Few-Shot Diffusion Models

Abstract

Denoising diffusion probabilistic models (DDPM) are powerful hierarchical latent variable models with remarkable sample generation quality and training stability. These properties can be attributed to parameter sharing in the generative hierarchy, as well as a parameter-free diffusion-based inference procedure. In this paper, we present Few-Shot Diffusion Models (FSDM), a framework for few-shot generation leveraging conditional DDPMs. FSDMs are trained to adapt the generative process conditioned on a small set of images from a given class by aggregating image patch information using a set-based Vision Transformer (ViT). At test time, the model is able to generate samples from previously unseen classes conditioned on as few as 5 samples from that class. We empirically show that FSDM can perform few-shot generation and transfer to new datasets. We benchmark variants of our method on complex vision datasets for few-shot learning and compare to unconditional and conditional DDPM baselines. Additionally, we show how conditioning the model on patch-based input set information improves training convergence.

1 Introduction

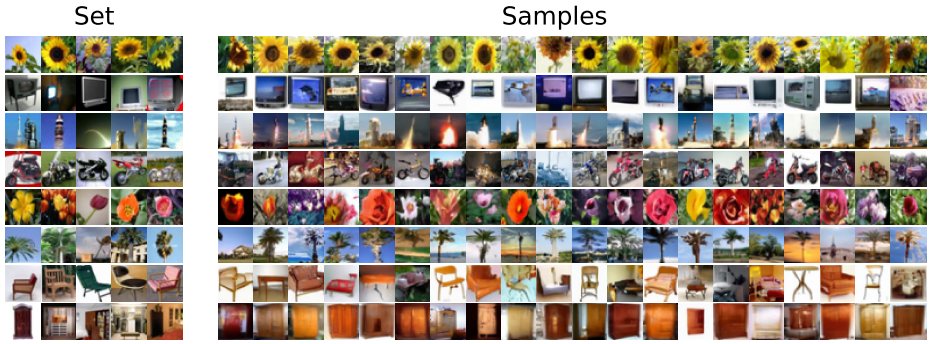


Figure 1: Set (left) and conditional samples (right) on CIFAR100 using a Few-Shot Diffusion Models on known classes. FSDM can extract content information from a handful of realistic examples and generate rich and complex samples from a variety of conditional distributions. More samples in Appendix Fig. 6.

Humans are exceptional few-shot learners able to grasp concepts and function of objects never encountered before [46, 85, 47]. This is because we build internal models of the world so we can combine our prior knowledge about object appearance and function to make well-educated inferences from very little data [86, 48, 88]. In contrast, traditional machine learning systems have to be trained tabula rasa and therefore need orders of magnitude more data. A particularly challenging problem is *few-shot adaptation in generative latent variable models* [16, 69, 71, 7]. Few-shot generation has been limited to simple datasets and shallow tasks, using handcrafted aggregation and conditioning mechanisms. Recently diffusion models [82, 29] have shown impressive generative performance for vision [55, 31, 30], language [33, 3], speech [44], biological data [34, 53, 96], and multimodal [56, 64] generation, providing an important step toward general and stable pure generative models. Unconditional diffusion models are expressive likelihood-based density estimators with high sample quality [40]. This expressivity arises from the Monte Carlo (layer) sampling that we can perform during training thanks to the special structure of the forward process: a parameter-free diffusion process for which the posterior can be computed at each step in closed form [29].

Under review at the NeurIPS 2022 Workshop on Score-Based Methods. Do not distribute.

However such effectiveness is at the cost of posterior flexibility and absence of latent space structure. For this reason the few-shot capacities of this class of models is largely unexplored and conditional adaptation is challenging. In this work we aim to study adaptation mechanisms and improve few-shot generation in latent variable models on realistic and complex visual data (Fig. 1). The setting we consider is that of learning from a large quantity of homogeneous sets, where each set is an un-ordered collections of samples of one concept or class. At test time, the model will be provided with sets of concepts never encountered during training. We consider explicit conditioning in a hierarchical formulation, where *global variables carry information about the set at hand*. The conditional hierarchical model can naturally represent a family of generative models, each specified by a different conditioning set-level variable. We formulate FSDM using vision transformers [14] and diffusion models [29, 55]. We propose to process the input set in patches and condition the generative model with a learnable attention mechanism using a tokenized representation for the input set. **Our contributions** are:

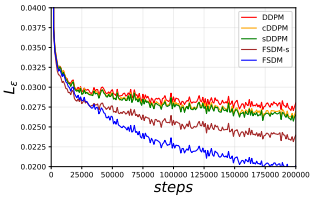
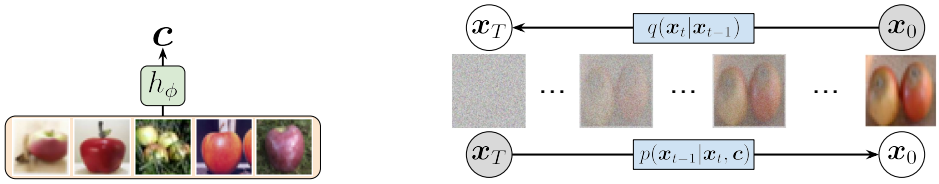


Figure 2: Estimated L_ϵ per layer on CIFAR100 during training. FSDM is data efficient during training and can denoise the data better and faster than unconditional and conditional DDPM baselines.

I) a new framework to perform few-shot generation for realistic sets of images in the DDPM framework. **II)** Learnable Attentive Conditioning (LAC), a conditioning mechanism where the input set is processed as a collection of patches and used to condition a DDPM through attention between sample-level and set-level variables. **III)** Experimental evidence that our model speeds up training, increases sample quality and variety, and improves transfer for conditional and few-shot generation compared to relevant unconditional and conditional DDPM-based baselines.

2 Few-Shot Diffusion Models

For background on diffusion models see Appendix A. For related work see Appendix B In this paper, our goal is to learn to quickly adapt to new generation tasks. That is, we want to perform few-shot generation conditioned on a set $\mathbf{X}$ containing previously unseen samples from a new task. We approach this problem using diffusion models: We learn a diffusion model $p_\theta(\mathbf{x}|\mathbf{X})$ conditioned on the set $\mathbf{X}$. We refer to our approach as Few-Shot Diffusion Models (FSDM). Our model can be broken down into two main parts: 1) A neural network h_ϕ that produces a context representation $\mathbf{c} = h_\phi(\mathbf{X})$ of the set $\mathbf{X}$, and 2) a conditional diffusion model that generates novel samples conditioned on the context $\mathbf{c}$. See Fig. 3 for an illustration.



(a) Context net h_ϕ: Takes the set $\mathbf{X}$ as input and outputs the context $\mathbf{c}$.

(b) Conditional diffusion: A denoising diffusion model where the generative path is conditioned on the context $\mathbf{c}$.

Figure 3: Few-Shot Diffusion Models.

