Appendix

A. Implementation Details

A.1. Training Setup

For the training of our models, we generally follow the training settings in [1], with a modification on the training objective, *i.e*. using the masked language modeling objective described in Section 3 instead of the original diffusion-based denoising objective. Additionally, we adopt a larger batch size of 2048 and correspondingly increase the learning rate to 4e-4 for the ImageNet dataset. The results on CC3M are based on a publicly available pre-trained Muse model from github[1].

A.2. Strategy Optimization

For generation strategy optimization, we adopt SGD [22] as our optimizer for the generation strategy and perform 100 epochs of gradient descent. We initialize the strategies with heuristic-based configurations and simply set $k = \lambda = 1$. Notably, initialization is not decisive for the effectiveness of our method, as we find a highly naive initialization (initializing $\tau_1(t) = \tau_2(t) = s(t) = 1$, and $r(t)$ to linearly decrease from 1 to 0) can also perform well:

	Baseline	Our-Init.	Naive-Init.
FID-50K	8.40	**4.30**	4.45 (+0.15)

Here we take AutoNAT-S on ImageNet-256 as an example, both our adopted initialization and the naive initialization significantly outperform the baseline result (in gray).

The learning rates are set to 0.1 for hyperparameters $\boldsymbol{\tau_1}, \boldsymbol{\tau_2}, \boldsymbol{s}$ and 0.001 for $\boldsymbol{r}$. The step sizes for numerical gradient estimation are set to 0.1 for $\boldsymbol{\tau_1}, \boldsymbol{\tau_2}, \boldsymbol{s}$ and 0.01 for $\boldsymbol{r}$. For results on ImageNet-512, we additionally incorporate a gradient clipping of 10 to stabilize the training process.

Regarding the optimization of the training strategy, we perform a greedy search in $(\alpha/\beta, \beta)$ space in practice, as we find this implementation technique results in a more stable optimization process. The step size for line search in the greedy optimization process is set to 1. To expedite the optimization process, we train models under each strategy for only 50K steps, assessing performance as a proxy indicator for a fully converged model.

B. Additional Results

Visualization of the optimized strategies. In Figure 1, we visualize the optimized strategies for AutoNAT-S on ImageNet-256 as an example ($T = 4$).

CLIP score-based AutoNAT. In Section 5, we utilize the Fréchet Inception Distance (FID) [16] for optimization evaluation. However, our AutoNAT framework is general and flexible, which can be easily extended beyond FID. The table below adopts the CLIP score as the evaluation metric, showcasing the adaptability of AutoNAT in adopting various evaluation metrics:

Dataset	T	U-ViT [1]	AutoNAT
MS-COCO [19]	8	0.296	**0.314**
Dataset	T	Muse [2]	AutoNAT-Muse
CC3M [23]	8	0.283	**0.292**

To maintain consistency with prior works [2, 18], here we adopt ViT-B/32 to calculate the CLIP score.

Additional visualization results. We provide a qualitative comparison between AutoNAT and baseline (heuristic configuration) on both class-conditional and text-to-image generation in Figure 2 and Figure 3. Here we combine FID with other metrics such as IS, CLIP score as the evaluation metric F as we find this yields better results. For text-to-image generation, we adopt a muse model trained on large-scale text-to-image dataset [21].

C. Limitations and Future Work

While AutoNAT effectively improves the configuration of non-autoregressive Transformers, future research could fruitfully focus on enhancing the interpretability of the identified strategies. This could provide clearer insights for improved training and generative paradigms. Additionally, extending the evaluation metric within the AutoNAT framework through a wider array of metrics [4, 17, 35] represents a promising direction. Moreover, exploring the application of AutoNAT across diverse generative tasks and domains [3, 5–7, 15] or even synthesizing datasets for traditional visual perception tasks [14, 24, 33, 34], offers the potential for broadening its impact. Finally, integrating advanced training methods [20, 25, 26, 28, 29, 31], architectural innovations [27, 30, 32], and adaptive inference techniques [8–13] could further enhance the capabilities and applicability of non-autoregressive Transformers.

