adaptation does not negatively impact the generations. The comparison reveals that the additional regularization term $\mathcal{L}_{reg}$ does not

Table 1. *MC-TI* is compared with various approaches. CLIP(RN50) and CLIP(ViT-L) are listed as references. Note that the concepts in CLIP are learned from millions of text-image pairs. We highlight the best performance with **bold** font and the second with *underlines*.

Method	*MC-TI* (Ours)						Ours + CoOp	TI	CLIP ViT-L Feat.	CoOp	ProG	PLOT	Synth.	SD Feat.	*DiC*	DM-ZSC	CLIP RN50	CLIP ViT-L
Semantic Prior	✗									✓								
N-shot	1	2	4	5	8	16	16	16	16	16	16	16	Zero	Full	Zero	Zero	Zero	Zero
Ox. Pets	65.2	77.8	84.6	88.7	89.8	91.7	**94.1**	50.7	73.8	87.2	89.0	87.0	31.3	75.9	87.3	72.5	85.4	*93.5*
Flowers	80.3	87.3	91.8	93.1	94.8	**95.9**	*95.8*	40.9	73.9	94.8	94.4	94.5	22.1	70.0	66.3	-	65.9	78.7
Food101	53.6	68.6	77.6	80.4	82.2	86.0	*88.2*	56.0	78.7	77.1	78.4	74.5	12.6	73.0	77.7	71.6	81.1	**92.9**
Aircrafts	24.9	32.2	39.0	40.0	45.5	*49.2*	**51.5**	16.3	40.6	31.5	31.1	31.4	9.4	35.2	26.4	-	19.3	36.1
Cars	54.9	65.8	71.5	73.1	77.6	*79.5*	**79.9**	43.3	72.1	72.8	73.5	73.6	-	-	-	-	55.8	77.3
CIFAR10	61.4	78.9	86.6	91.6	91.7	93.4	**96.5**	31.2	76.6	-	-	-	35.3	84.0	88.5	72.1	75.6	*96.2*
STL10	61.8	91.9	94.4	97.5	98.4	98.7	*98.8*	65.4	93.8	-	-	-	38.0	87.2	95.4	92.8	94.3	**99.3**
Caltech	79.8	85.4	89.3	89.9	92.4	**93.2**	*93.0*	59.8	87.4	92.2	92.2	92.0	-	-	-	-	82.1	92.6
DTD	37.3	46.5	52.4	56.9	60.5	**65.5**	*65.1*	34.7	54.8	63.3	64.0	62.5	-	-	-	-	41.7	55.3
EuroSAT	60.1	62.0	71.3	72.3	72.8	79.2	**85.2**	13.6	69.2	82.2	*83.7*	83.6	-	-	-	-	41.1	59.9
UCF101	55.0	58.2	61.6	61.9	69.4	77.2	**77.5**	33.0	62.4	76.9	*77.3*	76.9	-	-	-	-	63.6	76.2
ImageNet	41.3	53.6	63.9	66.4	71.2	*74.8*	67.9	35.0	58.9	63.0	63.5	61.9	18.9	56.6	61.4	61.9	59.6	**75.3**
Average	56.3	67.4	73.7	76.0	78.9	*82.0*	**82.8**	40.0	70.2	-	-	-	-	-	-	-	63.8	77.8

