CAT: Contrastive Adapter Training for Personalized Image Generation

*Jae Wan Park[1,2]   *Sang Hyun Park[1,2]   *Jun Young Koh[1,2]   *Junha Lee[1,2]

Min Song[1,2]

[1]Yonsei University   [2]Onoma AI

Abstract

The emergence of various adapters [1–5], including Low-Rank Adaptation (LoRA) applied from the field of natural language processing, has allowed diffusion models [6] to personalize image generation at a low cost. However, due to the various challenges including limited datasets and shortage of regularization and computation resources, adapter training often results in unsatisfactory outcomes, leading to the corruption of the backbone model's prior knowledge. One of the well known phenomena is the loss of diversity in object generation, especially within the same class which leads to generating almost identical objects with minor variations. This poses challenges in generation capabilities. To solve this issue, we present Contrastive Adapter Training (CAT), a simple yet effective strategy to enhance adapter training through the application of CAT loss. Our approach facilitates the preservation of the base model's original knowledge when the model initiates adapters. Furthermore, we introduce Knowledge Preservation Score (KPS) to evaluate CAT's ability to keep the former information. We qualitatively and quantitatively compare CAT's improvement. Finally, we mention the possibility of CAT in the aspects of multi-concept adapter and optimization.

1. Introduction

The advent of diffusion models like Stable Diffusion [7–9] has advanced the text-to-image generation field, meeting the growing demand for personalized image generation. This demand is driven by the integration of text-to-image technology in various production contexts, including comics, illustrations, and animations. Additionally, accurate personalization has spurred the application of diffusion models in other research areas, such as medical image data synthesis and augmentation [10, 11], and 3D model generation [12–14]. Despite these progresses, the complexity of achieving successful personalization poses considerable
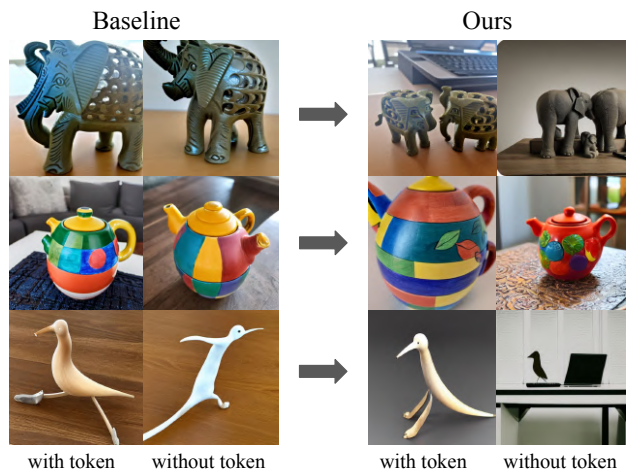
---
*These authors contributed equally to this work

Figure 1. **Comparison between Baseline [18] and Ours.** We used the following prompts *elephants, a colorful teapot,* and *a thin bird.* Baseline (left) displays knowledge shift and lacking diversity. CAT (right) does not display any knowledge shift and preserves models' ability in diverse manner.

challenges [15–17], primarily due to stringent data requirements and inherently unstable nature of adapters.

To lower the burden of dataset requirements and avoid fickle results, various researches utilizing limited datasets have been conducted [18–21]. Still, the following problems remain unsolved: underfitting and catastrophic forgetting due to overfitting. These problems cause the degradation of generation quality along with unsuccessful identity generation.

