Visual Prompt Tuning for Generative Transfer Learning

Kihyuk Sohn, Yuan Hao, José Lezama, Luisa Polania,
Huiwen Chang, Han Zhang, Irfan Essa, Lu Jiang

Google Research

CVPR2023

Presenter: Hsuan Yu Fan
Date: 2023/8/24

Outline

- Introduction
- Preliminary
- Methodology - Visual Prompt for Generative Transfer
- Experiments
- Analysis and Discussion
- Conclusion

Introduction

- The goal of image synthesis is to generate **diverse and plausible scenes** resembling the training images.
- The generalization ability is usually determined by **the amount of training images**.
- Recent efforts have shown success in transferring knowledge from pretrained Generative Adversarial Network (GAN) models [42, 53, 64, 68], **these demonstrations are limited by narrow visual domains, e.g., faces or cars** [42, 68].
- In this work, we approach the transfer learning for image synthesis using generative vision transformers, such as DALL·E [47], Taming Transformer [14], MaskGIT [6], CogView [12], NÜWA [67], or Parti [70]

[42] Utkarsh Ojha, Yijun Li, Jingwan Lu, Alexei A Efros, Yong Jae Lee, Eli Shechtman, and Richard Zhang. Few-shot Image generation via cross-domain correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10743–10752, 2021. 1, 8
[53] Mohamad Shahbazi, Zhiwu Huang, Danda Pani Paudel, Ajad Chhatkuli, and Luc Van Gool. Efficient conditional gan transfer with knowledge propagation across classes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12167–12176, 2021. 1, 2, 6, 7, 11, 15
[64] Yaxing Wang, Abel Gonzalez-Garcia, David Berga, Luis Herranz, Fahad Shahbaz Khan, and Joost van de Weijer. Minegan: effective knowledge transfer from gans to target domains with few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9332–9341, 2020. 1, 2, 6, 7, 11, 15
[68] Ceyuan Yang, Yujun Shen, Zhiyi Zhang, Yinghao Xu, Jiapeng Zhu, Zhirong Wu, and Bolei Zhou. One-shot generative domain adaptation. *arXiv preprint arXiv:2111.09876*, 2021. 1

- Our study employs the visual task adaptation benchmark (or VTAB) [71]
- We present a transfer learning framework using prompt tuning [34,36]. The technique has been used for transfer learning of discriminative models for vision tasks [1,26], this work appears to be **the first to adopt prompt tuning for transfer learning of image synthesis.**

[6] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. *arXiv preprint arXiv*:2202.04200, 2022. 1, 2, 3, 4, 6, 7, 10

[12] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems*, 34:19822–19835, 2021. 1, 2

[14] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12873–12883, 2021. 1, 2, 3, 4, 6

[47] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In Marina Meila and Tong Zhang, editors, *ICM*L, 2021. **1, 2**

[67] Chenfei Wu, Jian Liang, Lei Ji, Fan Yang, Yuejian Fang, Daxin Jiang, and Nan Duan. NÜWA: Visual synthesis pre-training for neural visual world creation. *arXiv preprint arXiv:2111.12417*, 2021. 1, 2

[70] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022. 1, 2

[71] Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, et al. A large-scale study of representation learning with the visual task adaptation benchmark. *arXiv preprint arXiv:1910.04867*, 2019. 1, 2, 5

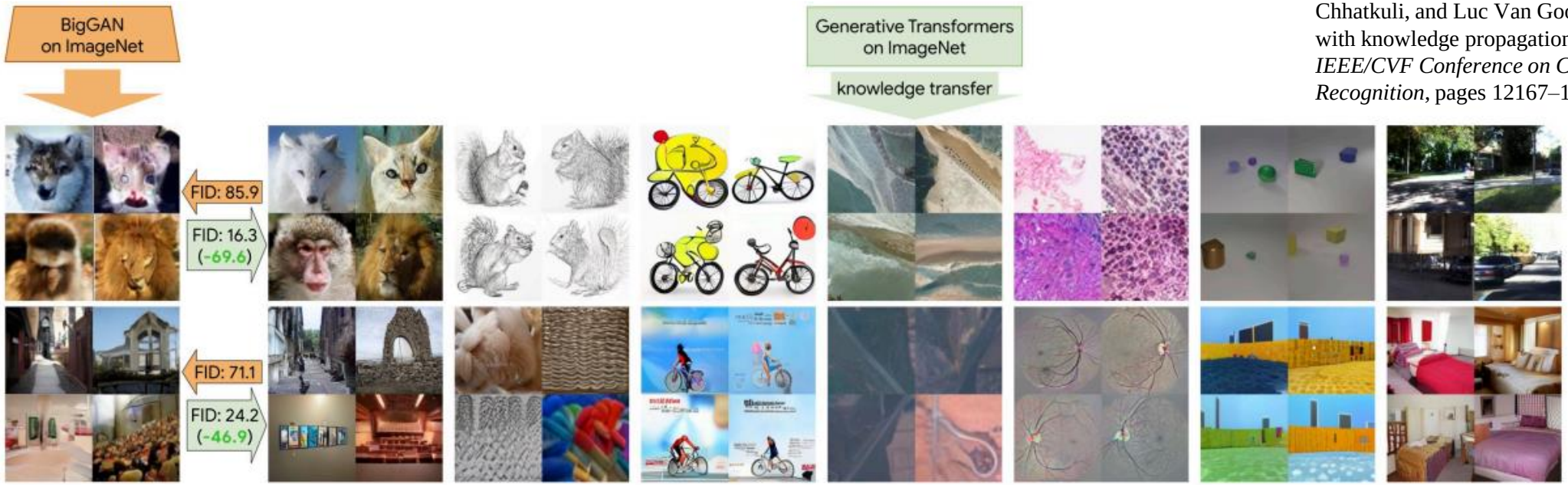
[34] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, 2021. 1, 2, 3, 4, 11

[36] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021. 1, 3, 4, 11

[1] Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. Visual prompting: Modifying pixel space to adapt pre-trained models. *arXiv preprint arXiv:2203.17274*, 2022. 1, 3, 11

[26] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. *arXiv preprint arXiv:2203.12119*, 2022. 1, 3, 4, 11

- We propose two technical innovations. First, a parameter efficient design of prompt token generator that admits condition variables (e.g., class, instance).
- Second, a marquee header prompt that engineers (e.g., composes and interpolates) learned prompts to enhance generation diversity.
- We show that generative vision transformers with prompt tuning **outperforms the prior state-of-the-art held by GANs** [53,64] through a vast margin.



(a) Generative model transfer based on GAN

(b) The proposed generative model transfer based on the generative vision transformers.