Generative Model. The generative model is a conditional diffusion model

$$p_\theta(\mathbf{x}_{0:T}|\mathbf{X}) = p_\theta(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{c}), \quad \mathbf{c} = h_\phi(\mathbf{X}), \tag{1}$$

conditioned on the set $\mathbf{X}$ through the context $\mathbf{c}$ produced by the context net h_ϕ. In practice, we use a Vision Transformer (ViT [14]) as the context net h_ϕ, but we also experiment with a UNet encoder. We discuss the use of ViT in Sec. C.1. The generative path $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{c})$ is parameterized by a UNet, as is common practice for diffusion models. However, since we have an additional context $\mathbf{c}$, we need the UNet to fuse the information in $\mathbf{x}_t$ and $\mathbf{c}$ to predict $\mathbf{x}_{t-1}$. In this work we consider two main mechanisms for this, 1) a mechanism based on FiLM [61] and 2) we propose Learnable Attentive Conditioning (LAC), inspired by [73]. We discuss these in greater detail in Sec. C.2. The prior could also have been conditioned on $\mathbf{c}$, i.e. $p_\theta(\mathbf{x}_T|\mathbf{c})$, but for simplicity we use a standard unconditional Gaussian $p_\theta(\mathbf{x}_T) = \mathcal{N}(\mathbf{0}, \mathbf{I})$.

Inference Model. Given the special structure of DDPM and FSDM, the inference model is parameter-free and we do not need to condition on c. This is a great simplification during training. In practice FSDM employs a diffusion parameter-free posterior that degrades the information in the data at each step adding noise as presented in Eq. 3.

Loss and Training. The negative ELBO can be expressed as a conditional version of Eq. 5, $L_{\texttt{FSDM}} = L_0^c + \sum_{t=2}^{T} L_{t-1}^c + L_T^c$. As for regular DDPMs, the loss can be decomposed into a sum of terms, one per layer, that can be computed independently. Training thus enjoys the same benefits where we can get efficient stochastic estimates of the objective by Monte Carlo sampling terms. The conditional per-layer loss L_{t-1}^c can then formulated as

$$L_{t-1,\epsilon}^c = \mathbb{E}_{q(\epsilon)}\left[\|\epsilon_\theta(\mathbf{x}_t, \mathbf{c}) - \epsilon\|_2^2\right], \quad \mathbf{x}_t(\mathbf{x}_0, \epsilon) = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\,\epsilon. \tag{2}$$

L_T^c is unconditional and fixed in our model formulation, and L_0^c is a negated conditional discretized normal likelihood, $L_0^c = \mathbb{E}_{q(\mathbf{x}_1|\mathbf{x}_0)}\left[-\log p_\theta(\mathbf{x}_0|\mathbf{x}_1, \mathbf{c})\right]$.

3 Experiments

Setup. We use as backbone the standard DDPM model proposed in [29, 55] with fixed $T = 1000$ and linear β schedule. We reduce the number of channels in the model to 64 obtaining a 25M parameter model. We train using $L = L_\epsilon + \lambda L_{\texttt{vlb}}$ with $\lambda = 0.001$ from [55] that we found gave us a good balance between sample quality and training stability. We employ a Unet (10M) and a ViT (5M) as set encoders. We use the standard unconditional DDPM and conditional DDPM variants as baselines. In general our approach can be applied to condition any unconditional diffusion model [29, 41, 83]. For this reason we limit our attention to standard DDPM models with a discrete number of layers. More details on datasets, baselines, and additional experiments in Appendix D.

Few-Shot Generation. We compare the generative models in terms of denoising capacity, summing over T steps Eq. 2, FID [27] for sample quality, sFID [54] to capture spatial relationships, Precision and Recall [45] for measuring variety and mode coverage. We consider two main scenarios: in-distribution (In), testing on classes seen during training; and out-distribution (Out), testing on classes unknown during training (the few-shot scenario). We perform qualitative experiments on Omniglot, CIFAR100 and CelebA, and quantitative experiments on CIFAR100 and miniImageNet in Table 1 and Table 3. In Table 1 FSDM outperforms the unconditional and conditional baselines on both datasets and scenarios, providing evidence that a token-based representation jointly with cross-attention conditioning are effective mechanisms for few-shot generation. FSDM is a better denoiser (L_ϵ) and image generator (FID, sFID, P, R) than strong conditional baselines. We notice that in-distribution, DDPM and the conditional baselines (cDDPM and sDDPM) perform well as expected, but tend to under-perform out-distribution. This is expected for unconditional DDPM. For the conditional variants global aggregation is not expressive enough to represent complex novel realistic classes. Processing the set using FSDM-s tends to work better than the baselines but under-performs FSDM. Additionally, FSDM is an efficient learner, being able to extract more information from less data and converge faster than the baselines (Fig. 2).

Transfer. The goal of FSDM is to perform few-shot generation on objects never seen during training. However there are multiple challenges when dealing with new classes, in particular if these new classes are from new categories and datasets. Imagine to train a model on `cats` and `lions`, and then provide `tiger` at test time: even if the model has never seen a `tiger`, the encoder can extract information from a small set of tigers leveraging classes with similar animals. In a way the model can "interpolate" between the set at hand and similar classes. But if we train on `apples` and `oranges` and test on `tiger`, the model is challenged in a more fundamental way. There is no way to interpolate with known classes and the model has to rely mostly on the conditioning set. Between these two extremes there is a spectrum of challenges in few-shot generation and we want to explore how far our model can adapt to new information. For this reason we test few-shot transfer on a different dataset. We take models trained on CIFAR100 and test them (without gradient based adaptation) on MinImageNet. This is a difficult generalization task. We report results in the bottom part of Table 1. All the models struggle compared to new classes from the same dataset. However FSDM is still able to extract more information than the baselines, providing additional evidence that our framework can be used in a variety of small and large adaptation tasks.

Sampling. FSGM can sample known and unknown classes conditioning on few samples. When sampling known classes we are simply sampling conditional iid from a set $\mathbf{X}$ summarized by a certain

Table 1: Few-Shot generative evaluation on different datasets and for different metrics. We test few-shot generation on CIFAR100 and miniImageNet and transfer from CIFAR100 to MinimageNet. In CIFAR100 we use the original split and all the test classes are from new categories. In: in-distribution - we evaluate the models on known classes. The context $\mathbf{c}$ can be: V: deterministic-vector. T: deterministic-tokens. vV: variational-vector. vT: variational-tokens. Out: out-distribution - we evaluate the models on unknown classes (few-shot task). L_ϵ: denoising loss; FID: Frechet score; sFID: spatial FID; P: precision; R: recall. We do not use augmentation to train these models. We use 10K samples for the metrics and 250 steps. LAC: learnable attentive conditioning. FSDM performs better than the baselines on known and unknown classes, providing evidence that the token-based representation and the cross-attention conditioning mechanism are effective for few-shot generation in diffusion models.