We propose a new training pipeline called **C**ontrastive **A**dapter **T**raining that efficiently tackles the aforementioned problems. We adopt a novel form of optimization function that does not require any data augmentation compared to former methods [19, 20]. CAT allows the model to focus contrastively on maintaining the original model's base knowledge by calculating the difference of noise prediction between the original model and adapter without any token conditioning. Lastly, we apply minor modifications to the former metrics and introduce a new metric to measure quantitatively the magnitude of identity generation with knowledge preservation.
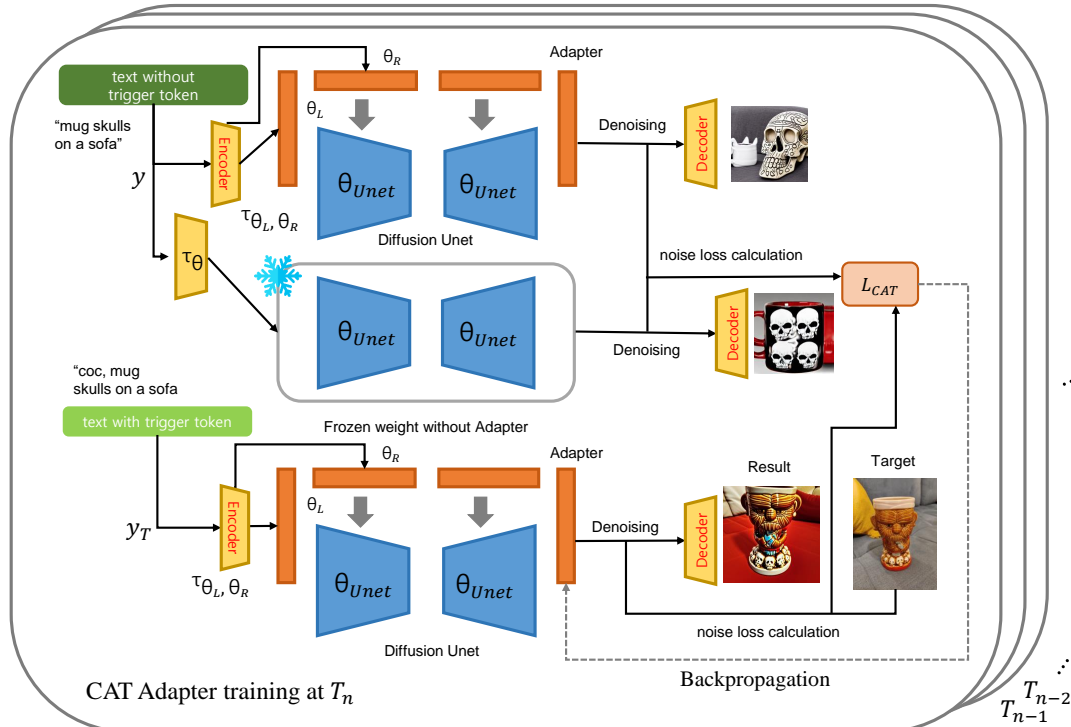
Figure 2. **The Basic Pipeline of CAT.** The CAT loss between the frozen U-net [22] and unconditioned adapter activation is calculated while training the adapter. The Adapters such as LoRA in the attention layers of U-net is depicted as orange boxes. All the figures share the same parameters unless specified.

The contributions concerning the proposed approaches are as follows.

- We develop a new pipeline that mitigates the underfitting and knowledge corruption problem in consistent generation adaptations.
- We propose a simplified new metric that assesses the knowledge preservation degree of text-to-image generation adaptations.
- With the proposed pipeline and metric, we qualitatively and quantitatively show the effectiveness of our proposed methodology.

We also anticipate varied improvements regarding the application and structure of CAT. We mention this part in Section 5.

2. Method

2.1. Preliminary

Latent Diffusion Models [6] are probabilistic architectures that model intractable density functions by sending images to latent spaces. For a certain number, n, of steps, T, the model passes the latent forward through adding or extracting some noise from Gaussian normal distribution. We also describe this process in the introduction of our method (Fig 2). During the train, the model gains the ability to predict how much noise to extract which is the equivalent of

modeling the density function. Researches improved the architecture by applying deep learning structures including the implementation of attention architecture [23]. This not only has increased the capacity of the model [8], but also enabled text-to-image guidance utilizing CLIP embedding [24]. The mathematical model often denotes the text prompts as y, the text encoder as τ_θ, and the original LDM loss function as

$$L_{LDM} := \mathbb{E}_{\epsilon_{(x)}, y, \epsilon \sim \mathcal{N}(0,1), t}[||\epsilon - \epsilon_\theta(z_t, t, \tau_\theta(y))||_2^2] \quad (1)$$

where ϵ is the additive Gaussian noise.

Diffusion Adapters have rapidly developed in text-to-image generation. Diffusion models have been a significant breakthrough in high quality image generation, yet the realm of generation control is still a field of active research. Many refers to one of the major approaches to enable the control as *personalization*. It is the capability of generating a consistent identity with variety, and to achieve it, adapters are deemed as reasonable solutions.

One of the well received personalization method is Dreambooth [19]. The method gains control over consistency using rare-token identifier. It creates a dataset of consistent identity and finetunes the backbone model, mostly diffusion models. The identity images are paired with a rare token: a combination of special, meaningless letters. From this, we denote the initialization of the adapter with

Figure 3. **Results of text-to-image generation with prompts**: *a round bird staring to the right with its beak closed*. LoRA fails to preserve original knowledge and limits its generation to target image. CAT achieves high fidelity identity generation, while retaining the original knowledge in token-less generation.

y_T which represents prompts with rare tokens added in text guidance. To prevent the mode collapse phenomenon [25], or just simply to keep the general capability of the model, it utilizes the self-generated image set of the identity class, called *regularization* set, and constructs a novel preservation loss function.