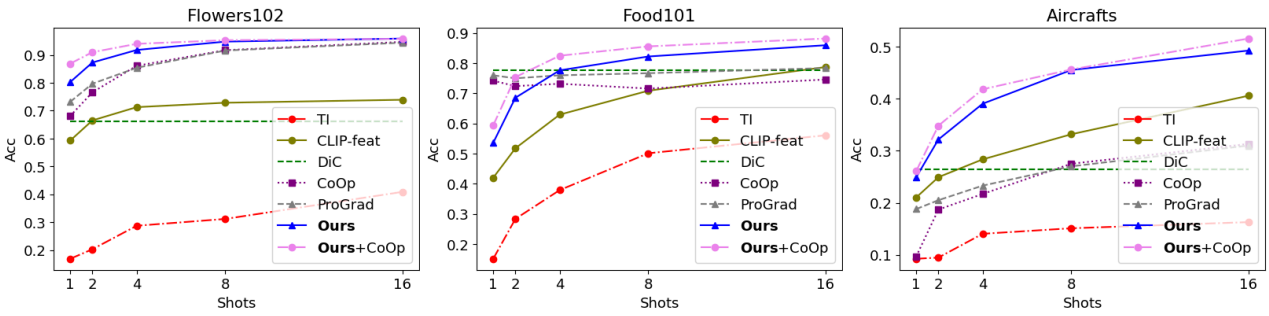

Figure 5. *MC-TI* is compared with the Textual Inversion (*TI*), the CLIP-feat baseline, Diffusion Classifier (*DiC*) and prompt learning methods (CoOp, ProGrad) by computing classification accuracies. We vary the N-shot ($N = 1, 2, 4, 8, 16$) numbers to draw the trend plots.

Table 2. Comparison between *MC-TI* and *TI* in image generation across various datasets by computing the CLIP similarity (%) between the training few-shot samples and the generated images.

	CLIP-Similarity (*MC-TI*/ *TI*)					
N-shot	1	2	4	5	8	16
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**
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

significantly affect the generation quality and may even lead to marginal gains. This conclusion is further supported by the generated samples in Fig. 6. Considering both discriminative and generative performance, *MC-TI* effectively achieves classification objectives without compromising its generative capabilities.

Time complexity. Our method exhibits similar time complexity as *TI*. For learning each concept, the single-concpet TI consumes 12 minute and *MC-TI* with 12m24s. Note that, both TI and *MC-TI* support parallel training for multiple new concepts and do not incur additional time cost in classification phase (as CLIP model). *DiC* is achieving classification based on the diffusion process, which incurs a significant sampling time. *DiC* ranges from 18s/image (Oxford-Pets) to 1000s/image (ImageNet), which leads to an unsatisfactory time complexity for real-time classification tasks.

Table 3. Ablation study on the Oxford-Pets dataset by varying the s, β, α, inference step, template TPL and also the SD versions.

HyperP		*MC-TI* (step=100, TPL=T_1, $\alpha = 1.0$)						HyperP			$s = 10.0, \beta = 1.0$
s	β	N-shot 1	2	4	5	8	16	step	TPL	α	N-shot 5
3.0	1.0	62.2	76.3	84.0	86.4	88.2	90.7	50/75	T_1	1.0	88.6 / 88.8
10.0	*1.0*	**65.2**	**77.8**	**84.6**	**88.7**	**89.8**	**91.7**	100	T_2/T_3	1.0	89.0
30.0	1.0	61.2	74.1	82.0	85.6	87.6	88.2	100	T_1	0.0	83.7
10.0	0.1	49.5	63.8	74.9	82.8	84.8	85.4	SD v1.4			88.5
10.0	10.0	63.8	76.4	84.2	87.6	88.8	91.1	SD v2.0			88.3