				$\downarrow L_\epsilon$		$\downarrow$ FID		$\downarrow$ sFID		$\uparrow$ P		$\uparrow$ R	
	Enc	Cond	$\mathbf{c}$	In	Out	In	Out	In	Out	In	Out	In	Out
Few-Shot Generation **CIFAR100** (32)													
DDPM	-	-	-	6.92	8.14	15.35	62.84	18.03	28.91	0.66	0.58	0.56	0.40
cDDPM	Unet	FiLM	V	6.58	8.08	11.84	38.50	17.64	22.21	0.70	0.55	0.56	0.46
sDDPM	ViT	FiLM	T	6.70	8.17	13.34	45.50	21.32	29.87	0.67	0.54	0.55	0.46
FSDM-s (Ours)	ViT	LAC	T	5.81	7.72	12.39	40.71	**17.26**	22.12	0.68	0.57	0.57	0.44
FSDM (Ours)	ViT	LAC	T	**5.56**	**6.88**	**10.21**	**35.07**	**17.48**	**20.95**	**0.72**	**0.62**	**0.65**	**0.53**
Few-Shot Generation **miniImageNet** (32)													
DDPM	-	-	-	9.73	10.08	22.84	41.37	**20.01**	23.37	0.60	0.58	0.54	0.47
cDDPM	Unet	FiLM	V	9.50	10.12	17.47	32.22	20.04	**21.57**	0.65	0.59	0.55	0.52
sDDPM	ViT	FiLM	V	9.44	10.18	18.21	35.86	20.92	22.49	0.64	0.56	0.53	0.49
FSDM-s (Ours)	ViT	LAC	T	8.46	9.47	22.40	35.83	20.79	22.19	0.65	0.59	0.52	0.52
FSDM (Ours)	ViT	LAC	T	**7.76**	**8.30**	**15.39**	**30.62**	**19.83**	**21.84**	**0.67**	**0.64**	**0.61**	**0.56**
Transfer **CIFAR100** (32) $\rightarrow$ **miniImageNet** (32)													
DDPM	-	-	-	-	10.68	-	63.13	-	33.23	-	0.61	-	0.30
cDDPM	Unet	FiLM	V	-	10.77	-	41.00	-	**25.61**	-	0.59	-	0.39
sDDPM	ViT	FiLM	V	-	10.85	-	47.73	-	32.90	-	0.56	-	0.37
FSDM-s (Ours)	ViT	LAC	T	-	10.37	-	42.32	-	**25.74**	-	0.61	-	0.37
FSDM (Ours)	ViT	LAC	T	-	**9.60**	-	**39.55**	-	27.99	-	**0.65**	-	**0.45**

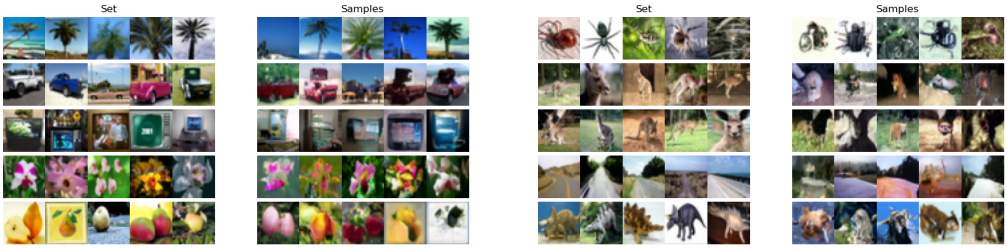


Figure 4: Few-Shot Conditional samples on CIFAR100 using a FSDM. Left side conditioning set and samples from in-distribution classes; right side conditioning set and samples from out-distribution classes. More samples in higher resolution in Appendix Fig. 7 and Fig. 8.

$\mathbf{c}$. When performing real few-shot sampling we condition on unknown classes using small sets of samples. In Fig 4 we show samples from known-classes (left panel) and from unknown-classes (right panel). The visual quality of the unknown classes is obviously worse than the known one. However the model can extract content information from few-samples and complex realistic classes in an effective way. We report additional visualizations on Omniglot and CelebA in Fig 10.

4 Conclusion

We presented Few-Shot Diffusion Models, a flexible framework to adapt quickly to different generative processes at test-time, leveraging advances in Vision Transformers and Diffusion Models. We show how conditioning a diffusion model with rich, expressive information gives superior performance in a wide range of experiments in and out-distribution. Few-shot generation is performed on realistic, complex sets of images, showcasing a promising direction for large-scale few-shot latent variable generative models.

References

[1] Marcin Andrychowicz, Misha Denil, Sergio Gomez, Matthew W Hoffman, David Pfau, Tom Schaul, Brendan Shillingford, and Nando De Freitas. Learning to learn by gradient descent by gradient descent. In *Advances in neural information processing systems*, pages 3981–3989, 2016.

[2] Oron Ashual, Shelly Sheynin, Adam Polyak, Uriel Singer, Oran Gafni, Eliya Nachmani, and Yaniv Taigman. Knn-diffusion: Image generation via large-scale retrieval. *arXiv preprint arXiv:2204.02849*, 2022.

[3] Jacob Austin, Daniel Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. *Advances in Neural Information Processing Systems*, 34, 2021.

[4] Fan Bao, Chongxuan Li, Jun Zhu, and Bo Zhang. Analytic-dpm: an analytic estimate of the optimal reverse variance in diffusion probabilistic models. *arXiv preprint arXiv:2201.06503*, 2022.

[5] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

[6] Dmitry Baranchuk, Ivan Rubachev, Andrey Voynov, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. *arXiv preprint arXiv:2112.03126*, 2021.

[7] Sergey Bartunov and Dmitry Vetrov. Few-shot generative modelling with generative matching networks. In *International Conference on Artificial Intelligence and Statistics*, pages 670–678, 2018.

[8] Andreas Blattmann, Robin Rombach, Kaan Oktay, and Björn Ommer. Retrieval-augmented diffusion models. *arXiv preprint arXiv:2204.11824*, 2022.

[9] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[10] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021.

[11] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. Ilvr: Conditioning method for denoising diffusion probabilistic models. *arXiv preprint arXiv:2108.02938*, 2021.

[12] 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.

[13] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34, 2021.

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[15] Conor Durkan and Yang Song. On maximum likelihood training of score-based generative models. *arXiv e-prints*, pages arXiv–2101, 2021.

[16] Harrison Edwards and Amos Storkey. Towards a neural statistician. *arXiv preprint arXiv:1606.02185*, 2016.

[17] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. *arXiv preprint arXiv:1703.03400*, 2017.

[18] Marta Garnelo, Dan Rosenbaum, Chris J Maddison, Tiago Ramalho, David Saxton, Murray Shanahan, Yee Whye Teh, Danilo J Rezende, and SM Eslami. Conditional neural processes. *arXiv preprint arXiv:1807.01613*, 2018.

[19] Marta Garnelo, Jonathan Schwarz, Dan Rosenbaum, Fabio Viola, Danilo J Rezende, SM Eslami, and Yee Whye Teh. Neural processes. *arXiv preprint arXiv:1807.01622*, 2018.

[20] Giorgio Giannone and Ole Winther. Hierarchical few-shot generative models. *arXiv preprint arXiv:2110.12279*, 2021.

[21] Jonathan Gordon, Wessel P Bruinsma, Andrew YK Foong, James Requeima, Yann Dubois, and Richard E Turner. Convolutional conditional neural processes. *arXiv preprint arXiv:1910.13556*, 2019.

[22] Erin Grant, Chelsea Finn, Sergey Levine, Trevor Darrell, and Thomas Griffiths. Recasting gradient-based meta-learning as hierarchical bayes. *arXiv preprint arXiv:1801.08930*, 2018.