The idea of linear layer Low-Rank Adaptation [18] has been successfully adapted to text-to-image model's attention layers. It reduced the cost of finetuning in computational aspect, allowing customer-level devices to deal with large language models such as LLAMA and GPT series. [26–28]. Especially in text-to-image models, LoRA has allowed personalization by decomposing mainly U-net [22] attention layers during the identity injection steps. LoRA has progressively overwhelmed Dreambooth by reducing the cost of train which solved the underfitting problem, common with Dreambooth training.

Basic decomposition of attention layers of diffusion model is defined such like

$$\theta_{n \times n} = \theta_{n \times r} \times \theta_{r \times n} \qquad (2)$$

However, LoRA does not explicitly state or restrict any form of regularization, and small sizes of datasets invoke catastrophic forgetting and cause model's knowledge shift.

Textual Inversion [21] took a whole new approach in personalization. The method does not attempt to update the weight of the original model. Rather, it tries to find the embedding in the text embedding space, such as BERT's hot vector space [29]. This train targets to generate the identity when the embedding is added to the prompt.

The Chosen One [20] utilizes both LoRA and Textual Inversion with K-MEANS++ [30] to enhance the targeted

identity. It generates a certain number of images like Dreambooth and adopts the machine learning method to cluster identities from the images and augment the intended personalization.

Textual Inversion and The Chosen One introduce unique ways for personalization, but the first still retains the issue of underfitting when the latter tends to overfit the other clustered identity to the intended one when the dataset is limited. Given these complications, we propose a method that sustains the general capacity with better initialization by implementing Contrastive Adapter Training loss.

2.2. Contrastive Adapter Training

Following the tradition of Parameter-Efficient Fine-Tuning (PEFT) [31], all weights are frozen except the target adapter weights. Here, we use LoRA from PEFT as minimal baseline. We adopt Mean Squared Error loss for U-Net's noise prediction between original model and adapted model and accumulate the loss to the original training loss (Fig 2). With this approach, CAT minimizes the corruption of the original knowledge and unwanted impact of the adaptation.

Using the rank decomposition notation (2) as a part of an adapter initialization, we denote CAT loss function as

$$L_{CAT} :=$$
$$\mathbb{E}_{\epsilon_{(x)},\, y,\, y_T,\, \epsilon \sim \mathcal{N}(0,1),\, t} [||\epsilon - \epsilon_\theta(z_t, t, \tau_{\theta_L, \theta_R}(y_T))||_2^2 \quad (3)$$
$$+\ \alpha * ||\epsilon_\theta(z_t, t, \tau_\theta(y)) - \epsilon_\theta(z_t, t, \tau_{\theta_L, \theta_R}(y))||_2^2]$$

while the constant α determines the strength of CAT.

3. Evaluation

We have conducted the evaluation regarding CAT mainly by comparing three metrics that uphold our assertion: identity similarity, prompt similarity, and knowledge preservation score. Various studies [19,21,32] already have accepted the first two, including TCO [20].

We have combined and modified the former metrics to make them more suitable to measure the quality of the proposed method. We named this new metric, **K**nowledge **P**reservation **S**core (KPS). Also, we added some adjustments in them when we aggregate the result to avoid outlier impacts. For notations, see A.1.

Prompt Similarity evaluates the objective quality of the image. The prompt and generated image without a token are embedded with CLIP encoder, θ_{clip}, and similarity is calculated without normalization. Let the index of the image and caption pair be denoted as i, and let P denote the caption part of the pair and I^{y_T} denote the image generated by adapter applied. Then, each pair's score stands for

$$s_i^{prompt} = S_C(\theta_{clip}(P_i), \theta_{clip}(I_i^{y_T})) \qquad (4)$$

With this, use harmonic mean to average the score of the whole inference.

$$prompt\ score = H(s_1^{prompt}, s_2^{prompt}, \ldots, s_i^{prompt}) \quad (5)$$

Identity Similarity compares the token generated image from the adapters with the original identity image. The original image is captioned along with a trigger token to note the identity, and the generated image is guided by this equivalent caption. Two images then are embedded by the CLIP encoder, and cosine similarity is calculated with normalization. Let the score of each pair of images denoted as

$$s_i^{id} = S_C(\theta_{clip}(I_i^{original}), \theta_{clip}(I_i^{y_T})) \qquad (6)$$

Again, for the final score, apply harmonic mean like above.

$$identity\ score = H(s_1^{id}, s_2^{id}, \ldots, s_i^{id}) \qquad (7)$$

Finally, to evaluate the degree of knowledge preservation, we implement *Knowledge Preservation Score* that compares the images generated by y and y_T. Denoted the following,

$$s_i^{prev} = S_C(\theta_{clip}(I_i^y), \theta_{clip}(I_i^{y_T})) \qquad (8)$$

And since s_i^{prev} is normalized within [0,1] and due to the usage of harmonic mean, we can subtract the total aggregation from 1. Adapters that lost the prior knowledge generate similar images whether the prompt has a trigger token or not as shown in Figure 3. Thus, the score will show that the increase in similarity points out overfitting and knowledge loss from the adapters.

$$knowledge\ preservation\ score =$$
$$1 - H(s_1^{prev}, s_2^{prev}, \ldots, s_i^{prev}) \quad (9)$$

We applied this metric to adapters trained in as unbiased environment as possible. We listed the results in Table 1. For details about the trains and settings, see A.2.