[53] Mohamad Shahbazi, Zhiwu Huang, Danda Pani Paudel, Ajad Chhatkuli, and Luc Van Gool. Efficient conditional gan transfer with knowledge propagation across classes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12167–12176, 2021. 1, 2, 6, 7, 11, 15

[64] Yaxing Wang, Abel Gonzalez-Garcia, David Berga, Luis Herranz, Fahad Shahbaz Khan, and Joost van de Weijer. Minegan: effective knowledge transfer from gans to target domains with few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9332–9341, 2020. 1, 2, 6, 7, 11, 15

Contributions

1. Present a generative visual transfer learning framework for vision transformers **with prompt tuning** [34], proposing **a novel prompt token generator design** and **a prompt engineering method** for image synthesis
2. Conduct a large-scale empirical study for generative transfer learning to validate our method on a variety of visual domains and scenarios (e.g., few-shot). To this end, we show state-of-the-art image synthesis performance.
3. The first to **employ prompt tuning for transfer learning of image synthesis**, and provide one-of-the-first substantial empirical evidence on **the necessity of knowledge transfer for data and compute efficient generative image modeling** using the standard visual transfer learning benchmark

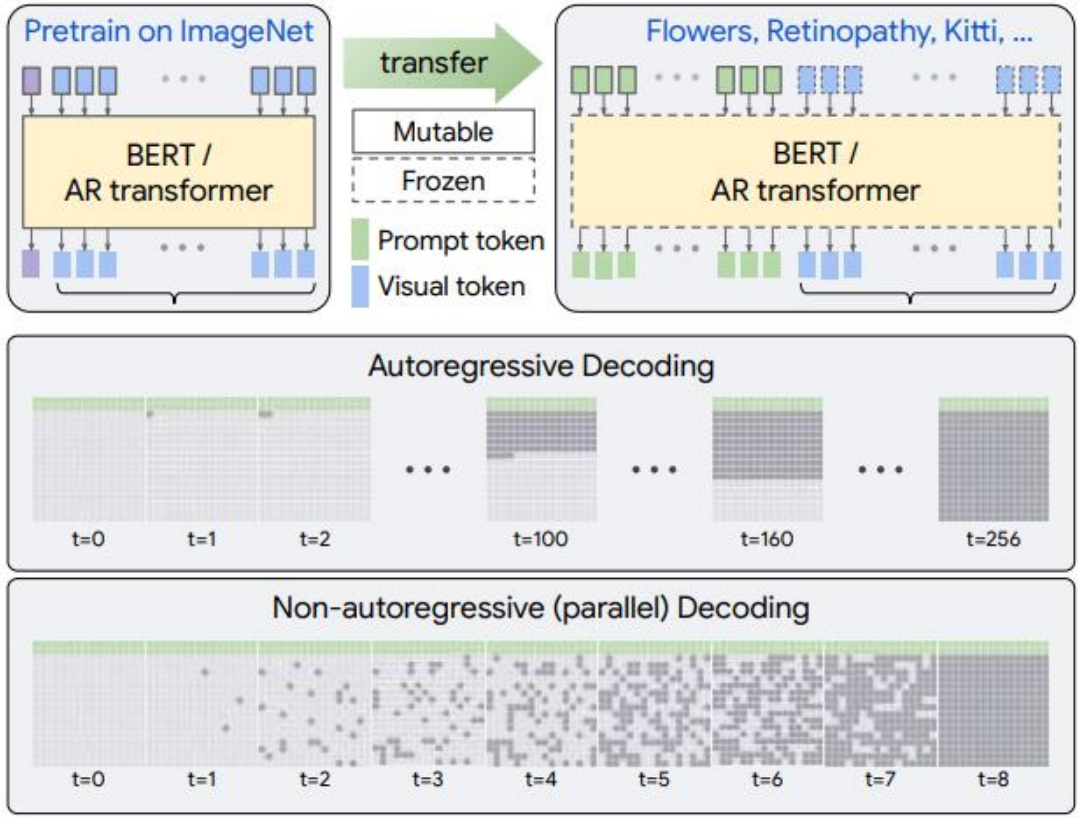
[34] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, 2021. 1, 2, 3, 4, 11

Preliminary 2.1Generative Vision Transformers

- Generally, there are two types of generative vision transformers: Auto Regressive (AR) and Non-Auto Regressive (NAR) transformers, both consisting of two stages [14,47]: image quantization and decoding.
- The first stage is the same between the two types of models in which the image is quantized into a grid of discrete tokens by a Vector-Quantized (VQ) autoencoder [14, 48, 60, 69]
- In the second stage of decoding, AR transformers follows a raster scan ordering, generating tokens from left to right, line-by-line. Finally, the generated tokens are mapped to the pixel space using the VQ decoder

[14] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12873–12883, 2021. 1, 2, 3, 4, 6
[47] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In Marina Meila and Tong Zhang, editors, *ICML*, 2021. **1, 2**
[48] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019. 2
[60] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *NeurIPS*, 2017. 1, 2
[69] Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. Vector-quantized image modeling with improved VQGAN. *arXiv preprint arXiv:2110.04627*, 2021. 2

- **NAR Transformer Process:** Begins with a fully masked canvas and generates images in approximately 8 steps. Each step predicts tokens simultaneously, preserving those with top prediction scores.
- We employ a prompt tuning [26,34,36] that uses a sequence of learnable tokens (e.g., green blocks with a solid line) to adapt to target distributions, while leaving transformer parameters frozen.



[26] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. arXiv preprint arXiv:2203.12119, 2022. 1, 3, 4, 11

[34] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, 2021. 1, 2, 3, 4, 11

[36] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021. 1, 3, 4, 11

2.2. Prompt Tuning

- Here, prompt is a sequence of additional tokens prepended to a token sequence
- In prompt tuning [34, 36], tokens are parameterized by learnable parameters and their parameters are updated via a gradient descent to adopt transformers to the downstream task.

[34] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, 2021. 1, 2, 3, 4, 11
[36] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190,* 2021. 1, 3, 4, 11

Methodology 3.1.1 Learning Visual Prompt

- Let $\mathcal{Z} = \{z_i\}_{i=1}^{H \times W}$ be a sequence of visual tokens (i.e., an output of VQ encoder followed by the vectorization) and $\mathcal{P}_\phi = \{p_{s;\phi}\}_{s=1}^{S}$ be a sequence of prompt tokens.
- For AR transformer, the loss is given as follows:

$$\mathcal{L}_{\mathrm{AR}} = \mathbb{E}_{x \sim P_{\mathcal{X}}} \left[-\log P_\theta(\mathcal{Z}|\mathcal{P}_\phi) \right] \quad (1)$$

$$P_\theta(\mathcal{Z}|\mathcal{P}_\phi) = \prod_{i=1}^{H \times W} P_\theta(z_i|z_{<i}, \mathcal{P}_\phi) \quad (2)$$

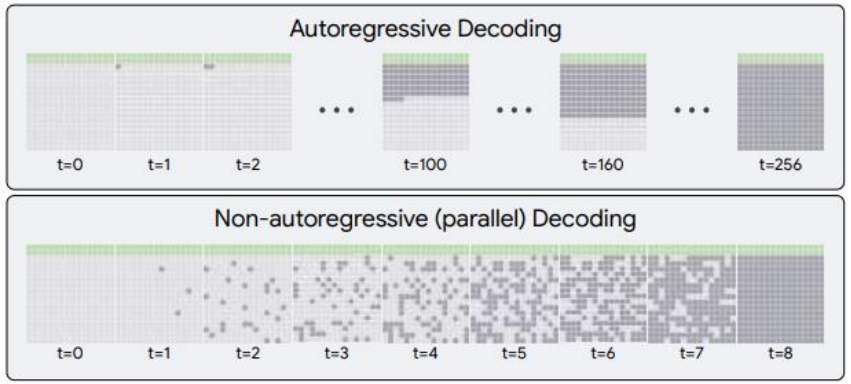
- For NAR transformer, we follow the loss of MaskGIT [6]:

$$\mathcal{L}_{\mathrm{NAR}} = \mathbb{E}_{x \sim P_{\mathcal{X}}, M \sim P_{\mathcal{M}}} \left[-\log P_\theta(\mathcal{Z}_M|\mathcal{Z}_{\overline{M}}, \mathcal{P}_\phi) \right] \quad (3)$$

$$P_\theta(\mathcal{Z}_M|\mathcal{Z}_{\overline{M}}, \mathcal{P}_\phi) = \prod_{i \in M} P_\theta(z_i|\mathcal{Z}_{\overline{M}}, \mathcal{P}_\phi) \quad (4)$$