[23] Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Jimenez Rezende, and Daan Wierstra. Draw: A recurrent neural network for image generation. *arXiv preprint arXiv:1502.04623*, 2015.

[24] William Harvey, Saeid Naderiparizi, Vaden Masrani, Christian Weilbach, and Frank Wood. Flexible diffusion modeling of long videos. *arXiv preprint arXiv:2205.11495*, 2022.

[25] Ali Hassani, Steven Walton, Nikhil Shah, Abulikemu Abuduweili, Jiachen Li, and Humphrey Shi. Escaping the big data paradigm with compact transformers. *arXiv preprint arXiv:2104.05704*, 2021.

[26] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377*, 2021.

[27] 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.

[28] Luke B Hewitt, Maxwell I Nye, Andreea Gane, Tommi Jaakkola, and Joshua B Tenenbaum. The variational homoencoder: Learning to learn high capacity generative models from few examples. *arXiv preprint arXiv:1807.08919*, 2018.

[29] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems 33*, 2020.

[30] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *Journal of Machine Learning Research*, 23(47):1–33, 2022.

[31] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022.

[32] Sepp Hochreiter, A Steven Younger, and Peter R Conwell. Learning to learn using gradient descent. In *International Conference on Artificial Neural Networks*, pages 87–94. Springer, 2001.

[33] Emiel Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forré, and Max Welling. Argmax flows and multinomial diffusion: Learning categorical distributions. *Advances in Neural Information Processing Systems*, 34, 2021.

[34] Emiel Hoogeboom, Victor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3d. *arXiv preprint arXiv:2203.17003*, 2022.

[35] Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. Meta-learning in neural networks: A survey. *arXiv preprint arXiv:2004.05439*, 2020.

[36] Chin-Wei Huang, Jae Hyun Lim, and Aaron C Courville. A variational perspective on diffusion-based generative models and score matching. *Advances in Neural Information Processing Systems*, 34, 2021.

[37] Bowen Jing, Gabriele Corso, Renato Berlinghieri, and Tommi Jaakkola. Subspace diffusion generative models. *arXiv preprint arXiv:2205.01490*, 2022.

[38] Alexia Jolicoeur-Martineau, Ke Li, Rémi Piché-Taillefer, Tal Kachman, and Ioannis Mitliagkas. Gotta go fast when generating data with score-based models. *arXiv preprint arXiv:2105.14080*, 2021.

[39] Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

[40] Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. On density estimation with diffusion models. *Advances in Neural Information Processing Systems*, 34, 2021.

[41] Diederik P Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *arXiv preprint arXiv:2107.00630*, 2021.

[42] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[43] Zhifeng Kong and Wei Ping. On fast sampling of diffusion probabilistic models. *arXiv preprint arXiv:2106.00132*, 2021.

[44] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761*, 2020.

[45] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *Advances in Neural Information Processing Systems*, 32, 2019.

[46] Brenden Lake, Ruslan Salakhutdinov, Jason Gross, and Joshua Tenenbaum. One shot learning of simple visual concepts. In *Proceedings of the annual meeting of the cognitive science society*, volume 33, 2011.

[47] Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

[48] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40, 2017.

[49] Seung Hoon Lee, Seunghyun Lee, and Byung Cheol Song. Vision transformer for small-size datasets. *arXiv preprint arXiv:2112.13492*, 2021.

[50] Yahui Liu, Enver Sangineto, Wei Bi, Nicu Sebe, Bruno Lepri, and Marco Nadai. Efficient training of visual transformers with small datasets. *Advances in Neural Information Processing Systems*, 34, 2021.

[51] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[52] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, 15(2018):11, 2018.

[53] Shitong Luo and Wei Hu. Diffusion probabilistic models for 3d point cloud generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2837–2845, 2021.

[54] Charlie Nash, Jacob Menick, Sander Dieleman, and Peter W Battaglia. Generating images with sparse representations. *arXiv preprint arXiv:2103.03841*, 2021.

[55] Alex Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. *CoRR*, 2021.

[56] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

[57] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021.

[58] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[59] Boris Oreshkin, Pau Rodríguez López, and Alexandre Lacoste. Tadam: Task dependent adaptive metric for improved few-shot learning. *Advances in Neural Information Processing Systems*, 31:721–731, 2018.

[60] Kushagra Pandey, Avideep Mukherjee, Piyush Rai, and Abhishek Kumar. Diffusevae: Efficient, controllable and high-fidelity generation from low-dimensional latents. *arXiv preprint arXiv:2201.00308*, 2022.

[61] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. *arXiv preprint arXiv:1709.07871*, 2017.

[62] Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. *arXiv preprint arXiv:2111.15640*, 2021.

[63] 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.

[64] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

[65] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021.

[66] Sachin Ravi and Alex Beatson. Amortized bayesian meta-learning. In *International Conference on Learning Representations*, 2018.

[67] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. 2016.

[68] 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.

[69] Scott Reed, Yutian Chen, Thomas Paine, Aäron van den Oord, SM Eslami, Danilo Rezende, Oriol Vinyals, and Nando de Freitas. Few-shot autoregressive density estimation: Towards learning to learn distributions. *arXiv preprint arXiv:1710.10304*, 2017.

[70] James Requeima, Jonathan Gordon, John Bronskill, Sebastian Nowozin, and Richard E Turner. Fast and flexible multi-task classification using conditional neural adaptive processes. *Advances in Neural Information Processing Systems*, 32, 2019.

[71] Danilo Jimenez Rezende, Shakir Mohamed, Ivo Danihelka, Karol Gregor, and Daan Wierstra. One-shot generalization in deep generative models. *arXiv preprint arXiv:1603.05106*, 2016.

[72] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014.

[73] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *arXiv preprint arXiv:2112.10752*, 2021.

[74] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[75] Andrei A Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. Meta-learning with latent embedding optimization. *arXiv preprint arXiv:1807.05960*, 2018.

[76] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022.

[77] Tom Schaul and Jürgen Schmidhuber. Metalearning. *Scholarpedia*, 5(6):4650, 2010.

[78] Pranav Shyam, Shubham Gupta, and Ambedkar Dukkipati. Attentive recurrent comparators. In *International Conference on Machine Learning*, pages 3173–3181. PMLR, 2017.

[79] Abhishek Sinha, Jiaming Song, Chenlin Meng, and Stefano Ermon. D2c: Diffusion-decoding models for few-shot conditional generation. *Advances in Neural Information Processing Systems*, 34, 2021.

[80] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, pages 4077–4087, 2017.

[81] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[82] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems 32*, 2019.

[83] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[84] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1199–1208, 2018.

[85] Joshua B Tenenbaum, Charles Kemp, Thomas L Griffiths, and Noah D Goodman. How to grow a mind: Statistics, structure, and abstraction. *science*, 331(6022):1279–1285, 2011.

[86] Joshua Brett Tenenbaum. *A Bayesian framework for concept learning*. PhD thesis, Massachusetts Institute of Technology, 1999.

[87] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021.

[88] Tomer D Ullman and Joshua B Tenenbaum. Bayesian models of conceptual development: Learning as building models of the world. *Annual Review of Developmental Psychology*, 2:533–558, 2020.