Method	PS(5)	IS(7)	KPS(9)
LoRA [18]	0.2894	**0.9056**	0.0808
Dreambooth [19]	0.2964	0.8394	0.0883
Textual Inversion [21]	**0.3054**	0.8405	0.0973
TCO [20]	0.2406	0.8416	0.0907
CAT(**Ours**)	0.2938	0.8968	**0.1231**

Table 1. **Comparison between common adapters**. The prompt score and identity score do not show any drastic gap, however, CAT overwhelms other adaptations in knowledge preservation score.

4. Conclusion

We accomplish the following two primary objectives. Firstly, we propose Contrastive Adapter Training (Fig 2 and Eq 3), an on-the-fly regularization technique that effectively transforms knowledge shift into knowledge injection. Secondly, we introduce the Knowledge Preservation Score (KPS) as a means to assess the controllability of token generation by utilizing the harmonic mean of similarity scores from trained concepts. Our experiments demonstrate that CAT surpasses competing methods in preserving knowledge, thereby enabling precise control over concept generation. This model notably prioritizes token-specific conditions, avoiding the common pitfall of class-wise mode collapse.

5. Limitation and Future Works

We note that we have not included CLIP score-based diversity and fidelity calculation due to its instability [33–35]. Additionally, we have yet to examine the impact of discrepancies between the model's domain knowledge and the training domain with no thorough investigation on CAT structure and application. In future work, we aim to establish a reliable benchmark for consistent character generation based on our current Knowledge Preservation Score (9). Also, we intend to inspect the impact of CAT (Fig 7) and optimizing its structure (Fig 5) more carefully. Both tasks will require various examinations such as Figure 8. Mainly, we plan to enhance CAT to support multi-concept training, incorporating per-token loss for individual items demonstrated in Section A.3. This expansion holds promise for achieving significant advancements in multi-concept generation [36–38] and enhancing the efficiency of personalization strategies.

References

[1] Haokun Liu et al. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. 2022. 1

[2] Neil Houlsby et al. Parameter-efficient transfer learning for nlp. 2019. 1

[3] Mou et al. T2i- adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. 2023. 1

[4] Zhang et al. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. 2023. 1

[5] Mahabadi et al. Parameter-efficient multi-task fine-tuning for transformers via shared hypernetworks. 2021. 1

[6] Robin Rombach et al. High-resolution image synthesis with latent diffusion models. 2022. 1, 2

[7] J Sohl-Dickstein. Deep unsupervised learning using nonequilibrium thermodynamics. *ICML 2015*, 18(1):234–778, 2015. 1

[8] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *arxiv:2006.11239*, 2020. 1, 2

[9] Alex Nichol Prafulla Dhariwal. Diffusion models beat gans on image synthesis. *arXiv:2105.05233*, 2021. 1

[10] CSaloni Dash et al. Medical time-series data generation using generative adversarial networks. 2020. 1

[11] M. AbdulRazek and M. Belal G. Khoriba. Gan-ga: A generative model based on genetic algorithm for medical image generation. 2023. 1

[12] Yichun Shi et al. Mvdream: Multi-view diffusion for 3d generation. 2024. 1

[13] Hansheng Chen et al. Single-stage diffusion nerf: A unified approach to 3d generation and reconstruction. 2023. 1

[14] Jiatao Gu et al. Nerfdiff: Single-image view synthesis with nerf-guided distillation from 3d-aware diffusion. 2023. 1

[15] Hu Ye et al. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *CVPR*, 2023. 1

[16] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22560–22570, October 2023. 1

[17] Yoad Tewel et al. Training-free consistent text-to-image generation. *https://arxiv.org/abs/2402.03286*, 2023. 1

[18] Edward J Hu et al. Lora: Low-rank adaptation of large language models. 2022. 1, 3, 4, 7, 8, 10

[19] Nataniel Ruiz Hu et al. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. 2023. 1, 2, 3, 4, 7, 8

[20] Avrahami et al. The chosen one: Consistent characters in text-to-image diffusion models. 2023. 1, 3, 4, 7, 8

[21] Rinon Gal et al. An image is worth one word: Personalizing text-to-image generation using textual inversion. 2022. 1, 3, 4, 7

[22] Olaf Ronneberger et al. U-net: Convolutional networks for biomedical image segmentation. 2015. 2, 3