$M \subset \{1, ..., H \times W\}$ is a set of visual token indices sampled from a masking schedule distribution $P_{\mathcal{M}}$, $\overline{M}$ is its complement, and $\mathcal{Z}_M = \{z_i\}_{i \in M}$.

- We generate visual tokens for image synthesis by iterative decoding.
- For AR transformer:



1: **for** $i \leftarrow 1$ to $H \times W$ **do**
2: $\quad \hat{z}_i \sim P_\theta(z_i | \hat{z}_{<i}, \mathcal{P}_\phi)$
3: **end for**

- For NAR model, scheduled parallel decoding [6] is used:

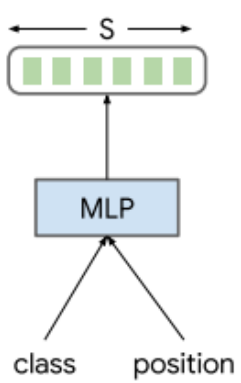
Require: $\overline{M} = \{\}, T, \{n_1, ..., n_T\}, \sum_{t=1}^{T} n_t = H \times W$
1: **for** $t \leftarrow 1$ to T **do**
2: $\quad \hat{z}_i \sim P_\theta(z_i | \widehat{\mathcal{Z}}_{\overline{M}}, \mathcal{P}_\phi), \forall i \in M$
3: $\quad \overline{M} \leftarrow \overline{M} \cup \{\arg\text{topk}_{i \in M}\left(P_\theta(z_i | \widehat{\mathcal{Z}}_{\overline{M}}, \mathcal{P}_\phi), k = n_t\right)\}$
4: **end for**

where $\{n_1, ..., n_T\}$ is a masking schedule that decides the number of tokens to decode at each decoding step.

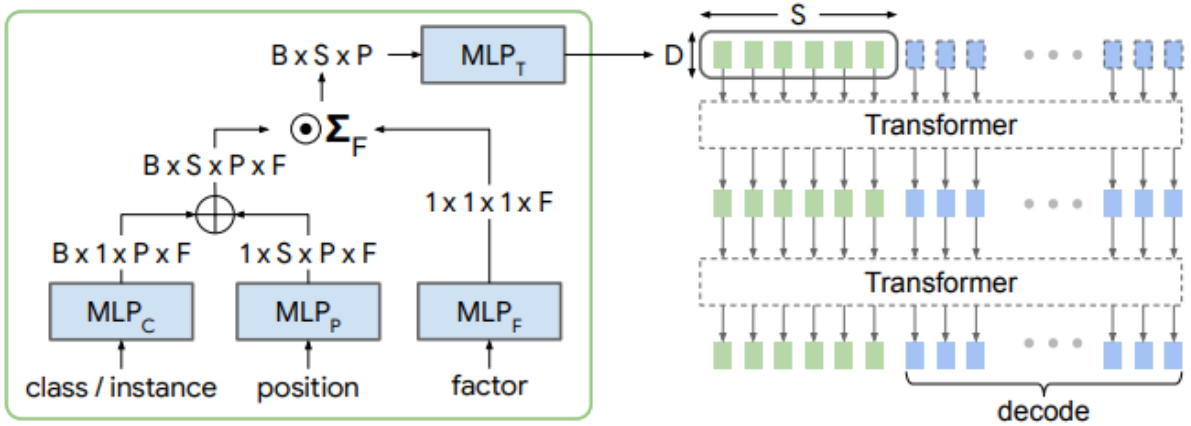
[6] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. *arXiv preprint arXiv*:2202.04200, 2022. 1, 2, 3, 4, 6, 7, 10

3.1.2 Prompt Token Generator Design

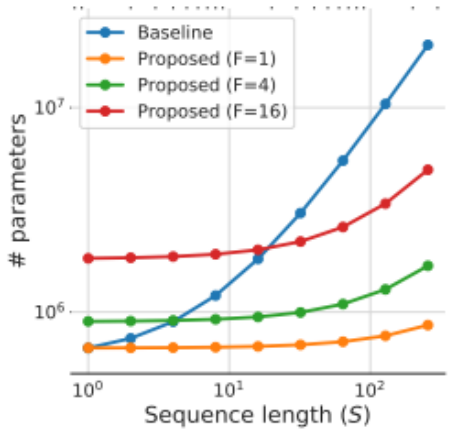
- We accomplish condition variables (e.g., class, attribute) with rather a straightforward extension of existing prompt designs using a class-condition, Pφ(c), as in Fig. a.
- To make it parameter efficient, we propose a factorized token generator, as in Fig. b
- Specifically, we encode class and sequence position index via MLP_C and MLP_P with F factors, respectively.



(a) Baseline prompt token generators of length S conditioned on class.



(b) The proposed parameter efficient prompt token generator via factorization of class / instance and position. $\oplus$ is an element-wise sum, $\odot$ is an element-wise product, Σ_F is a sum over F dimension. S: sequence length, B: batch size, P: feature dimension, D: token dimension.



(c) Number of parameters with respect to the sequence length and different number of factors F.