References

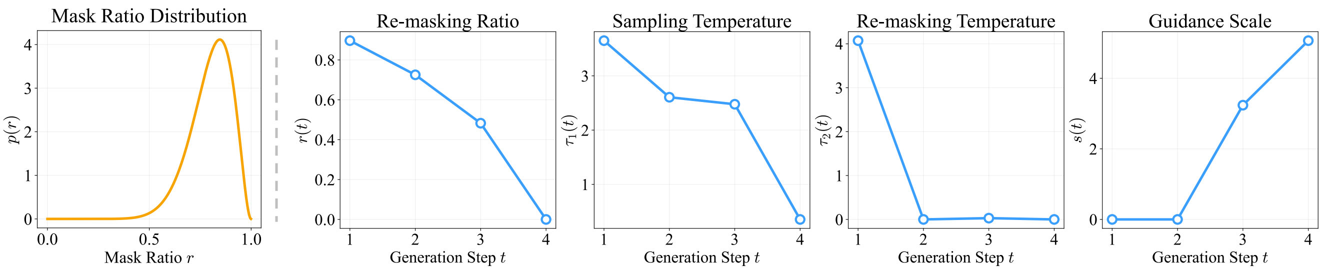
[1] Fan Bao, Chongxuan Li, Yue Cao, and Jun Zhu. All are worth words: a vit backbone for score-based diffusion models. In *CVPR*, 2023. 1

[2] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. In *ICML*, 2023. 1

[3] Rinon Gal, Or Patashnik, Haggai Maron, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. 2022. 1

[4] Jiayi Guo, Chaoqun Du, Jiangshan Wang, Huijuan Huang, Pengfei Wan, and Gao Huang. Assessing a single image in reference-guided image synthesis. In *AAAI*, 2022. 1

[1] https://github.com/baaivision/MUSE-Pytorch



(a) Training Strategy (b) Generation Strategy

Figure 1. **Visualization of optimized training and generation strategies** for AutoNAT-S on ImageNet-256 ($T = 4$).

Figure 2. **Qualitative comparison** on class-conditional image generation.

Figure 3. **Qualitative comparison** on text-to-image generation.

[5] Jiayi Guo, Hayk Manukyan, Chenyu Yang, Chaofei Wang, Levon Khachatryan, Shant Navasardyan, Shiji Song, Humphrey Shi, and Gao Huang. Faceclip: Facial image-to-video translation via a brief text description. *TCSVT*, 2023. 1

[6] Jiayi Guo, Chaofei Wang, You Wu, Eric Zhang, Kai Wang, Xingqian Xu, Humphrey Shi, Gao Huang, and Shiji Song. Zero-shot generative model adaptation via image-specific prompt learning. In *CVPR*, 2023.

[7] Jiayi Guo, Xingqian Xu, Yifan Pu, Zanlin Ni, Chaofei Wang, Manushree Vasu, Shiji Song, Gao Huang, and Humphrey Shi. Smooth diffusion: Crafting smooth latent spaces in diffusion models. *CVPR*, 2024. 1

[8] Yizeng Han, Gao Huang, Shiji Song, Le Yang, Honghui Wang, and Yulin Wang. Dynamic neural networks: A survey. *TPAMI*, 2021. 1

[9] Yizeng Han, Gao Huang, Shiji Song, Le Yang, Yitian Zhang, and Haojun Jiang. Spatially adaptive feature refinement for efficient inference. *TIP*, 2021.

[10] Yizeng Han, Yifan Pu, Zihang Lai, Chaofei Wang, Shiji Song, Junfeng Cao, Wenhui Huang, Chao Deng, and Gao Huang. Learning to weight samples for dynamic early-exiting networks. In *ECCV*, 2022.

[11] Yizeng Han, Zhihang Yuan, Yifan Pu, Chenhao Xue, Shiji Song, Guangyu Sun, and Gao Huang. Latency-aware spatial-wise dynamic networks. *Advances in Neural Information Processing Systems*, 35:36845–36857, 2022.

[12] Yizeng Han, Dongchen Han, Zeyu Liu, Yulin Wang, Xuran Pan, Yifan Pu, Chao Deng, Junlan Feng, Shiji Song, and Gao Huang. Dynamic perceiver for efficient visual recognition. In *ICCV*, 2023.