[23] Ashish Vaswani et al. Attention is all you need. 2017. 2

[24] Alec Radford et al. Learning transferable visual models from natural language supervision. 2021. 2

[25] Z Zhao et al. Modified generative adversarial networks for image classification. 2021. 3

[26] Hugo Touvron et al. Llama: Open and efficient foundation language models. 2022. 3

[27] Tom B. Brown et al. Language models are few-shot learners. 2020. 3

[28] Alec Radford et al. Language models are unsupervised multitask learners. 2019. 3

[29] Jacob Devlin et al. Bert: Pre-training of deep bidirectional transformers for language understanding. 2019. 3

[30] David Arthur and Sergei Vassilvitskii. k-means++: the advantages of careful seeding. 2019. 3

[31] Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. Peft: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`, 2022. 3

[32] Omri Avrahami et al. Break-a-scene: Extracting multiple concepts from a single image. *SIGGRAPH Asia 2023*, 2023. 4

[33] Shin-Ying Yeh et al. Navigating text-to-image customization:from lycoris fine-tuning to model evaluation. 2024. 4

[34] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Winoground: Probing vision and language models for visio-linguistic compositionality. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 4

[35] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *International Conference on Learning Representations*, 2023. 4

[36] Kihyuk Sohn et al. Styledrop: Text-to-image generation in any style. *https://arxiv.org/abs/2306.00983*, 2023. 4

[37] Zhe Kong et al. Omg: Occlusion-friendly personalized multi-concept generation in diffusion models. *https://arxiv.org/abs/2403.10983*, 2024. 4

[38] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *CVPR*, 2023. 4

[39] Bastian Leibe and Bernt Schiele. Analyzing appearance and contour based methods for object categorization. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, volume 2, pages II–409. IEEE, 2003. 7

[40] Ilya Loshchilov et al. Decoupled weight decay regularization. *ICLR*, 2019. 8

[41] Dustin Podell et al. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *https://arxiv.org/abs/2307.01952*, 2023. 8

[42] Zeju Qiu, Weiyang Liu, Haiwen Feng, Yuxuan Xue, Yao Feng, Zhen Liu, Dan Zhang, Adrian Weller, and Bernhard Schölkopf. Controlling text-to-image diffusion by orthogonal finetuning. In *NeurIPS*, 2023. 10
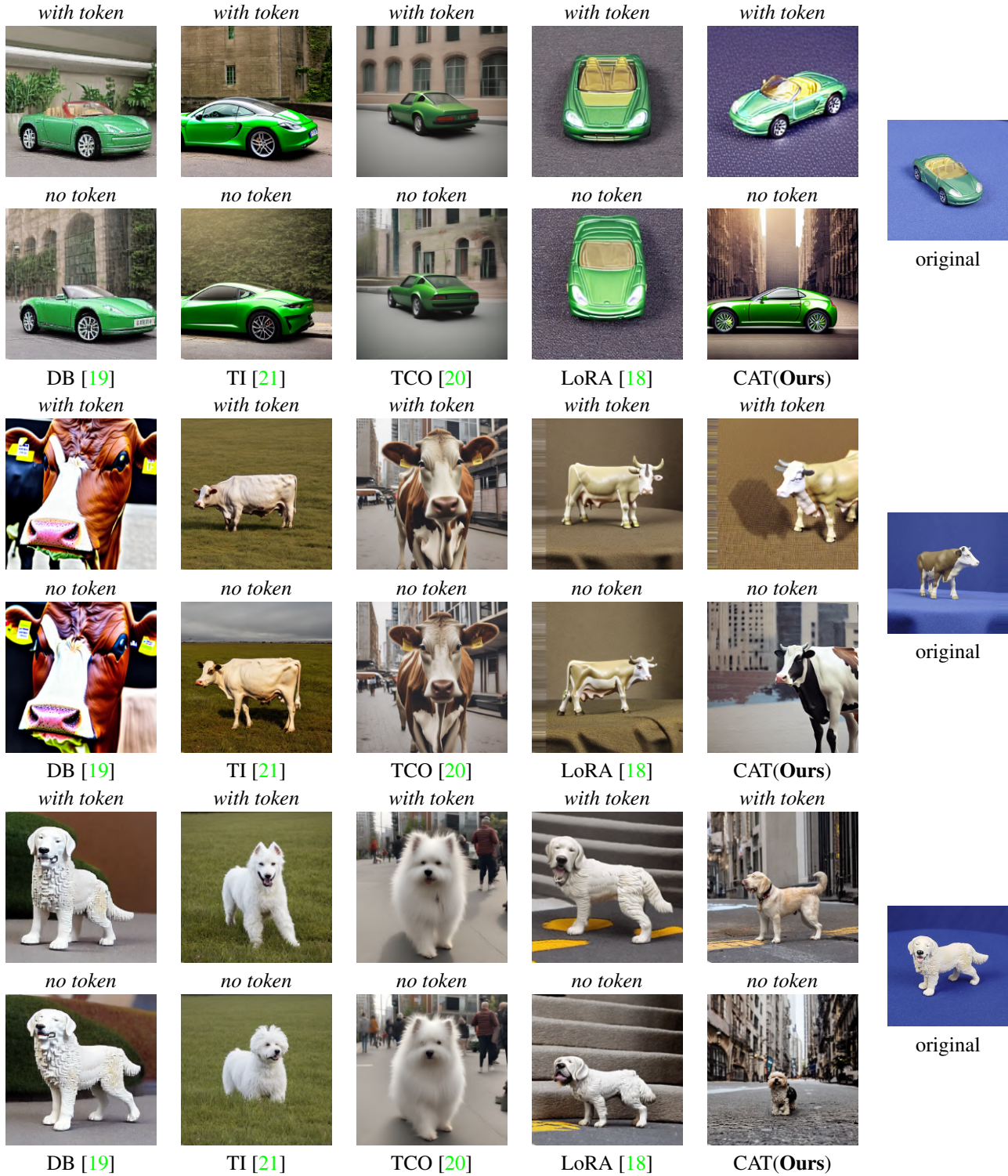
Appendix

