MBTI: Metric-Based Textual Inversion for Fine-Grained Image Generation

Supplementary Material

A. Additional Experiments

First, we visualize the image representation space to understand how different models capture and structure visual information. Next, we analyze model responses to various prompts to assess their ability to generate relevant outputs. These analyses collectively offer valuable insights into the capabilities of modern generative models.

A.1. Image Representation Space Visualization

We aim to assess how well the results generated by each model maintain the key characteristics of the original image. To achieve this goal, we conducted a PCA analysis on the generated images to accurately assess the extent to which different methods retain these crucial features. This analysis clearly visualizes each approach's effectiveness, highlighting how well the models balance the preservation of key characteristics with the necessary variation. As shown in Figure S1, Real-Guidance's (red dots) images deviate significantly from the original image, indicating high variation and excessive alterations. In contrast, the images produced by DA-Fusion (green dots) remain close to the original, suggesting minimal alteration and blur to the original. Our method (blue dots) strikes a balance, positioning the generated images between those of DA-Fusion and the Real-Guidance model. This indicates that our method preserves essential features while introducing enough variation to enhance fine-grained classification. This balance is crucial for improving model performance while maintaining the integrity of the original.

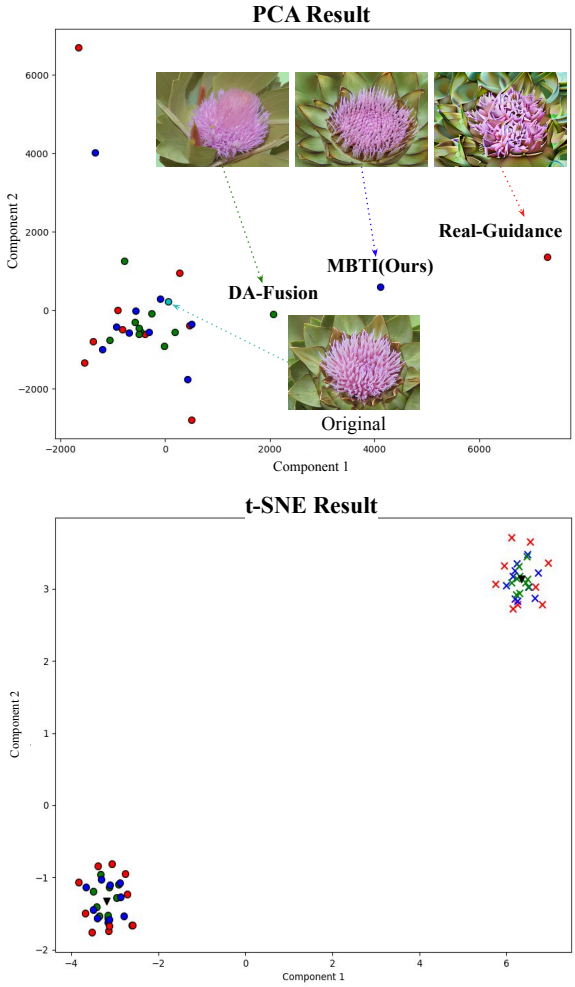

Figure S1. PCA analysis of 10 images generated by DA-Fusion (green dots), our MBTI (blue dots), and Real-Guidance (red dots). Real-Guidance produces the most unrealistic textures, placing it farthest from the original. DA-Fusion, while closest to the original, offers limited variation. Our model strikes a balance, achieving appropriate variation between the two.

A.2. Hyperparameters

Table S1 summarizes the hyperparameter settings used in our experiments, which follow those of DA-Fusion. The experiments were conducted on a single A6000 GPU with a batch size of 4, an initial learning rate of 0.005, and 1000 training steps. The classifier architecture employed was ResNet50, with a learning rate of 0.001 and a batch size of 32. The classifier was trained for 10,000 steps, with an early stopping interval set at 200 steps. The results were generated based on these configurations. For image-to-image generation, the stable-diffusion-v1-4 model was used, with a resolution of 512 and 1000 denoising steps. The Standard Prompt, "a photo of a $\langle w_i \rangle$," was used along with the input image to generate 10 synthetic images. For text-to-image generation, we use the stable-diffusion-v1-4 model with various prompts without input images.