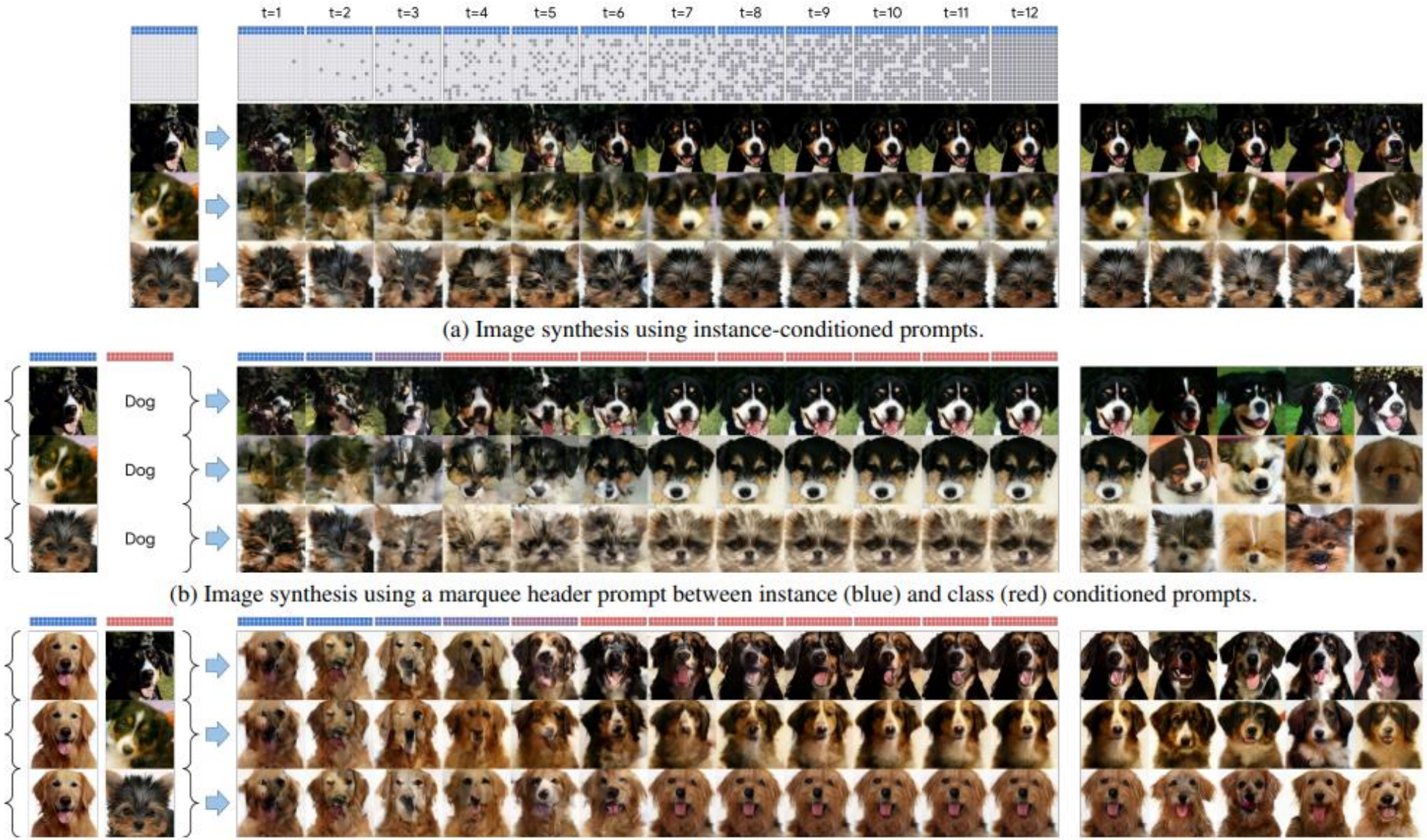
3.2. Engineering Learned Prompts

- We propose a novel prompt engineering strategy, a “Marquee Header” prompt, of the iterative transformer decoding, for enhancing the generation diversity.
- We interpolate the learned prompt representations (e.g., outputs of MLP_C).
- We provide a marquee header prompt formulation:

$$\textsc{Pmt}(t) = (1 - w_t)\textsc{Pmt}_1 + w_t\textsc{Pmt}_2 \quad (6)$$

$$w_t = \min\left\{\left(\frac{t-1}{T_{\text{cutoff}} - 1}\right)^2, 1\right\} \quad (7)$$

where t = 1, ..., T is a decoding step, $T_{\text{cutoff}} \leq T$ is a cutoff step, and PTM_i is a prompt representation (e.g., an output of MLP_C).
The schedule in Eq. (7) makes a smooth transition of prompts from PTM_1 to PTM_2.



(a) Image synthesis using instance-conditioned prompts.

(b) Image synthesis using a marquee header prompt between instance (blue) and class (red) conditioned prompts.

(c) Image synthesis using a marquee header prompt between instance-conditioned prompts (blue and red).

Figure 4. Iterative decoding of NAR transformers. (4a) instance prompts generate images of high-fidelity but with low diversity. Marquee header prompts enhance generation diversity by interpolating (4b) from instance to class prompts or (4c) between instance prompts.

Experiments 4.1. Generative Transfer on VTAB

- **VTAB Benchmark:** Utilized the VTAB [71], consisting of 19 visual recognition tasks across 16 datasets.
- **Generative Transfer Techniques:** Examined generative transfer via AR and NAR transformers. Used class-conditional Taming Transformer [14] and MaskGIT [6] trained on 256×256 ImageNet images.
- **Comparison with GAN-based Methods:** Our method vs. MineGAN [64] & cGANTransfer [53]. Both algorithms leverage BigGAN [2] from ImageNet.
- **Efficiency Analysis:** Compared with transformers trained from scratch on VTAB. Emphasized compute efficiency by maintaining a similar compute budget for both model training types.

- We use Frechet Inception Distance (FID) [24] as a quantitative metric. We generate 20k images from each model and compare with images from a respective dataset.
- Compared with using only 1 token, we find that using 128 tokens for the prompt increases the overall generation time by 25%.

Model		(# tr params)	Mean	C101	Flowers	Pet	DTD	Kitti	SUN	EuroSAT	Resisc
MineGAN [64]		(88M)	151.5	102.4	132.1	130.1	87.4	117.9	77.5	111.5	81.0
cGANTransfer [53]		(105M)	85.1	89.6	61.6	48.6	70.3	48.9	31.1	45.6	50.3
Non-Autoregressive	Prompt ($S=1$)	(0.67M)	53.7	13.5	13.8	11.9	**25.8**	32.3	**7.3**	45.9	28.5
	Prompt ($S=16$)	(0.68M)	39.9	12.7	**13.2**	11.1	26.0	30.0	7.4	**35.8**	24.9
	Prompt ($S=128$)	(0.76M)	**36.4**	**12.9**	13.4	**10.9**	**25.9**	**29.9**	7.7	38.4	**24.8**
	Scratch	(172M)	42.7	72.7	57.2	70.3	66.1	33.8	9.2	39.5	32.0
Autoregressive	Prompt ($S=1$)	(0.86M)	58.4	45.5	28.9	42.4	37.1	66.9	18.9	37.3	35.1
	Prompt ($S=16$)	(0.88M)	45.8	41.4	19.6	36.6	33.4	41.4	16.4	32.6	28.8
	Prompt ($S=256$)	(1.06M)	39.0	39.6	17.3	34.9	32.5	37.1	15.0	29.6	26.7
	Prompt ($S=256, F=16$)	(5.16M)	**36.9**	**27.2**	**14.1**	**27.2**	**30.0**	34.6	**12.8**	26.4	**22.2**
	Scratch	(306M)	39.6	76.0	56.1	52.5	92.7	**31.6**	13.5	**19.4**	29.5

Table 1. FIDs (lower the better) of image generation models on VTAB tasks. The number of trainable parameters (second column) are computed assuming 100 classes. The mean FID over 19 VTAB tasks (third column) and those for dataset with a small to mid-scale training data are reported. Complete results are in Appendix B.1.3. The **best** and the second best results are highlighted in each column.

(a) SUN397 (FID=7.7; NAR + Prompt) (b) DTD (FID=25.9; NAR + Prompt) (c) Resisc (FID=24.8; NAR + Prompt)

(d) SUN397 (FID=12.8; AR + Prompt) (e) DTD (FID=30.0; AR + Prompt) (f) Resisc: (FID=22.2; AR + Prompt)

Figure 5. Class conditional generation using NAR (top; S=128) and AR (bottom; S=256, F=16) transformers with prompt tuning.