Figure 4. **Qualitative results of various adapter training.** We used ETH-80 dataset [39] for training and validation. The prompts used for generation are *a green sports car, a cow in a city,* and *a dog in a city.*

A. Background

A.1. Mathematical Notations

We denoted cosine similarity used in equation 4, 6, and 8 as

$$S_C(A, B) = \frac{A \cdot B}{\|A\|_2 \|B\|_2} \tag{10}$$

For harmonic mean noted in 5, 7, and 9,

$$H(x_1, x_2, \ldots, x_n) = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} \tag{11}$$

A.2. Train Settings

For Figures 1, 3, 5 and Table 1, we utilized Textual Inversion dataset to validate our training. Each dataset was composed of 5-6 images. For the table, we trained total 9 adapters for each type and generated 10-11 images for each adapter to evaluate the metrics. For Figure 4, we collected 6 images from each ETH-80 dataset for the experiment. For these inferences, we trained all adapters with equal 600 steps with mentioned datasets except the regularization sets which were an essential part of some adapter's algorithm. For Figure 7, we only used a single dataset of *red teapot* among the datasets above and trained 4 adapters with different CAT variables, α, and 1000 steps. We applied uniform learning rate of 1e-4 for U-Net LoRA, Dreambooth and Textual Inversion. We applied AdamW optimizer [40] with weight decay λ=0.01 and betas β=(0.9, 0.999), epsilon ϵ=1e-08, utilizing max 18GB of VRAM in average, as described in Figure 5 with detailed training time. For hardware, we used 2 GeForce RTX 4090 gpus for all training.
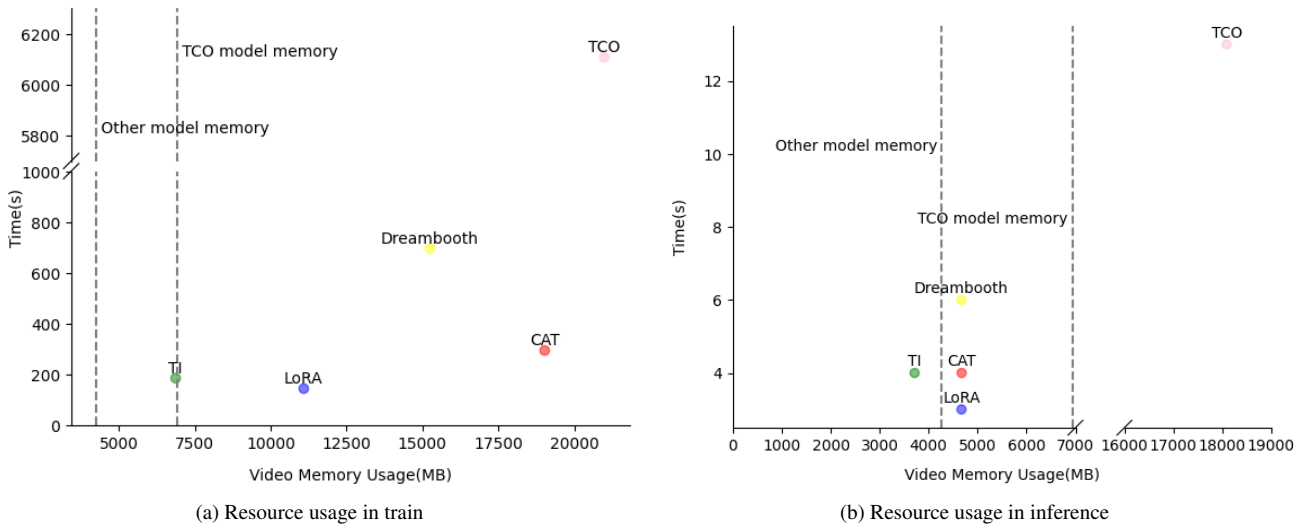