Pipeline Figure S2 S2 illustrates our model's generation process. Initially, MBTI trains text embeddings from a small set of real images to capture fine-grained details. These embeddings are then combined with a pre-trained Stable Diffusion model to generate augmented versions for each real image across c classes, effectively enriching the dataset for improved downstream performance.

Performance Comparison of Transformer-Based Classifiers To validate the effectiveness of MBTI across different network architectures, we conduct additional exper-

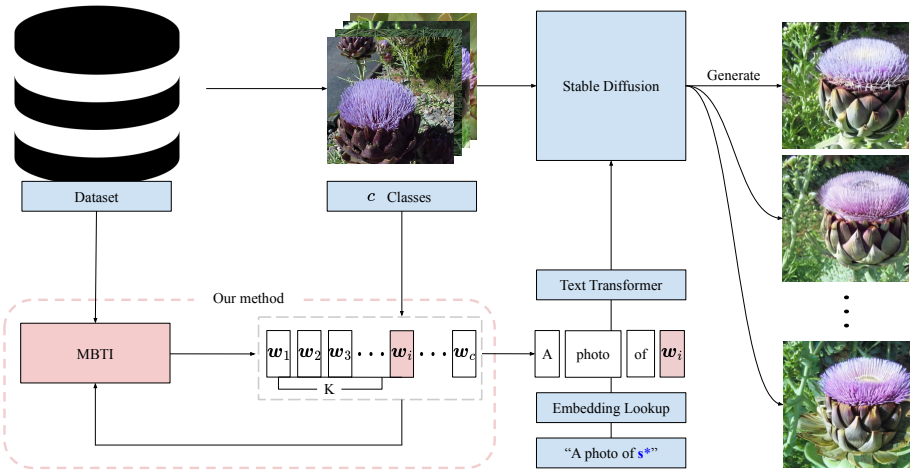

Figure S2. Overall Generation Process of the Model. The support embeddings are set to $k = 2$.

Hyperparameter	Value
Synthetic Probability α	0.5
Real Guidance Strength t_0	0.5
Num Intensities k	4
Intensities Distribution t_0	0.5
Synthetic Images Per Real M	10
Token Initialization	"the"
Batch Size	4
Learning Rate	0.005
Training Steps	1000
Class Agnostic Prompt	"a photo"
Standard Prompt	"a photo of a $\langle w_i \rangle$"
Stable Diffusion Checkpoint 1	CompVis/sd-v1-4
Stable Diffusion Guidance Scale	7.5
Stable Diffusion Resolution	512
Stable Diffusion Denoising Steps	1000
Classifier Architecture	ResNet50
Classifier Learning Rate	0.001
Classifier Batch Size	32
Classifier Training Steps	10000
Classifier Early Stopping Interval	200

Table S1. Hyperparameters and their values, All experiments were performed on a single NVIDIA A6000 GPU.

iments using more recent models. The main manuscript employs ResNet-50 for image classification. We further evaluate MBTI using the Data-Efficient Image Transformer (DeiT) on the Flowers102 dataset. Table S2 presents the results of this evaluation.

Notably, although all approaches show improved accuracy when transitioning from ResNet-50 to DeiT, MBTI consistently outperforms other methods on DeiT by achieving superior performance. These results demonstrate

Dataset	Flowers102			
Example per class	**1**	**2**	**4**	**8**
Baseline	0.473	0.638	0.762	0.852
Real-Guidance	0.492	0.643	0.765	0.851
DA-Fusion	0.463	0.640	0.758	0.846
MBTI (Ours)	**0.500**	**0.648**	**0.786**	**0.857**

Table S2. Evaluation of few-shot classification performance on the Flowers102 dataset with DeiT, a recent Transformer-based backbone.

MBTI's versatility in enhancing classification accuracy across various backbone networks.

A.3. Analysis of Generated Images using Grad-CAM.