[89] Arash Vahdat, Karsten Kreis, and Jan Kautz. Score-based generative modeling in latent space. *Advances in Neural Information Processing Systems*, 34, 2021.

[90] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[91] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Advances in neural information processing systems*, pages 3630–3638, 2016.

[92] Daniel Watson, Jonathan Ho, Mohammad Norouzi, and William Chan. Learning to efficiently sample from diffusion probabilistic models. *arXiv preprint arXiv:2106.03802*, 2021.

[93] Mike Wu, Kristy Choi, Noah D Goodman, and Stefano Ermon. Meta-amortized variational inference and learning. In *AAAI*, pages 6404–6412, 2020.

[94] Tian Xie, Xiang Fu, Octavian-Eugen Ganea, Regina Barzilay, and Tommi Jaakkola. Crystal diffusion variational autoencoder for periodic material generation. *arXiv preprint arXiv:2110.06197*, 2021.

[95] Jin Xu, Jean-Francois Ton, Hyunjik Kim, Adam R Kosiorek, and Yee Whye Teh. Metafun: Meta-learning with iterative functional updates. *arXiv preprint arXiv:1912.02738*, 2019.

[96] Minkai Xu, Lantao Yu, Yang Song, Chence Shi, Stefano Ermon, and Jian Tang. Geodiff: A geometric diffusion model for molecular conformation generation. *arXiv preprint arXiv:2203.02923*, 2022.

[97] Daquan Zhou, Bingyi Kang, Xiaojie Jin, Linjie Yang, Xiaochen Lian, Zihang Jiang, Qibin Hou, and Jiashi Feng. Deepvit: Towards deeper vision transformer. *arXiv preprint arXiv:2103.11886*, 2021.

A Background

Few-Shot Generation. Few-shot generation is the task of adapting quickly to new classes or objects at test time given a small amount of instances from a novel category. In standard few-shot learning [46, 47, 35], given a support set $\mathbf{X}_s = \{\mathbf{x}_s\}_{s=1}^S$ and a query sample $\mathbf{x}_q$, we condition a learner on $\mathbf{X}_s$ and predict on $\mathbf{x}_q$ with a model of the form $p(\mathbf{x}_q|f_\phi(\mathbf{X}_s))$. The conditioning can be explicit on the representations [19, 18, 91, 80, 59, 84, 78] or implicit on the parameters like in meta-learning [77, 32, 17, 22, 66] and optimization [75, 67, 1]. Few-shot generation [47] borrows a similar setting but for the more challenging task to generate objects given few samples of that object at test time. In particular, given context information $\mathbf{X}$, we want to learn a conditional generative model that adapts quickly to new objects. There are two main ways to do this: learn a set-based generative model $p(\mathbf{X}) = \int p(\mathbf{X}|\mathbf{c})p(\mathbf{c})d\mathbf{c}$ and perform few-shot generation as a downstream task leveraging the per-set posterior $q_\phi(\mathbf{c}|\mathbf{X}_{\texttt{new}})$ as in [16, 28, 20]. Alternatively, we can learn a conditional model of the form $p(\mathbf{x}|\mathbf{X}) = \int p(\mathbf{x}|\mathbf{c})p(\mathbf{c}|\mathbf{X})d\mathbf{c}$ similarly to [7, 71, 69] and perform few-shot generation using the model directly for $p(\mathbf{c}|\mathbf{X}_{\texttt{new}})$. We consider the deterministic case of this, where the encoder $p(\mathbf{c}|\mathbf{X}) = \delta(\mathbf{c} - h_\phi(\mathbf{X}))$ is a deterministic set-based neural network $\mathbf{c} \leftarrow h_\phi(\mathbf{X})$.

Diffusion Denoising Probabilistic Models (DDPM). Let $\mathbf{x}_0$ denote the observed data which is either continuous $\mathbf{x}_0 \in \mathbb{R}^D$ or discrete $\mathbf{x}_0 \in \{0, ..., 255\}^D$. Let $\mathbf{x}_1, ..., \mathbf{x}_T$ denote T latent variables in $\mathbb{R}^D$. We now introduce, the *forward or diffusion process* q, the *reverse or generative process* p_θ and the objective L. The forward or diffusion process q is defined as [29]:

$$q(\mathbf{x}_{1:T}|\mathbf{x}_0) = \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1}), \quad q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t|\sqrt{1-\beta_t}\,\mathbf{x}_{t-1}, \beta_t I) \tag{3}$$

The beta schedule $\beta_1, \beta_2, ..., \beta_T$ is chosen such that the final latent image $\mathbf{x}_T$ is nearly Gaussian noise. The generative or inverse process p_θ is defined as:

$$p_\theta(\mathbf{x}_{0:T}) = p(\mathbf{x}_T)\prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t), \quad p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}|\mu_\theta(\mathbf{x}_t, t), \sigma_t^2 I), \tag{4}$$

where $p(\mathbf{x}_T) = \mathcal{N}(\mathbf{x}_T|0, I)$, and σ_t^2 often is fixed (e.g. to $\sigma_t^2 = \beta_t$). The neural network $\mu_\theta(\mathbf{x}_t, t)$ is shared among all time steps and is conditioned on t. The model is trained with a re-weighted version of the ELBO that relates to denoising score matching [82]. The negative ELBO L can be written as

$$L = \mathbb{E}_q\left[-\log\frac{p_\theta(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\right] = L_0 + \sum_{t=2}^{T} L_{t-1} + L_T, \tag{5}$$

where $L_0 = \mathbb{E}_{q(\mathbf{x}_1|\mathbf{x}_0)}\left[-\log p(\mathbf{x}_0|\mathbf{x}_1)\right]$ is the likelihood term (parameterized by a discretized Gaussian distribution) and, if $\beta_1, ...\beta_T$ are fixed, $L_T = \mathbb{KL}[q(\mathbf{x}_T|\mathbf{x}_0), p(\mathbf{x}_T)]$ is a constant. The terms L_{t-1} for $t = 2, ..., T$ can be written as $L_{t-1} = \mathbb{E}_{q(\mathbf{x}_t|\mathbf{x}_0)}[\mathbb{KL}[q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) \mid p(\mathbf{x}_{t-1}|\mathbf{x}_t)]]$. By further applying the reparameterization trick [42], the terms $L_{1:T-1}$ can be rewritten as a prediction of the noise ϵ added to $\mathbf{x}_0$ in $q(\mathbf{x}_t|\mathbf{x}_0)$. Parameterizing μ_θ using the noise prediction ϵ_θ, we can write

$$L_{t-1,\epsilon} = \mathbb{E}_{q(\epsilon)}\left[w_t\|\epsilon_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\,\epsilon, t) - \epsilon\|_2^2\right] + \text{const.} \tag{6}$$

where $w_t = \frac{\beta_t^2}{2\sigma_t^2\alpha_t(1-\bar{\alpha}_t)}$, which corresponds to the ELBO objective. The weights w_t can also be written in terms of signal-noise-ratio as proposed in [41]. Empirically [29] shows superior sample quality and stable training when using a re-weighted ELBO objective using $w_t = 1$ with a predictable drop in likelihood performance. We call this loss L_ϵ in this paper.