[53] Mohamad Shahbazi, Zhiwu Huang, Danda Pani Paudel, Ajad Chhatkuli, and Luc Van Gool. Efficient conditional gan transfer with knowledge propagation across classes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12167–12176, 2021. 1, 2, 6, 7, 11, 15

[64] Yaxing Wang, Abel Gonzalez-Garcia, David Berga, Luis Herranz, Fahad Shahbaz Khan, and Joost van de Weijer. Minegan: effective knowledge transfer from gans to target domains with few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9332–9341, 2020. 1, 2, 6, 7, 11, 15

[2] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2018. 1, 6, 11

[24] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 6

[71] Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, et al. A large-scale study of representation learning with the visual task adaptation benchmark. *arXiv preprint arXiv:1910.04867*, 2019. 1, 2, 5

[6] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. *arXiv preprint arXiv*:2202.04200, 2022. 1, 2, 3, 4, 6, 7, 10

[14] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12873–12883, 2021. 1, 2, 3, 4, 6

4.2. Few-shot Generative Transfer

- We limit our study to transfer of an NAR transformer, i.e., MaskGIT [6], but with more comparisons to existing few-shot image generation models, either with [53, 64] or without [56, 73] knowledge transfer.
- We study few-shot generative transfer learning on Places [74], ImageNet [9], and Animal Face [54].
- For Places and ImageNet, we select **5 classes and use 500 images per class** for training.
- For Animal Face, we consider two scenarios – following [53], we use **100 images per class for training from 20 classes** (denoted as "Animal Face" in Tab. 2); alternatively, following [56, 73], we use **all images of dog (389) and cat (160) classes** for training.
- We tested our methods on challenging tasks from DomainNet [45] Infograph and Clipart (345 classes), and ImageNet sketch (1000 classes) [63], **using only 2 training images per class.**

- **Baselines:** GAN-based generative transfer learning methods, e.g., MineGAN [64] and cGANTransfer [53], are used as baselines. Moreover, we compare to few-shot image generation models, e.g., DiffAug [73] and LeCam GAN [56].
- **FID Calculation**: FIDs are computed using 10k generated images. However, for dog and cat face experiments, only 5k images are used, following the approach in [73].

Dataset (shot)	ImageNet (500)	Places (500)	Animal Face (100)	Dog Face (389)	Cat Face (160)
MineGAN [64]	61.8†	82.3†	–	93.0*	54.5*
cGANTransfer [53]	–	71.1‡	85.9‡	–	–
DiffAug [73]	–	–	–	58.5*	42.4*
LeCam GAN [56]	–	–	–	54.9*	34.2*
Ours (class)	**16.9**	24.2	16.3	65.4	40.2
Ours (instance)	19.6	**19.5**	**13.3**	**26.0**	**31.2**

Table 2. FIDs of image generation models on few-shot benchmark. Numbers with †, ‡, ∗ are from [64], [53], [56], respectively.

[6] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. *arXiv preprint arXiv*:2202.04200, 2022. 1, 2, 3, 4, 6, 7, 10

[53] Mohamad Shahbazi, Zhiwu Huang, Danda Pani Paudel, Ajad Chhatkuli, and Luc Van Gool. Efficient conditional gan transfer with knowledge propagation across classes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12167–12176, 2021. 1, 2, 6, 7, 11, 15

[64] Yaxing Wang, Abel Gonzalez-Garcia, David Berga, Luis Herranz, Fahad Shahbaz Khan, and Joost van de Weijer. Minegan: effective knowledge transfer from gans to target domains with few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9332–9341, 2020. 1, 2, 6, 7, 11, 15

[56] Hung-Yu Tseng, Lu Jiang, Ce Liu, Ming-Hsuan Yang, and Weilong Yang. Regularizing generative adversarial networks under limited data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7921–7931, 2021. 1, 7

(a) DomainNet Clipart (2 shot; FID=22.4) (b) DomainNet Infograph (2 shot; FID=20.6) (c) ImageNet Sketch (2 shot; FID=14.4)

Figure 6. Class conditional generation of few-shot transfer models. Images in red boxes are two training images of each class.

[73] Shengyu Zhao, Zhijian Liu, Ji Lin, Jun-Yan Zhu, and Song Han. Differentiable augmentation for data-efficient gan training. *Advances in Neural Information Processing Systems*, 33:7559–7570, 2020. 1, 7

[74] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. *Advances in neural information processing systems*, 27, 2014. 2, 7

[9] 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. 7

[54] Zhangzhang Si and Song-Chun Zhu. Learning hybrid image templates (hit) by information projection. *IEEE Transactions on pattern analysis and machine intelligence*, 34(7):1354– 1367, 2011. 2, 7

[53] Mohamad Shahbazi, Zhiwu Huang, Danda Pani Paudel, Ajad Chhatkuli, and Luc Van Gool. Efficient conditional gan transfer with knowledge propagation across classes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12167–12176, 2021. 1, 2, 6, 7, 11, 15

[45] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1406–1415, 2019. 7

[63] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 32, 2019. 7

- **Data Efficiency:** We studied data efficiency by training models on ImageNet, Places, and Animal Face datasets with 5, 10, 50, and 100 images per class, using a class-condition for image generation.