(a) Resource usage in train

(b) Resource usage in inference

Figure 5. **Video memory and time consumption of various adapters in train and inference.** The red dots show our method's performance. TCO [20] uses Stable Diffusion XL model [41] requiring a larger VRAM consumption. We included the preparation time for regularization set with Dreambooth [19] and TCO in the total training time. We plan to conduct further optimizations to decrease memory consumption on the same level as LoRA [18].

A.3. Multi-Concept Training

For Figure 6, we selected 44 images from Textual Inversion to compose a single dataset. We trained only one LoRA based CAT adapter with total 32,000 steps and the same setting from Section A.2. The experiment is to elucidate multi-concept training capability. This showed the result of a successful training trend of custom concepts in a single adapter, suggesting a large scale knowledge injection capability. For simplicity, we fixed all trigger tokens as first phrase, letting them be replaced dynamically.
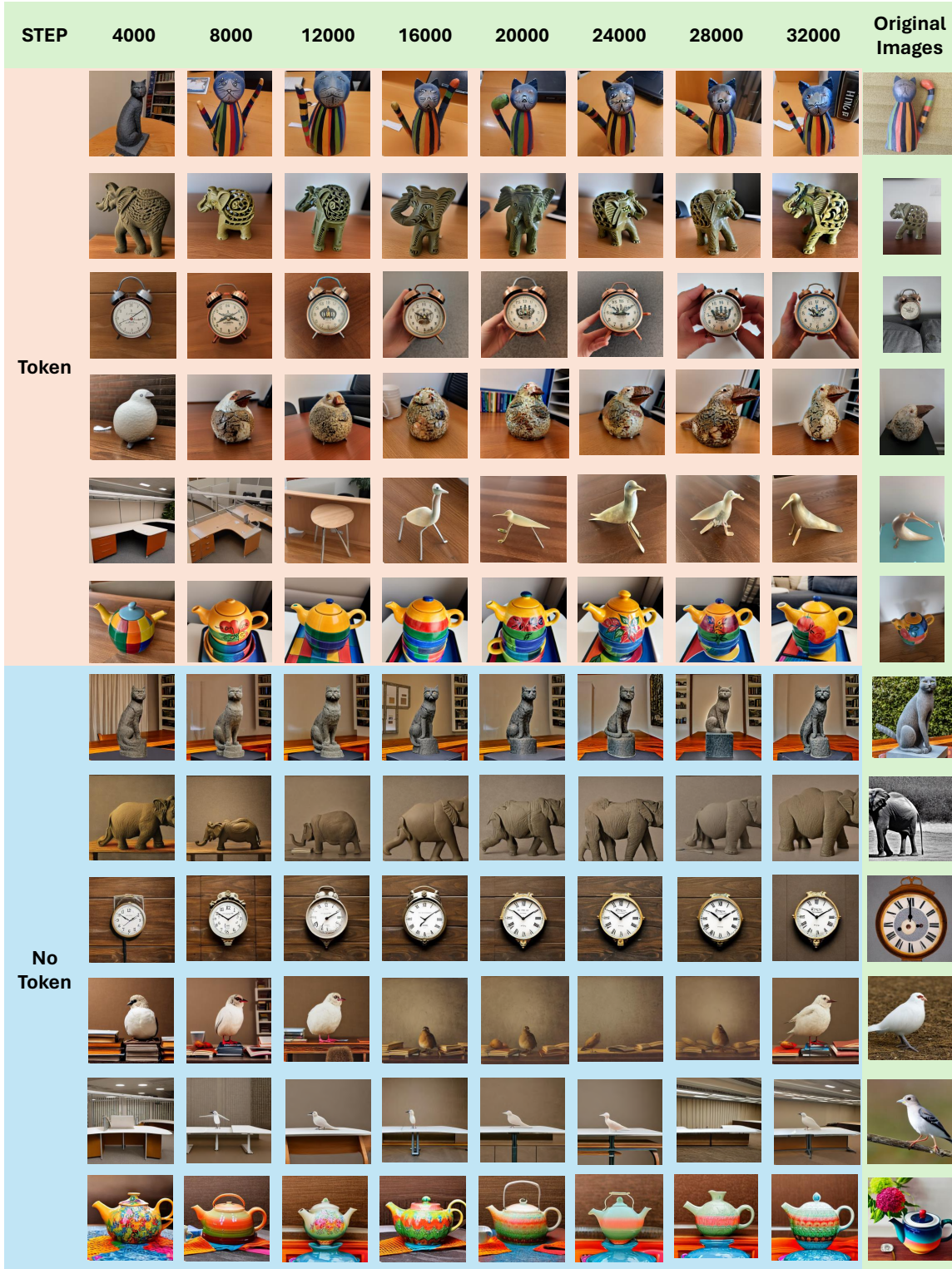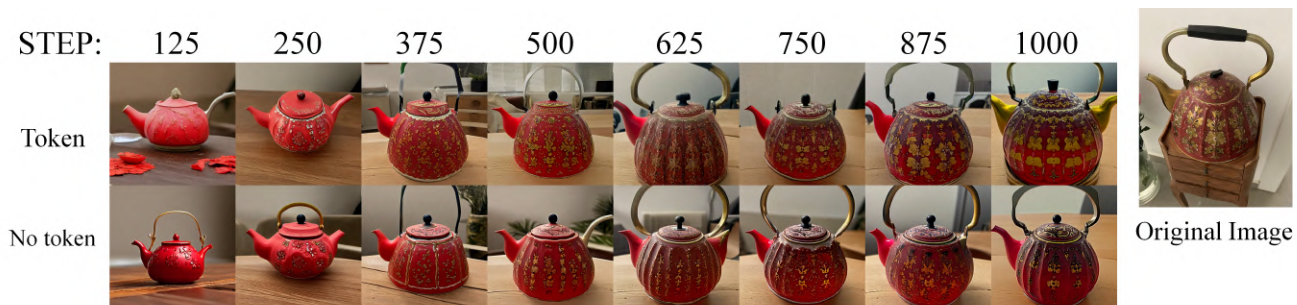
Figure 6. **Single Adapter Multi-Concept training samples.** The result shows the capability of learning multiple concepts as desired special tokens in single adapter.