B Related Work

Few-Shot Latent Variable Models. Historically the machine learning community has focused its attention on supervised few-shot learning [46, 47], solving a classification or regression task on new classes at test time given a small number of labeled examples. The problem can be tackled using metric based approaches [91, 80, 59, 84, 78], gradient-based adaptation [17, 1, 32], optimization [67], and posterior inference [18, 21, 70, 22, 66]. More generally, the few-shot learning task can be recast as Bayesian inference in hierarchical modelling [22, 66]. In such models, typically parameters or representations are conditioned on the task, and conditional predictors are learned for such task. In [95] an iterative attention mechanism is used to learn a query-dependent task representation for supervised few-shot learning. Modern few-shot generation in machine learning was introduced in [46]. The Neural Statistician [16] is one of the first few-shot learning models in the context of VAEs [42, 72]. The model has been improved further increasing expressivity for the conditional prior using powerful autoregressive models [28], a non-parametric formulation for the context [93], hierarchical learnable aggregation for the input set [20], and exploiting supervision [19]. [71] proposed a recurrent and attentive sequential generative model for one-shot generation based on [23]. Powerful autoregressive decoders and gradient-based adaptation are employed in [69] for one-shot generation. The context $\mathbf{c}$ in this model is a deterministic variable. In GMN [7] a variational recurrent model learns a per-sample context-aware latent variable.

Vision Transformers. Transformers [90] have shown remarkable performance on unstructured text based data. Vision Transformers ([14], ViT) have recently emerged as a transformer variant to process image-like data using a patch-based approach. Then these patches are encoded as tokens and fed to a standard transformer encoder. ViT has been used for discriminative tasks with remarkable results [87, 97, 10, 51]. However less work has been done to use ViT in the context of generative latent variable models. Recently masked autoencoders [26], based on the ViT formulation, have been proposed for self-supervised learning and pretraining [5]. ViT variants for small and little data [49, 50, 25] have been proposed and our patch aggregation relies on similar ideas.

Conditional Diffusion Models. The standard DDPM [29] can be improved using likelihood-based training [15, 36], continuous time modeling [83, 89], learnable noise scales [41], efficient sampling mechanism [81, 4, 37, 76, 38, 92, 43], and exploiting powerful (variational) autoencoders for dimensionality reduction [73, 62, 60]. Methods to condition DDPM have been proposed, conditioning at sampling time [11], learning a class-conditional score [83], explicitly conditioning on class information [55, 57], physical properties [96, 94, 34], side information [6, 30], and temporal structure [31, 24]. We present a more general class of methods to condition diffusion models based on set-conditioning: We learn a parametric conditioning mechanism at the set-level and a conditional diffusion process at the sample-level. A conditional DDPM has been proposed for point-cloud generation [53]. However they limit their attention to a specific application, where we consider the general idea to encode sets of generic data and use for few-shot generation and transfer. Retrieval-based approaches [8, 2] use an external database and pre-trained contrastive embeddings [63]. We similar increase expressivity using a collection of sample similar to the input selecting a set from the same class without relying on a retrieval mechanism, pre-trained model and an external database. In [73] DDPM leverages a large vector quantized [58] pretrained autoencoder to encode the data in latent space, and such encoder can be used to condition on generic data. This approach is effective but expensive and rely on large pre-trained models on relevant datasets not always easily available. VAE-DDPM models [89, 60] have been proposed to learn conditional models in latent space. Text-to-image diffusion models [56, 64] have been recently proposed for guided generation. Our approach can be easily adapted to work with text tokens instead of visual tokens, simply changing the patch encoder.

C Conditioning the model

C.1 ViT as Set Encoder

Transformers [90] are the de-facto standard for natural language processing tasks and text generation [12, 9]. Recently Vision transformers (ViT, [14]) unlocked the general power of attention for vision tasks. However the use of transformers in latent variable models for generation is still limited. ViT gives us a flexible way to process images at the patch level. We adapt ViT to handle sets of images (sViT) similarly to [49]: in particular we want to handle small sets (1-10) of images. The fundamental idea is that we want to extract global information from the set and each patch should contain global information for a specific region in the image (Fig. 5). In general we can condition the ViT encoder on the layer embedding t using $\texttt{ViT}(\mathbf{X}, t; \phi)$: doing so we obtain a cheap way to learn a per layer-dependent context, coarse for large T, and more refined for small t. The use of tokens as input opens the door for a general domain-agnostic few-shot latent variable generator: our approach can be effortlessly employed for few-shot and conditional generation with any modality (text, speech, vision) simply tokenizing the input set and fine-tuning the patch embedding layer, without the need of any modification to the set-encoder, conditioning mechanism or generative process.

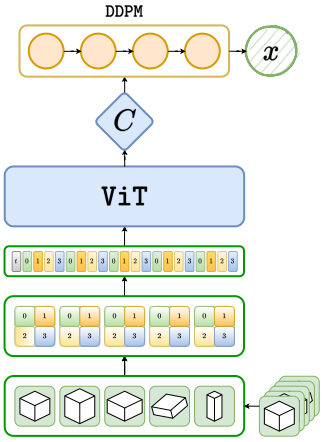


Figure 5: sViT architecture. The input is a set $\mathbf{X}$ of images. These are split in non-overlapping patches and fed to a transformer encoder using a shared positional encoding, as indicated by the patch colors. The sViT outputs a context as a vector (V) or collection of visual tokens (T). The DDPM is conditioned on this information using FiLM or attention.

C.2 Conditioning the Generative Process

After processing the patches and obtaining $\mathbf{c}$, we need to find a way to condition the DDPM generative path. We work with $\mathbf{c}$ in two different forms: A *vector* $\mathbf{c} \in \mathbb{R}^d$ or a collection of N *tokens*, $\mathbf{c} \in \mathbb{R}^{N \times d}$ as summarized in Table 2. In this work we consider two main conditioning mechanisms: FiLM [61] and Learnable Attentive Conditioning (LAC), inspired by [73].

Table 2: Different conditioning mechanisms.

	$\mathbf{c}$	Cond
FiLM	$\mathbb{R}^d$	$m(\mathbf{c})\mathbf{u} + b(\mathbf{c})$
LAC	$\mathbb{R}^{N \times d}$	$\texttt{att}(\mathbf{u}, \{\mathbf{c}_p\}_{p=1}^{N_p})$

Vector (V). One approach can be to condition the intermediate feature maps $\mathbf{u}$ in the DDPM UNet on $\mathbf{c}$, for example using a FiLM like mechanism [61], $\mathbf{u} = m(\mathbf{c})\mathbf{u} + b(\mathbf{c})$, where m and b are learnable and context-dependent. Given the special structure of the generative model, where all the layers share parameters θ and differ only through an embedding of the step t, the conditioning mechanism can be generically written as $\mathbf{u} = f(\mathbf{u}, \mathbf{c}, t) = f(\mathbf{u}_t, \mathbf{c}_t)$. Merging together $\mathbf{c}$ with the step embedding we can condition each layer, defining a generic per-step conditioning mechanism. In practice we found $\mathbf{u}(\mathbf{c}, t) = m(\mathbf{c})\mathbf{u}(t) + b(\mathbf{c})$ being the best performing and flexible approach.