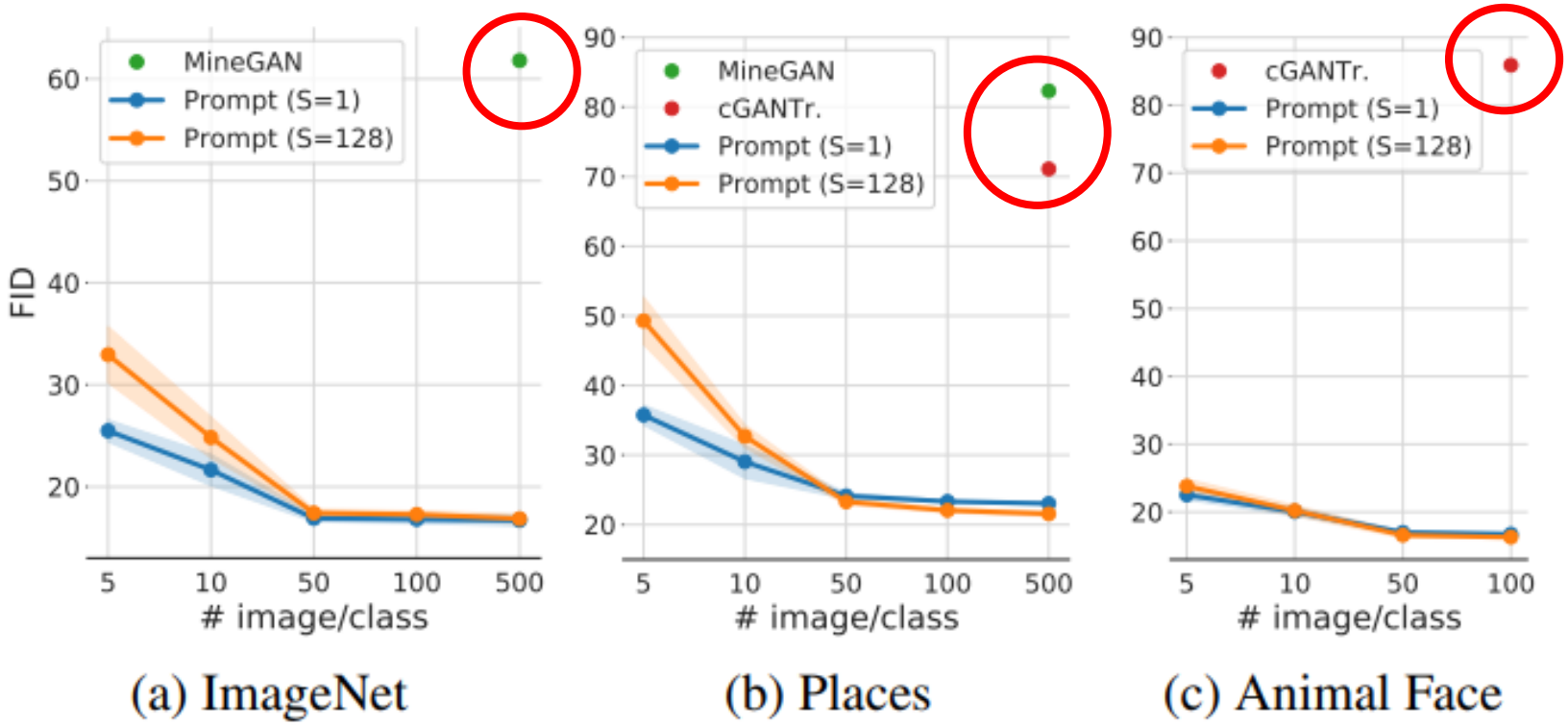
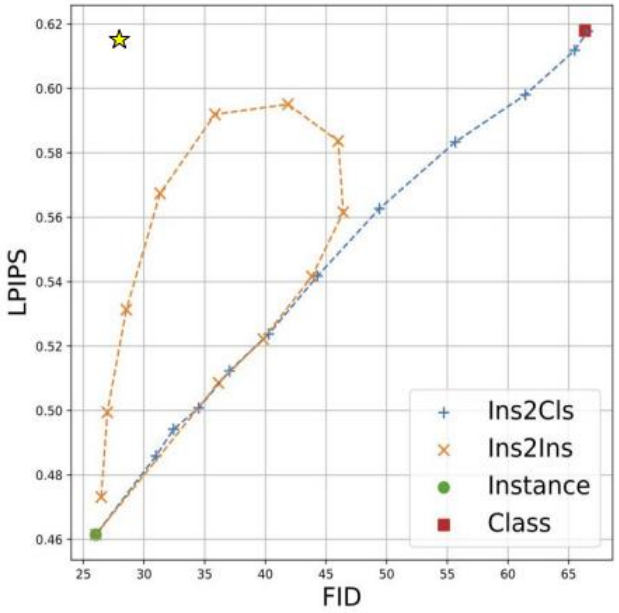


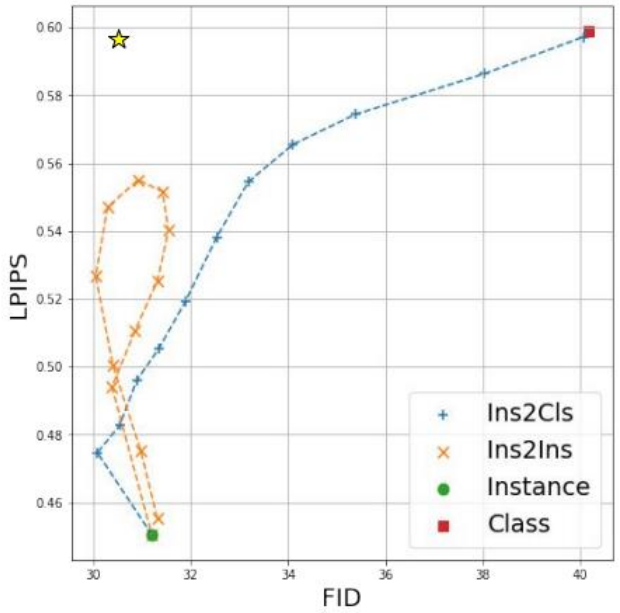
Figure 7. FIDs for models trained with varying numbers of images per class for class-conditional few-shot generative transfer.

Enhancing Generation Diversity via Prompt Engineering:

- Our model offers a way to enhance generation diversity by composing prompts.
- We conduct experiments on the dog and cat faces dataset using marquee header prompts with different T_{cutoff} values.
- For the fidelity metric, we compute the FID.
- To measure the diversity, we follow [42] and report a intra-cluster pairwise LPIPS distance, where we generate 5k samples and map them into one of training images.



(a) Dog Faces

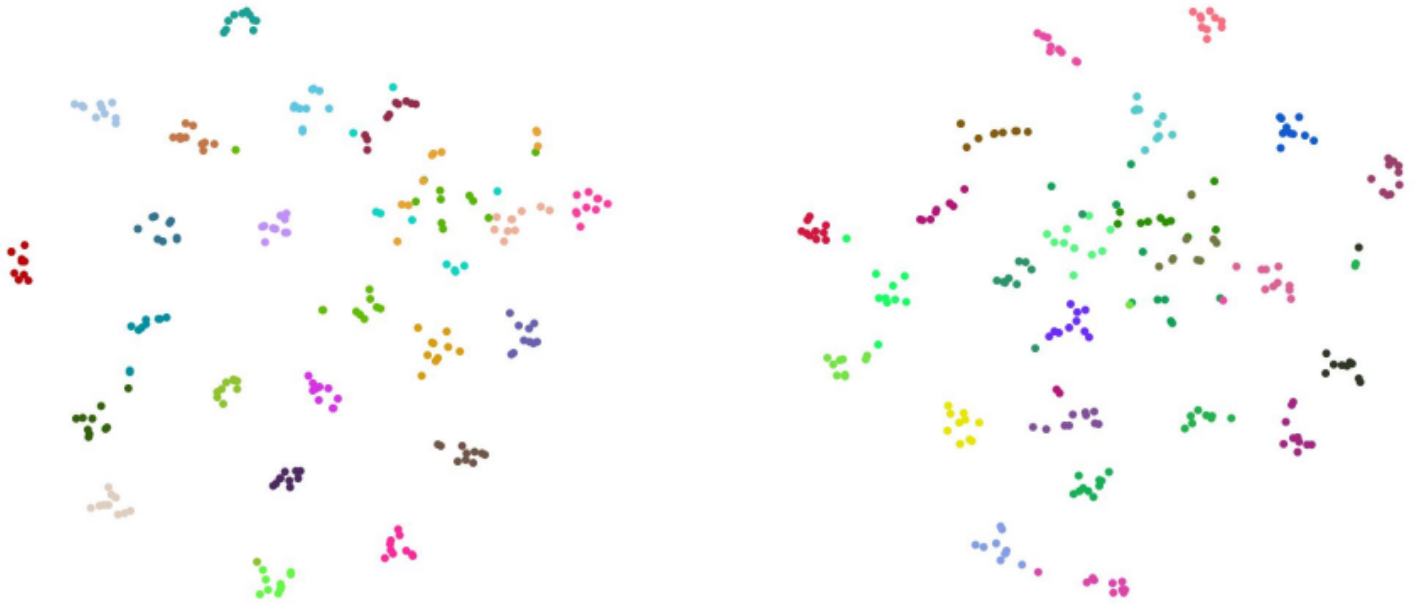


(b) Cat Faces

[42] Utkarsh Ojha, Yijun Li, Jingwan Lu, Alexei A Efros, Yong Jae Lee, Eli Shechtman, and Richard Zhang. Few-shot Image generation via cross-domain correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10743–10752, 2021. 1, 8

Analysis and Discussion
5.1. What does the Prompt Learn?