Figure 7. **The qualitative result of CAT factor experiment using LoRA [18].** When the CAT factor is low, the original class generation tends to be off-grade, which is known as hyperspherical energy diverge [42]. On the other hand, with proper tuning of CAT factor, the adapter generates the original class with better quality, also the injection of the original knowledge accelerates.
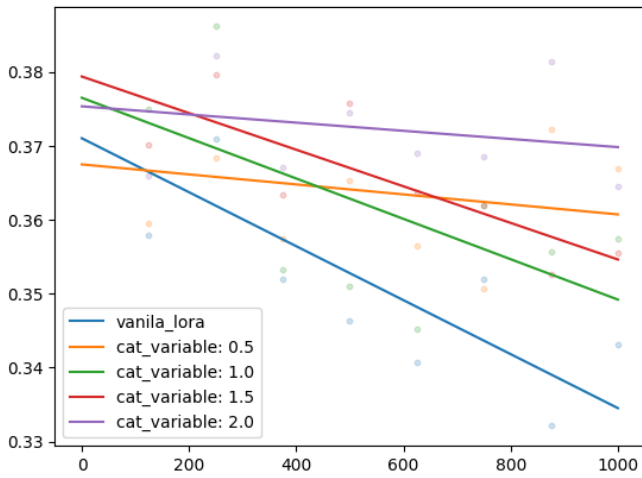
(a) Prompt Score(5) Graph regarding the experiment in Figure 7
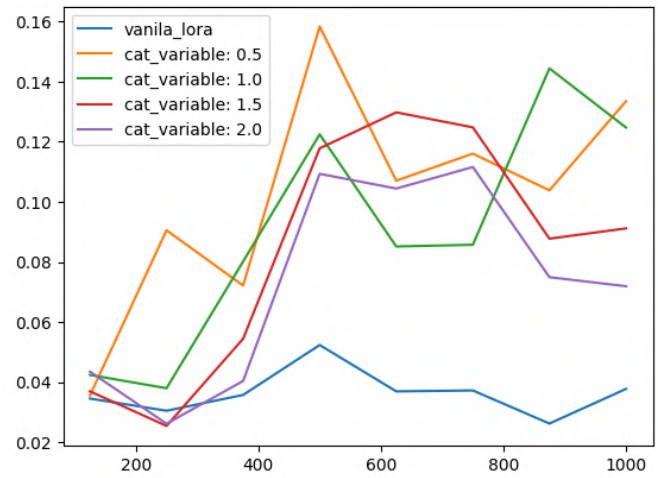
(b) KPS(9) Graph regarding the experiment in Figure 7

Figure 8. **Metric Graph on CAT variable test**. We used the metrics mentioned in Section 3 to evaluate the experiment in Figure 7. As stated in the experiment, prompt score shows the higher capability of original knowledge generation. Also, KPS is always higher than the baseline LoRA as long as we apply CAT despite the variable's magnitude.