MBTI aims to generate images that preserve and enhance features crucial for classification. We analyzed the generated images using Grad-CAM to validate this. This technique visualizes the regions of an image that most influence the model's classification decision. As shown in Figure S3, our MBTI successfully generates images that maintain the highlighted regions crucial for classification. In contrast, the Textual Inversion fails to reproduce these important areas accurately. This is why MBTI shows superior performance in few-shot classification tasks. By generating diverse images that emphasize key features. This leads to more robust model learning and better generalization, even with few original images. We conduct ablation study on the FGVC-Aircraft dataset to validate the effectiveness of the proposed components in MBTI the results of these studies).

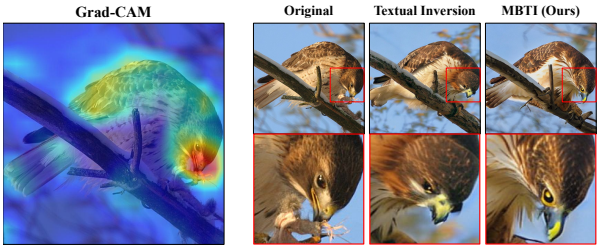

Figure S3. Grad-CAM analysis for fine-grained classification highlights class-specific features. MBTI more accurately captures the specific characteristics that Grad-CAM focuses on compared to Textual Inversion.

Dataset	**Flowers102**			
examples per class	**1**	**2**	**random(1,2,4)**	**4**
DA-Fusion	0.499	0.641	0.655	0.711
MBTI (Ours)	**0.517**	**0.658**	**0.714**	**0.773**
Difference	+0.018	+0.017	+0.059	+0.062

Table S3. Evaluation of classification with few examples under an imbalanced setting on a Flowers102 dataset. The experimental configuration follows the setup of Table 1.

Dataset	**Post2023Cars**			**BMW-New**		
Examples per class	**1**	**2**	**4**	**1**	**2**	**4**
Baseline	0.205	0.348	0.507	0.182	0.255	0.347
Real Guidance	0.251	0.347	0.515	0.181	0.243	0.304
DA-Fusion	0.265	0.333	0.511	0.225	0.265	0.385
MBTI (Ours)	**0.268**	**0.378**	**0.545**	**0.236**	**0.297**	**0.411**
Difference	+0.003	+0.031	+0.030	+0.011	+0.032	+0.026

Table S4. Few-shot classification performance on the custom car datasets (Post2023Cars and BMW-New).

A.4. Custom Datasets

We curate two car datasets, *Post2023Cars* and *BMW-New*, to minimize pretraining leakage and assess generalization to novel concepts. *Post2023Cars* contains models released after 2023, while *BMW-New* focuses on recent BMW lineups. Each dataset has 10 classes with 10 images per class. We use these sets for few-shot classification experiments; detailed results are provided below.

A.5. Robustness to data imbalance.

Table S3 evaluates robustness to label imbalance by constructing an imbalanced training set in which each class is randomly assigned a number of examples per class from 1,2,4, following the experimental setup of Table 1 and comparing against the original experiments. In the random(1,2,4) examples-per-class setting, MBTI achieves a higher macro-F1 than DA-Fusion, underscoring its effectiveness and robustness to label imbalance.

A.6. Additional Qualitative Results

Personalized concept image generation. Figure S4 compares images generated by Textual Inversion and MBTI using a customized celebrity dataset derived from Kaggle's Celebrity Face Image Dataset. The dataset includes 18 classes, with four examples per class used to train text embeddings. MBTI demonstrates superior performance in capturing fine-grained facial features essential for realism, particularly around the eyes. For instance, in images of Johnny Depp, TTextual Inversion generates pupils that are too large, while MBTI shows them in a more natural size. Similarly, for Leonardo DiCaprio, Textual Inversion fails to reproduce his distinctive eye shape, while MBTI successfully retains his recognizable facial structure. These results suggest that MBTI is more effective in preserving identity-specific details during image generation, making it better suited for high-fidelity, personalized concept synthesis.