[13] Yizeng Han, Zeyu Liu, Zhihang Yuan, Yifan Pu, Chaofei Wang, Shiji Song, and Gao Huang. Latency-aware unified dynamic networks for efficient image recognition. *arXiv preprint arXiv:2308.15949*, 2023. 1

[14] Chunming He, Kai Li, Yachao Zhang, Longxiang Tang, Yulun Zhang, Zhenhua Guo, and Xiu Li. Camouflaged object detection with feature decomposition and edge reconstruction. In *CVPR*, 2023. 1

[15] Chunming He, Kai Li, Yachao Zhang, Yulun Zhang, Zhenhua Guo, Xiu Li, Martin Danelljan, and Fisher Yu. Strategic preys make acute predators: Enhancing camouflaged object detectors by generating camouflaged objects. 2024. 1

[16] 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. In *NeurIPS*, 2017. 1

[17] Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking fid: Towards a better evaluation metric for image generation. *arXiv preprint arXiv:2401.09603*, 2023. 1

[18] Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and WOOK SHIN HAN. Draft-and-revise: Effective image gen-

eration with contextual rq-transformer. In *NeurIPS*, 2022. 1

[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 1

[20] Zanlin Ni, Yulin Wang, Jiangwei Yu, Haojun Jiang, Yue Cao, and Gao Huang. Deep incubation: Training large models by divide-and-conquering. In *ICCV*, 2023. 1

[21] Suraj Patil, William Berman, Robin Rombach, and Patrick von Platen. amused: An open muse reproduction. *arXiv preprint arXiv:2401.01808*, 2024. 1

[22] Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016. 1

[23] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018. 1

[24] Yansong Tang, Zanlin Ni, Jiahuan Zhou, Danyang Zhang, Jiwen Lu, Ying Wu, and Jie Zhou. Uncertainty-aware score distribution learning for action quality assessment. In *CVPR*, 2020. 1

[25] Yulin Wang, Xuran Pan, Shiji Song, Hong Zhang, Gao Huang, and Cheng Wu. Implicit semantic data augmentation for deep networks. 2019. 1

[26] Yulin Wang, Gao Huang, Shiji Song, Xuran Pan, Yitong Xia, and Cheng Wu. Regularizing deep networks with semantic data augmentation. *TPAMI*, 2021. 1

[27] Yulin Wang, Rui Huang, Shiji Song, Zeyi Huang, and Gao Huang. Not all images are worth 16x16 words: Dynamic transformers for efficient image recognition. 2021. 1

[28] Yulin Wang, Zanlin Ni, Shiji Song, Le Yang, and Gao Huang. Revisiting locally supervised learning: an alternative to end-to-end training. In *ICLR*, 2021. 1

[29] Yulin Wang, Yang Yue, Rui Lu, Tianjiao Liu, Zhao Zhong, Shiji Song, and Gao Huang. Efficienttrain: Exploring generalized curriculum learning for training visual backbones. In *ICCV*, 2023. 1

[30] Zhuofan Xia, Xuran Pan, Shiji Song, Li Erran Li, and Gao Huang. Vision transformer with deformable attention. In *CVPR*, 2022. 1

[31] Zhuofan Xia, Xuran Pan, Xuan Jin, Yuan He, Hui Xue, Shiji Song, and Gao Huang. Budgeted training for vision transformer. In *ICLR*, 2023. 1

[32] Zhuofan Xia, Xuran Pan, Shiji Song, Li Erran Li, and Gao Huang. Dat++: Spatially dynamic vision transformer with deformable attention. *arXiv preprint arXiv:2309.01430*, 2023. 1

[33] Zhuofan Xia, Dongchen Han, Yizeng Han, Xuran Pan, Shiji Song, and Gao Huang. Gsva: Generalized segmentation via multimodal large language models. In *CVPR*, 2024. 1

[34] Yong Xien Chng, Henry Zheng, Yizeng Han, Xuchong Qiu, and Gao Huang. Mask grounding for referring image segmentation. 2024. 1

[35] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 1