- The goal of this section is to investigate what prompts have learned.
- For this study, we train instance conditioned prompt models on flowers dataset of VTAB, with S = 1 and 128.
- We used a visualization technique called t-SNE [61] to **display the relationships between these prompts.** The results showed that **similar prompts (representing the same flower categories) cluster together.**
- We quantify our observation using a normalized mutual information (NMI) computed by clustering prompts.
- The results are better than the number obtained using an embedding from ImageNet pretrained ResNet-50 [21] (NMI=0.734).



(a) $S = 1$ (NMI=0.848) (b) $S = 128$ (NMI=0.800)

Figure 9. t-SNE plots of instance-conditioned prompt representations on flowers dataset. Points of the same color are from the same class. We also report normalized mutual information (NMI) score by clustering prompt representations using KMeans.

[61] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 9
[21] 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. 1, 9

5.2. Adaptation-Diversity Trade-Off

- In this section, we study prompts with various lengths, but on a single image.
- With short prompts, the model produces diverse but less detailed images. This implies that the short prompt learns concepts, while the long prompt learns fine details of training data.

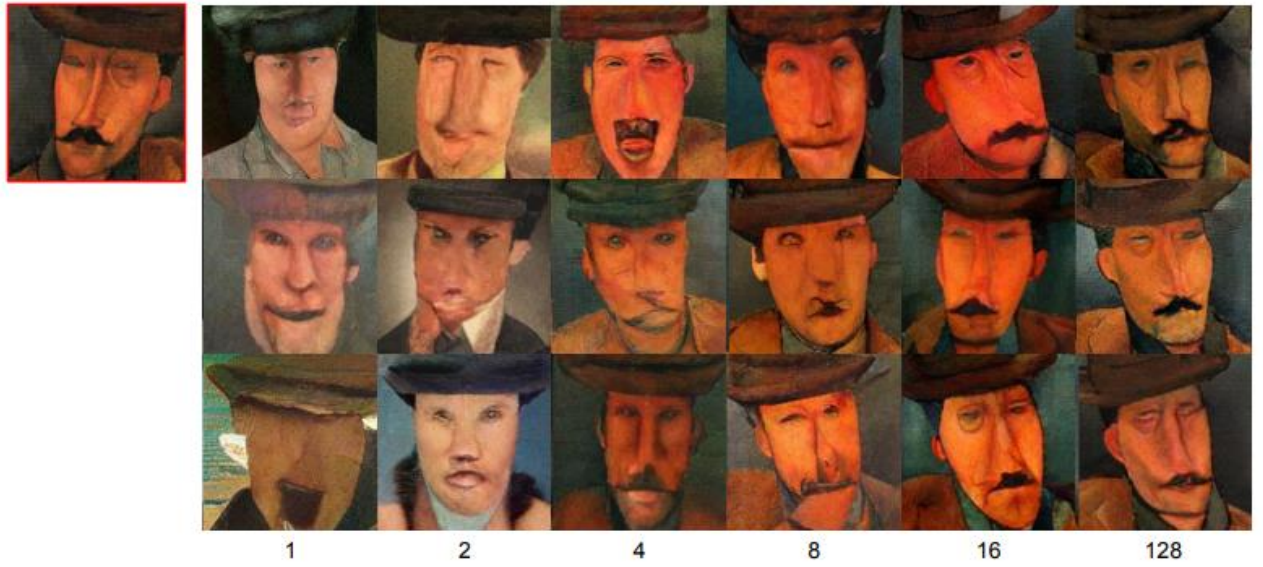


Figure 10. A single training image in red box and those generated by models using prompts of various lengths from 1 to 128.

- We visualize images generated by models of Sec. 5.1.

(a) Oxford Flowers, "Grape hyacinth" ($S = 1$)

(b) Oxford Flowers, "Grape hyacinth" ($S = 128$)

Figure 11. Instance-conditioned generation. For each row, leftmost image in red box is a training image and next five images are generated. When instance conditioned, generated images follow finer-grained details of the reference training image, such as color, shape, or background, beyond class information. Adaptation and diversity could be further controlled by the prompt length.

5.3. Ablation on Prompt Token Generators

- For models using prompts with the proposed factorization closely matches those using the baseline, non-factorized prompts.
- AR transformers achieve on par results with the baseline using less than 30% of parameters

NAR		# params	Small	Medium	Large	Natural	Struct.	Spec.
S=16	baseline	1.81M	18.6	34.6	89.1	23.8	50.9	41.7
	F=1	0.68M	18.6	36.1	89.5	25.2	51.9	41.5
	F=4	0.95M	18.6	35.5	88.4	24.4	51.5	41.4
	F=16	2.02M	18.5	35.0	86.8	24.3	50.8	40.4
S=128	baseline	10.4M	18.2	30.8	86.4	22.0	46.9	39.9
	F=1	0.76M	18.5	30.6	88.9	22.5	47.1	40.5
	F=4	1.30M	18.1	31.5	88.0	23.3	48.2	38.0
	F=16	3.39M	17.9	30.8	86.5	22.6	47.4	37.7
AR		# params	Small	Medium	Large	Natural	Struct.	Spec.
S=16	baseline	2.02M	30.5	41.9	82.7	28.5	61.9	41.7
	F=1	0.88M	34.5	43.3	83.9	32.3	62.9	42.9
	F=4	1.14M	31.9	42.3	82.7	29.9	62.0	42.0
	F=16	2.21M	31.2	41.9	82.6	28.9	61.9	41.6
S=256	baseline	20.4M	25.7	32.7	71.6	23.7	52.1	35.9
	F=1	1.06M	32.3	33.5	70.5	29.0	49.1	36.4
	F=4	1.88M	31.2	41.9	82.6	28.9	61.9	41.6
	F=16	5.16M	26.6	32.6	69.9	24.5	48.9	34.6