Image-to-Image Generation Figures S5, S6, and S7 show that MBTI generates more detailed and diverse features than Textual Inversion by reducing semantic overlap with related classes in the text-embedding space.

Figure S5 evaluates robustness to style variation using the prompt "A drawing of a $\langle \boldsymbol{w}_i \rangle$." While Textual Inversion often misses fine details—such as pistils in flowers or airplane wing structure—MBTI preserves these attributes by reducing semantic overlap with related classes in the text-embedding space, resulting in more faithful renderings even under stylistic changes.

Figure S6 assesses the variation in generated features using the prompt "A photo of $\langle w_i \rangle$," which does not include the higher-level class. During this process, Textual Inversion failed to accurately depict fine details, such as the texture of flower petals, resulting in blurry or less detailed images. Conversely, the proposed model successfully rendered these fine details, producing more refined images.

In Figure S7, the prompt "A $\langle w_i \rangle$ flower" was used, which includes the higher-level class. While Textual Inversion only captured general characteristics of flowers, the proposed model generated images with more precise and detailed features.

Furthermore, base models often learn from data where people, especially women, are depicted holding or accompanied by flowers. This can unintentionally introduce such contexts into the generated images, adding elements not originally specified in the prompt. However, by generating unique embeddings, the proposed model accurately portrayed details regardless of changes in style or context, demonstrating superior performance compared to Textual Inversion.

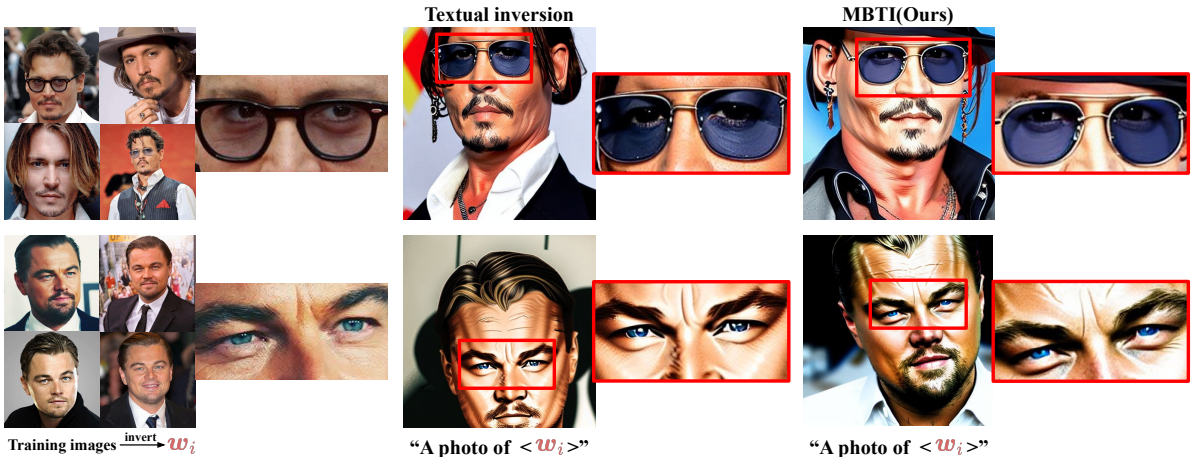

Figure S4. Visual comparisons for image generation. Generated image comparisons between Textual Inversion and MBTI using the prompt "A photo of $\langle w_i \rangle$," based on a custom dataset from Kaggle's Celebrity Face Image Dataset. In this example, support embeddings are fixed at $k = 2$.