Tokens (T). Alternatively, $\mathbf{c}$ can be a collection of variables $\mathbf{c} = \{\mathbf{c}_{sp}\}_{s=1,p=1}^{N_s,N_p}$, where N_p is the number of patches per sample, and N_s the number of samples in the set. In this case, attention can be used to fuse information between the context $\mathbf{c}$ and the feature maps $\mathbf{u}$. In principle we could use the patches directly, i.e. $\mathbf{u} = \texttt{att}(\mathbf{u}, \{\mathbf{c}_{sp}\}_{s=1,p=1}^{N_s,N_p})$. However, this approach scales badly with the number of samples in the set N_s. Another option is to use a per-patch aggregation where we average of the set dimension $\mathbf{c}_p = \frac{1}{N_s}\sum_{s=1}^{N_s} \mathbf{c}_{sp}$ to obtain N_p tokens $\{\mathbf{c}_p\}_{p=1}^{N_p}$ that we feed to the ViT. We then use cross-attention [73] on the per-patch averaged tokens $\mathbf{u} = \texttt{att}(\mathbf{u}, \{\mathbf{c}_p\}_{p=1}^{N_p})$ to condition DDPM. Using per-patch aggregation, we can process any number of samples without increasing the number of tokens used to condition DDPM and, more importantly, aggregate information from different samples in the context $\mathbf{c}$.

C.3 Variational FSDM

Alternatively to our formulation of FSDM, we could specify a latent variable model where the context $\mathbf{c}$ is a latent variable and the set $\mathbf{X}$ is generated conditioned on $\mathbf{c}$. We refer to this model as Variational

FSDM (VFSDM) and write it like

$$p_\theta(\mathbf{X}_{0:T}, \mathbf{c}) = p_\theta(\mathbf{c}) \left[\prod_{s=1}^{S} p_\theta(\mathbf{x}_{0:T}^{(s)} | \mathbf{c}) \right], \quad p_\theta(\mathbf{x}_{0:T} | \mathbf{c}) = p_\theta(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{c}). \tag{7}$$

In this case, the inference model will be a combination of the parameter-free diffusion posterior and a parameterized encoder for $\mathbf{c}$,

$$q_\phi(\mathbf{X}_{1:T}, \mathbf{c} | \mathbf{X}_0) = \underbrace{q_\phi(\mathbf{c} | \mathbf{X}_0)}_{\texttt{Set Encoder}} \underbrace{\left[\prod_{s=1}^{S} q(\mathbf{x}_{1:T}^{(s)} | \mathbf{x}_0^{(s)}, \mathbf{c}) \right]}_{\texttt{Diffusion}}, \quad q(\mathbf{x}_{1:T} | \mathbf{x}_0, \mathbf{c}) = \prod_{t=1}^{T} q(\mathbf{x}_t | \mathbf{x}_{t-1}, \mathbf{c}). \tag{8}$$

Furthermore, the negative ELBO will contain an extra KL term between the encoder $q_\phi(\mathbf{c}|\mathbf{X}_0)$ and the prior $p(\mathbf{c})$,

$$L_{\texttt{VFSDM}} = \mathbb{E}_{q_\phi(\mathbf{c}|\mathbf{X}_0)}[L_{\texttt{FSDM}}] + \mathbb{KL}\left[q_\phi(\mathbf{c}|\mathbf{X}_0) \| p_\theta(\mathbf{c})\right] \tag{9}$$

We originally worked with this model, but found the training to be more challenging, resulting in under-performance or poor conditioning properties.

D Experiments

D.1 Experimental Setup

Baselines. We compare FSDM with unconditional and conditional baselines. For the conditional baselines, we adapt conditional diffusion models in the literature [13, 79, 89] to the few-shot class generation scenario. We use a DDPM [29, 55] as unconditional baseline. We then compare with two main conditional diffusion models: a cDDPM, where a Unet encoder [74] processes independently the images in $\mathbf{X} = \{x_s\}_{s=1}^{S}$ and aggregates using a mean operator, $\mathbf{c} = \frac{1}{S}\sum_{s=1}^{S} f_\phi(x_s)$; the Unet encoder has the same structure of the guiding classifier in [13] used to learn class-conditional models. We then consider a sDDPM adapting ideas in [79] without contrastive learning, where a ViT encoder [14] splits $\mathbf{X}$ in patches and processes them jointly, as depicted in Fig. 5, and aggregates all the patches using a mean operator, $\mathbf{c} = \frac{1}{N_s N_p}\sum_{s=1}^{N_s}\sum_{p=1}^{N_p} \mathbf{c}_{sp}$. We also train a variational variant inspired by [89], vDDPM, that uses standard amortized variational inference [39, 42] on the per-set latent variable. For cDDPM, sDDPM and vDDPM, $\mathbf{c}$ is a vector and the conditioning mechanism is FiLM based [61]. We also compare explicit conditioning with test-time (or sampling-time) conditioning as proposed in [11].

We consider also two variants for FSDM, called FSDM-s and vFSDM: FSDM-s employees a different way to extract and aggregate set-information using ViT: we stack all the samples on the channel dimension and process them as one entity as proposed in [49]. vFSDM is a variational formulation where we learn a distribution over a set of tokens in the conditioning mechanism and we deal with amortized variational inference with quantized latents. We adapt the relaxation proposed in [65, 56] and the vector quantization proposed in [58, 68] to learn this model.

Datasets. We extensively test the baselines and our approach on 4 image datasets with different complexity and size: Omniglot (28) [46] using the binarization provided in [7]. FS-CIFAR100 (32) [59] using the original class split and mixing all the classes together (CIFAR100mix). miniImageNet (32) [91, 67] dataset to test few-shot and transfer capacity. CelebA (64) [52] for additional visualizations. We refer to FS-CIFAR100 as CIFAR100 in the following.

D.2 Test-time Conditioning

In this paper we argue the case that explicit adaptation during training is a powerful way to condition diffusion models. However inference-time adaptation [11] has also been shown to be an effective mechanism to adapt diffusion models to new distributions and does not require retraining the model or a parametric encoder. In Table 3 we compare the ILVR method, a powerful conditioning mechanism at sampling time, with FSDM, that condition the model during training, in the few-shot generation scenario, i.e. conditioning on samples from unknown classes during testing. We use the same datasets and evaluation procedure proposed in Table 1. We see that ILVR improves the result compared to a standard unconditional DDPM in terms of few-shot generation. However the task is challenging and the adaptation we require is on new classes and not only on new attributes. FSDM outperforms ILVR in this setting, providing additional evidence that explicit conditioning during training is essential for few-shot generation of realistic and complex objects.

Table 3: Few-Shot metrics test set (new classes) for different datasets. We compare unconditional DDPM, test-time adaptation with ILVR and FSDM.

	↓ FID	↑ P	↑ R
CIFAR100			
DDPM [29]	62.84	0.58	0.40
ILVR [11]	45.83	**0.62**	0.38
FSDM (Ours)	**35.07**	**0.62**	**0.53**
miniImageNet			
DDPM [29]	41.37	0.58	0.47
ILVR [11]	41.68	0.59	0.46
FSDM (Ours)	**30.62**	**0.64**	**0.56**
CIFAR100 ↓ miniImageNet			
DDPM [29]	63.13	0.61	0.30
ILVR [11]	53.12	0.58	0.31
FSDM (Ours)	**39.55**	**0.65**	**0.45**

D.3 Additional Experiments

In this section we discuss additional experiments and visualizations for FSDM.