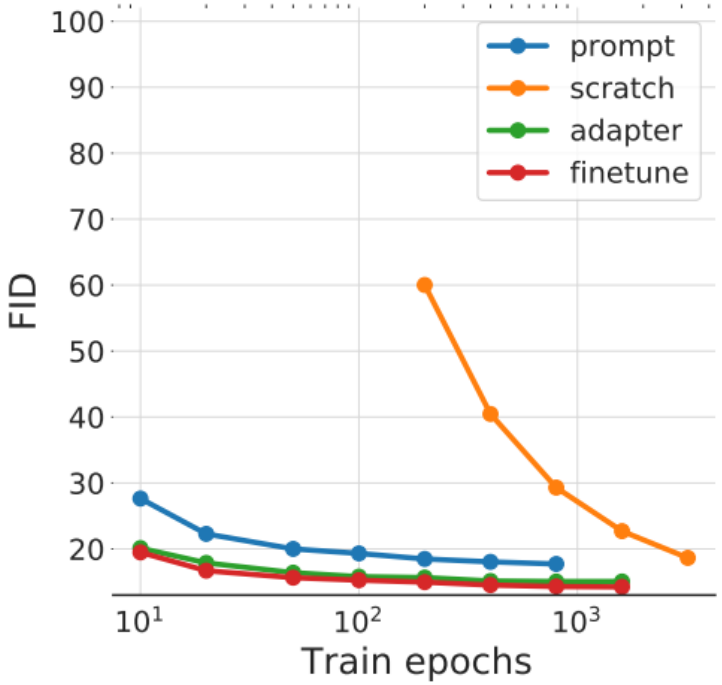
5.4. Beyond Prompt Tuning for Generative Transfer

- We conducted a comprehensive study on various learning methods for generative vision transformers, **ranging from 10 to 3200 training epochs, emphasizing the significance of training efficiency.**
- Adapter tuning [25] introduces learnable adapter modules to each transformer block, while fine-tuning unfreezes and updates pretrained weights..
- We integrated class-conditional prompts of length 1, randomly initialized, for both adapter tuning and fine-tuning.
- **Prompt tuning showcased superior efficiency**, with its trainable parameters being **less than 0.5%** of those in fine-tuning and learning from scratch.

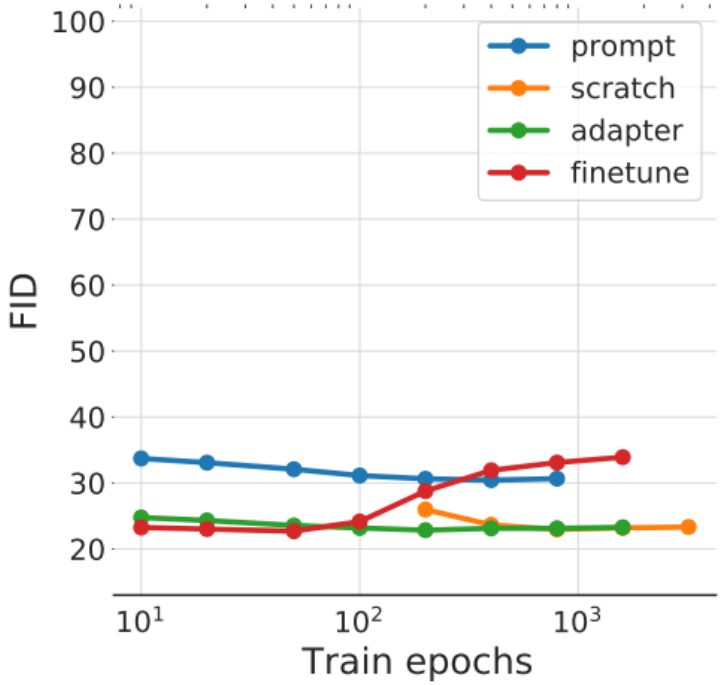
[25] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR, 2019. 3, 10, 11, 15

	# params	train / step	generation
Prompt tuning ($S = 128$)	0.76M	1×	1×
Adapter tuning	5.43M	1.04×	0.84×
Fine-tuning, Scratch	172M	1.67×	0.80×

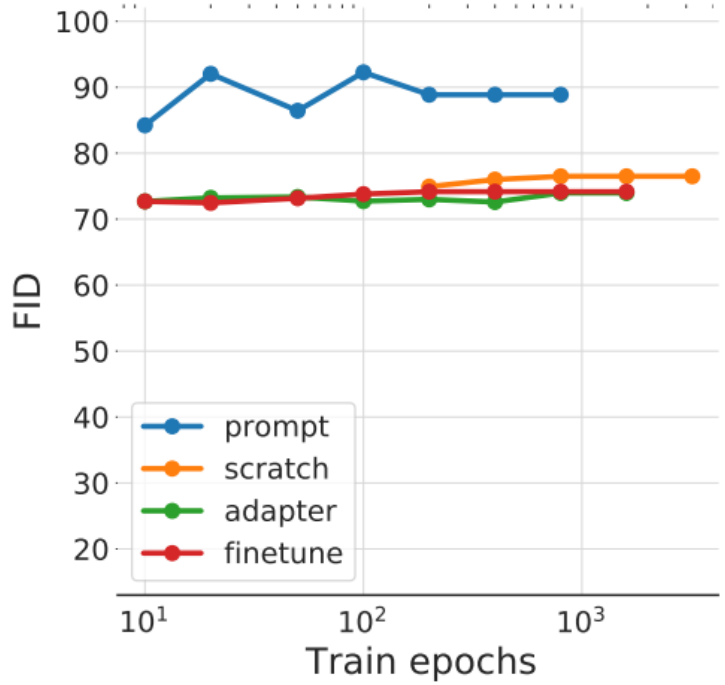
- Fig. 12 compares the generation performance in FID on VTAB.
- It requires almost 800 epochs for models learned from scratch to reach FIDs of the prompt tuning models trained for 10 epochs for tasks with a small data.
- **For tasks with a small training data, fine-tuning shows the best FIDs.** On the other hand, we find that **fine-tuning behaves unstable on some datasets** (e.g., smallnorb), and the performance diverges as training goes.



(a) VTAB small ($<10k$)



(b) VTAB medium ($<100k$)



(c) VTAB large ($>100k$)

6. Related Work

- TransferGAN [65], by fine-tuning on the target dataset, has shown that leveraging pre-trained knowledge enhances performance with limited data. Freezing certain discriminator layers [40] further boosts and stabilizes the training.
- MineGAN [64] introduces a miner, which projects random noise into the embedding space of the pretrained generator, and trains it with discriminator while fixing generator parameters.
- cGANTransfer [53] makes explicit transfer of knowledge on classes of the source dataset to new classes.

[53] Mohamad Shahbazi, Zhiwu Huang, Danda Pani Paudel, Ajad Chhatkuli, and Luc Van Gool. Efficient conditional gan transfer with knowledge propagation across classes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12167–12176, 2021. 1, 2, 6, 7, 11, 15

[64] Yaxing Wang, Abel Gonzalez-Garcia, David Berga, Luis Herranz, Fahad Shahbaz Khan, and Joost van de Weijer. Minegan: effective knowledge transfer from gans to target domains with few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9332–9341, 2020. 1, 2, 6, 7, 11, 15

[65] Yaxing Wang, Chenshen Wu, Luis Herranz, Joost van de Weijer, Abel Gonzalez-Garcia, and Bogdan Raducanu. Transferring gans: generating images from limited data. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 218–234, 2018. 11

7. Conclusion

- We introduce a method to learn image generation from diverse data using knowledge transfer from a large-dataset-trained source model.
- A tweak in prompt token design aids in learning efficient class and instance-specific image generation models of autoregressive and non-autoregressive vision transformers.
- Comprehensive experimental results of image synthesis are provided across diverse visual domains, tasks, and the number of training images.
- We also show how to use learned prompts for novel image synthesis in the form of marquee header prompts, especially beneficial when generating from limited image data.

Thank you for listening