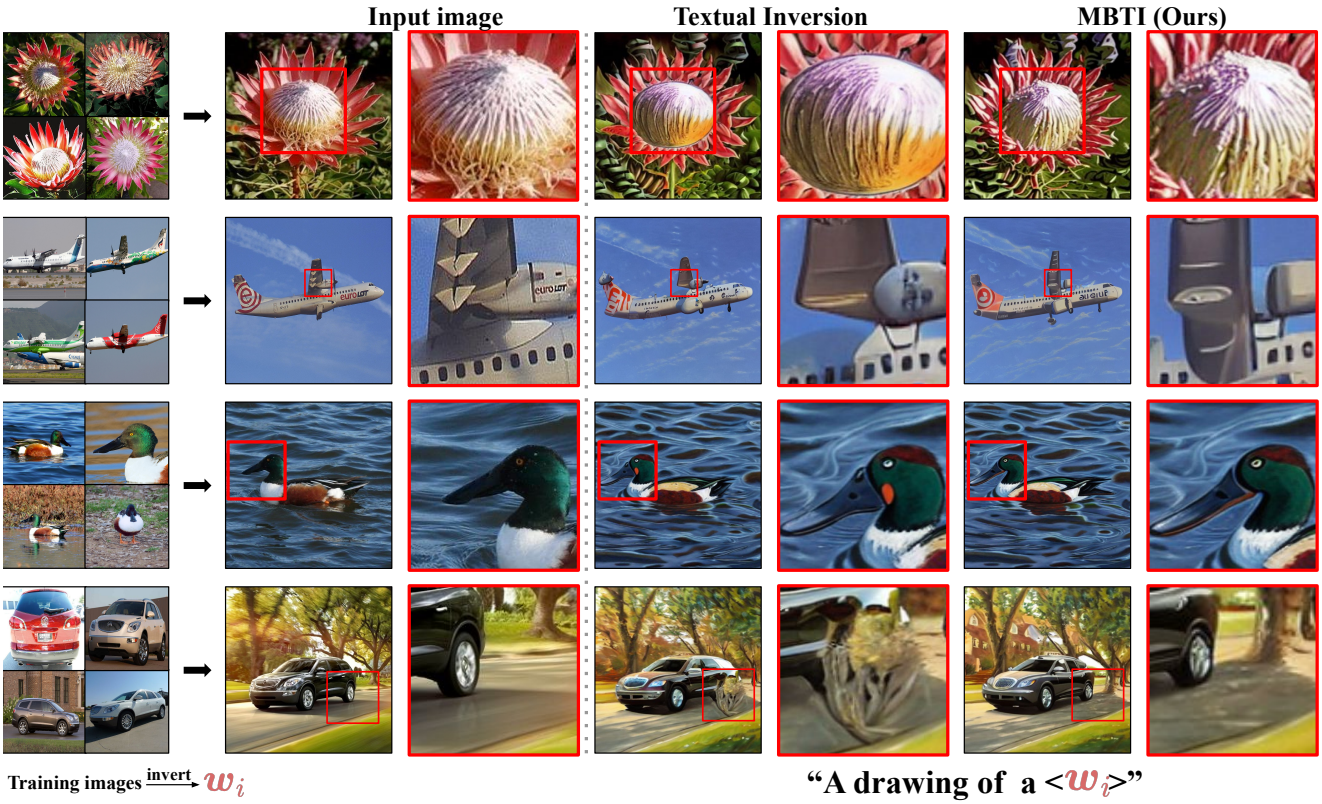

Figure S5. **Visual Comparison of Image Generation Using the Prompt "A drawing of a $\langle w_i \rangle$."** This figure demonstrates the effectiveness of the text embedding w_i in generating accurate and detailed images, even with varying styles. Textual Inversion struggles to retain fine-grained, class-specific details, such as the pistils of a flower or the wings of an airplane. In contrast, the proposed model successfully preserves these specific elements, producing more precise and detailed images despite style variations. The support embeddings are set to $k = 2$.