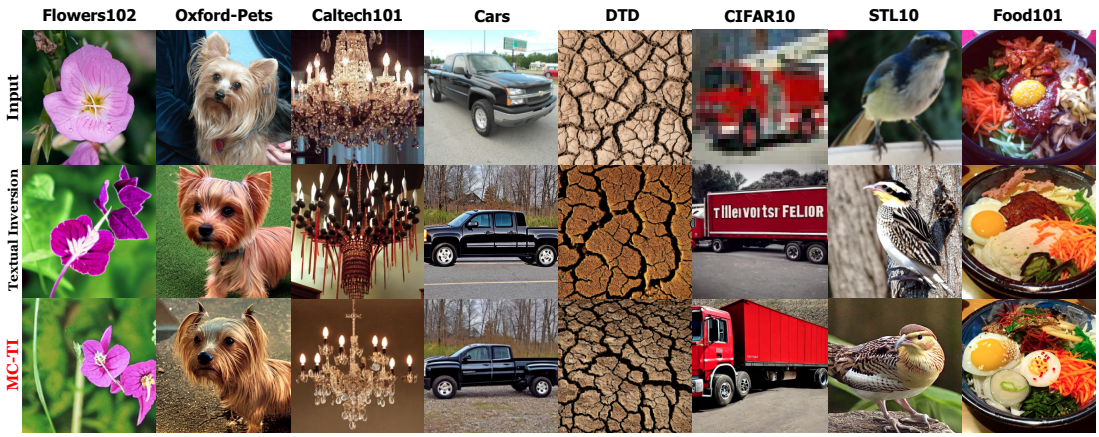

Figure 6. *MC-TI* does not compromise the generation performance of *TI* as both approaches produce objects similar to the inputs. The illustrated examples are from the 5-shot setup.

4.3. Ablation Study

Ablation studies are conducted on the Oxford-Pets dataset, with summarized results in Table 3. The scale factor s is varied from 3.0 to 30, revealing that lower values of s correspond to less performance degradation. Similarly, the hyperparameter β ranges from 0.1 to 10.0, with *MC-TI* showing a preference for higher values of β, although excessively high values can lead to a decrease in the generation performance. We also investigate the influence of varying the training steps of our method, observing minimal impact on performance, indicating that *MC-TI* converges robustly with only 100 steps. Lastly, we explore changes in textual templates during inference. By default, we use the same zero-shot classification template as in CLIP [35], denoted as T_1="a photo of a S^*". Additionally, we experiment with T_2="a photo of the nice S^*" and T_3="a cropped photo of the S^*". The results demonstrate the robustness of our method to the templates. Subsequently, we set $\alpha = 0.0$ and update the tokens solely based on the loss of regularization $\mathcal{L}_{reg}$, without initialization from *TI* and without applying the loss of noise reconstruction $\mathcal{L}_{mse}$. In this configuration, the reconstruction quality cannot be guaranteed, and we observe a performance drop of nearly 5%. This indicates that the generation quality contributes to learning meaningful semantic information, thereby influencing performance. Finally, we also assess *MC-TI* using different backbones, namely SD v1.4 and SD v2.0. The classification performances do not exhibit significant changes, as also observed in *DiC* [25].

5. Conclusion

Existing prompt learning methods face challenges while there is no *semantic-aware* prior knowledge of the few samples, where in most cases the *class names* are unknown. To learn a textual token to represent the class name, *single-concept* T2I adaptation methods excel at learning new concepts from minimal image data. However, they often neglect the *discriminative* potential of newly acquired tokens. This study delves into *single-concept Textual Inversion* as a representative of the T2I adaptation. Our investigation uncovers the dual nature of *multiple concept* tokens, possessing both generative and discriminative capabilities. However, token updates may lack directionality without proper constraints. To mitigate this issue, we introduce an additional regularization term. Our *Multi-Class Textual Inversion* method, named *MC-TI*, achieves a robust *Semantic-Agnostic Classification* by incorporating discriminative regularization while retaining the generative prowess of modifier tokens. Extensive evaluations of diverse datasets consistently show superior results in both classification and generation performance.