Fig. 6 shows conditioning sets with cardinality 5 (left) and 20 conditional samples (right) for a FSDM using a large number of in-distribution classes from CIFAR100. We can see that samples are high quality and have large variability.

In Fig. 7 we compare conditional samples on in-distribution classes using CIFAR100 (top) and CIFAR100-mix (bottom). In Fig. 8 we perform the same experiment on out-of-distribution classes. When using CIFAR100 the out-of-distribution sets are not only from novel classes never seen during training but also from new categories, i.e. training on cats and testing on cars. When using

CIFAR100-mix all the classes are mixed and the out-of-distribution sets are from new classes but not necessarily from novel categories, i.e. training on `cats` and testing on `tigers`.

Sample quality and variety decrease for out-of-distribution samples as expected. When FSDM is presented with out-of-distribution sets from CIFAR100, the model cannot rely on similar classes and few-shot samples have reduced variability. But when presented with out-of-distribution sets from CIFAR100mix, the model can rely on similar classes and the few-shot generation task is easier, giving rise to better samples. Smarter conditioning mechanisms and expressive set-level latent variables can help in improving generalization and we will explore such possibilities in future work.

In Fig. 9 we use L_ϵ as a proxy signal to evaluate the capacity of DDPM and FSDM to distinguish between in-distribution and out-of-distribution samples. FSDM performs better than DDPM for this task, consistently distinguishing better between in-distribution and out-of-distribution classes.

In Fig. 10 we show visualizations for Omniglot and CelebA using short training runs. We train FSDM only for 100K iterations. The goal is to explore how fast the conditioning mechanism can extract and aggregate context information. As expected for a simple dataset as Omniglot the conditioning quality is high. For CelebA the conditioning mechanism struggles using so few training iterations. However, most of the context information is extracted successfully and samples are compatible with the content in the conditioning sets.

Figure 6: Set (left) and conditional samples (right) on CIFAR100 using a Few-Shot Diffusion Models. FSDM can extract content information from an handful of realistic examples and generate rich and complex samples from a variety of conditional distributions.

Figure 7: Conditional samples on CIFAR100 and CIFAR100mix using a FSDM. Left side conditioning set. Right side samples. Top in-distribution (known classes) on CIFAR100. Bottom in-distribution (known classes) on CIFAR100mix.

Figure 8: Few-Shot samples on CIFAR100 and CIFAR100mix using a FSDM. Left side conditioning set. Right side samples. Top out-distribution for CIFAR100 (unknown classes from unknown category). Bottom: out-distribution for CIFAR100mix (unknown classes from known category) .

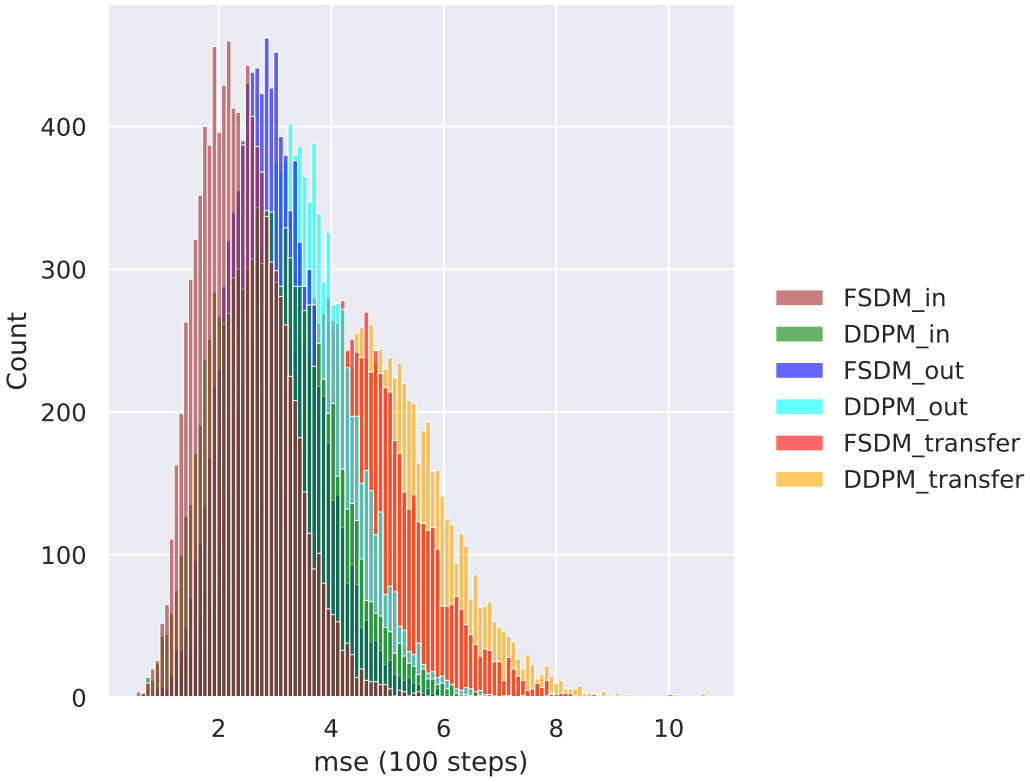


Figure 9: CIFAR100 in-distro, out-distro, transfer using L_ϵ computed with 100 steps of denoising.

Figure 10: Conditional samples on Omniglot and CelebA using a FSDM with deterministic context. Random samples. We train the models for only 100K iterations with batch size 32. Even training for such short time, we can see that FSDM can extract content information from small complex sets and condition the generative path in a consistent way. Left side conditioning set; right side samples.

E Experimental Details

Table 4: Relevant Hyperparameters for FSDM. $L_{\mathtt{h}}$: Loss hybrid [57].

	Omniglot	CelebA	FS-CIFAR100	miniImageNet
Dimension	1x28x28	3x64x64	3x32x32	3x32x32
Number classes	1623	6349	100	96
Classes Train	1000	4444	60	60
Classes Val	200	635	20	16
Classes Test	423	1270	20	20
Batch size	32	16	32	32
Channels $\mathbf{c}$	128	128	128	128
Channels model	64	64÷128	64	64
Channel multiplier	(1, 1, 2, 4)	(1, 1, 2, 4)	(1, 1, 2, 4)	(1, 1, 2, 4)
Channels $\mathbf{z}$	64	64	64	64
Classes per set	1	1	1	1
Heads	12	12	12	12
Iterations	100K	100K	200K	200K
Layers	6	6	6	6
Learning rate	$2e^{-4}$	$2e^{-4}$	$2e^{-4}$	$2e^{-4}$
Likelihood	$\mathcal{N}$	$\mathcal{N}$	$\mathcal{N}$	$\mathcal{N}$
Loss	$L_{\mathtt{h}}$	$L_{\mathtt{h}}$	$L_{\mathtt{h}}$	$L_{\mathtt{h}}$
Optimizer	Adam	Adam	Adam	Adam
Set size	5	5	5	5