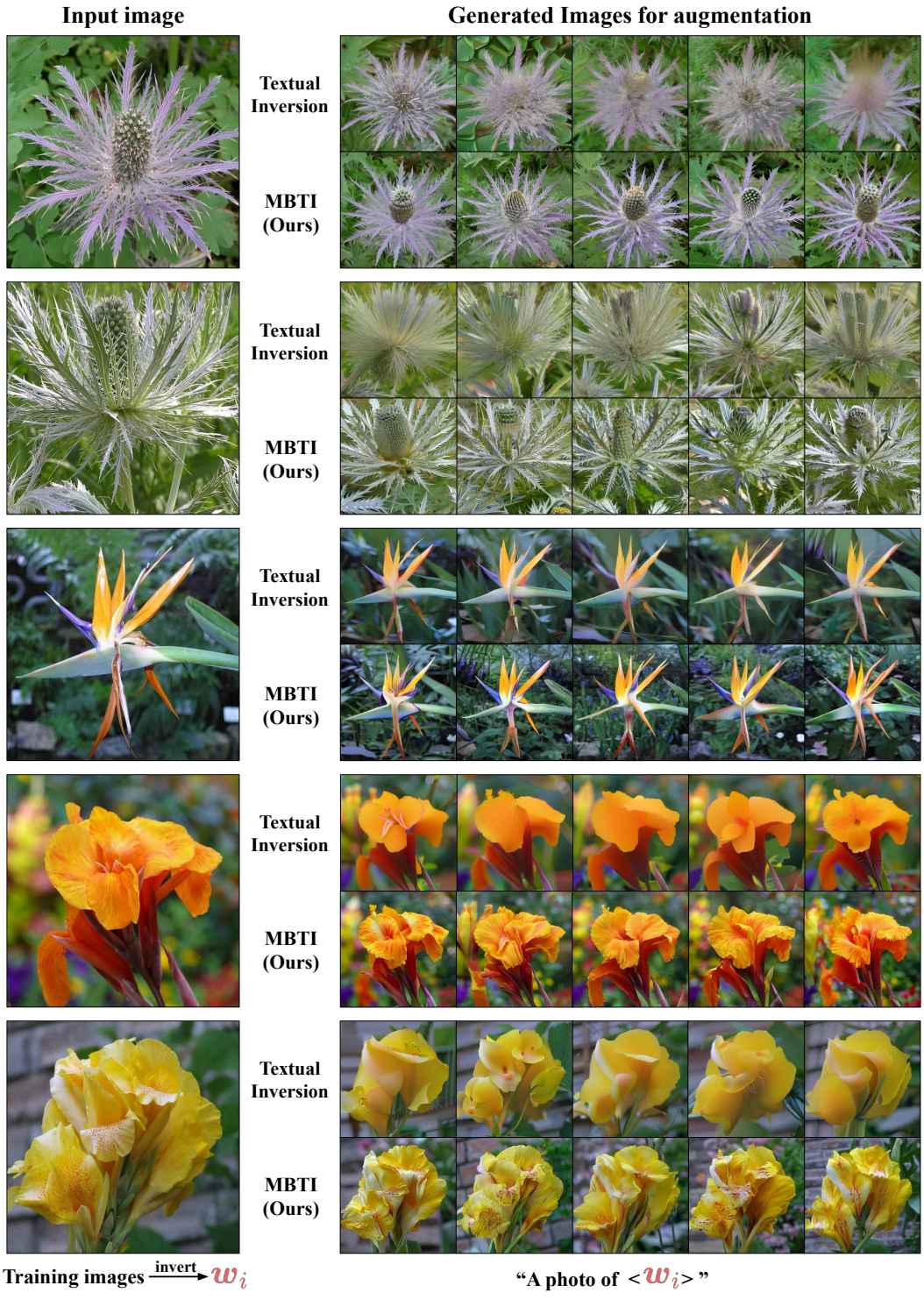

Figure S6. **Image synthesis for observing variation.** This figure demonstrates the ability to generate diverse scenarios and features. The top row displays images generated by Textual Inversion, while the bottom row shows images produced by the proposed MBTI method. The comparison highlights the higher level of variation and resulting diversity in the images generated by the MBTI approach. For this example, Support embeddings are set to $k = 2$.

Figure S7. **Image Synthesis with Upper-Class Prompt.** This figure presents images generated based on the prompt "A $\langle w_i \rangle$ flower," using the same settings as Figure S6. The images demonstrate the effectiveness of our approach in maintaining class-specific details despite the inclusion of an upper-class in the prompt.