Acknowledgements. This work is funded by Grants TED2021-132513B-I00 and PID2022-143257NB-I00 funded by MCIN/AEI/10.13039/501100011033, by the European Union NextGenerationEU/PRTR and by ERDF A Way of Making Europa, the Departament de Recerca i Universitats from Generalitat de Catalunya with reference 2021SGR01499, and the Generalitat de Catalunya CERCA.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 1

[2] 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. 6

[3] Muhammad Atif Butt, Kai Wang, Javier Vazquez-Corral, and Joost van de Weijer. Colorpeel: Color prompt learning with diffusion models via color and shape disentanglement. In *European Conference on Computer Vision*, pages 456–472. Springer, 2025. 3

[4] Guangyi Chen, Weiran Yao, Xiangchen Song, Xinyue Li, Yongming Rao, and Kun Zhang. Prompt learning with optimal transport for vision-language models. *International Conference on Learning Representations*, 2023. 3, 6

[5] 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. 6

[6] Kevin Clark and Priyank Jaini. Text-to-image diffusion models are zero shot classifiers. *Advances in Neural Information Processing Systems*, 36, 2024. 6

[7] 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. 6

[8] 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. 6

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2

[10] 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. 6

[11] 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 generation using textual inversion. *International Conference on Learning Representations*, 2023. 1, 3, 4, 6

[12] Rinon Gal, Moab Arar, Yuval Atzmon, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Designing an encoder for fast personalization of text-to-image models. *arXiv preprint arXiv:2302.12228*, 2023. 3, 6

[13] 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. 3

[14] Yuchao Gu, Xintao Wang, Jay Zhangjie Wu, Yujun Shi, Yunpeng Chen, Zihan Fan, Wuyou Xiao, Rui Zhao, Shuning Chang, Weijia Wu, et al. Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[15] Inhwa Han, Serin Yang, Taesung Kwon, and Jong Chul Ye. Highly personalized text embedding for image manipulation by stable diffusion. *arXiv preprint arXiv:2303.08767*, 2023. 3

[16] Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. Svdiff: Compact parameter space for diffusion fine-tuning. *Proceedings of the International Conference on Computer Vision*, 2023. 3

[17] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 2

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[19] 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. 6

[20] Geoffrey E Hinton. To recognize shapes, first learn to generate images. *Progress in brain research*, 165:535–547, 2007. 1, 2

[21] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 1

[22] 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. 6

[23] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 6

[24] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 1, 2, 3, 4, 6

[25] 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. 1, 2, 3, 4, 6, 8

[26] Senmao Li, Joost van de Weijer, Taihang Hu, Fahad Shahbaz Khan, Qibin Hou, Yaxing Wang, and Jian Yang. Stylediffusion: Prompt-embedding inversion for text-based editing, 2023. 3

[27] Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. *arXiv preprint arXiv:2312.04461*, 2023. 3

[28] Zhiheng Liu, Ruili Feng, Kai Zhu, Yifei Zhang, Kecheng Zheng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. Cones: Concept neurons in diffusion models for customized generation. *International Conference on Machine Learning*, 2023. 3, 6

[29] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the International Conference on Computer Vision*, December 2015. 2

[30] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[31] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 6

[32] Andrew Ng and Michael Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Advances in Neural Information Processing Systems*, 14, 2001. 2

[33] 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. 6

[34] 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. 6

[35] 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2, 3, 6, 8

[36] Marc'Aurelio Ranzato, Joshua Susskind, Volodymyr Mnih, and Geoffrey Hinton. On deep generative models with applications to recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2857–2864. IEEE, 2011. 2

[37] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 06 2022. 2, 4, 6

[38] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 4

[39] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 1, 3, 4

[40] 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. 6

[41] Kai Wang, Fei Yang, Shiqi Yang, Muhammad Atif Butt, and Joost van de Weijer. Dynamic prompt learning: Addressing cross-attention leakage for text-based image editing. *Advances in Neural Information Processing Systems*, 2023. 3

[42] Hantao Yao, Rui Zhang, and Changsheng Xu. Visual-language prompt tuning with knowledge-guided context optimization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6757–6767, 2023. 3

[43] 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, 5, 6

[44] Beier Zhu, Yulei Niu, Yucheng Han, Yue Wu, and Hanwang Zhang. Prompt-aligned gradient for prompt tuning. In *Proceedings of the International Conference on Computer Vision*, pages 15659–15669, 2023. 1, 3, 6