A. Procedural Warm-Up Details

We pretrain the ViT-T/16 configuration used throughout the paper for 15k steps using masked-token prediction. Table 9 summarizes the training hyperparameters.

Setting	Value
Batch size	256
Training steps	15,000
Mask ratio	0.5 (close-only)
Optimizer	AdamW
Learning rate	$2e^{-3}$
Weight decay	0.05
Betas	(0.9, 0.999)
LR schedule	Cosine decay
Warmup steps	1,000

Table 9. Training hyperparameters for the procedural warm-up stage.

For sampling from the k-DYCK and k-DYCK SHUFFLE languages, we draw opening tokens from k_{open} types and closing tokens from k_{close} types, using an opening probability of $p_{\text{open}} = 0.6$ whenever opening is structurally permissible.

In Section 4.3, when varying the language type, we modify only the underlying symbolic generator. When varying the pretraining length, we train a separate model for each duration by adjusting the number of training steps accordingly.

B. Image-Based Training Details

For the experiments in Section 4.1 (Table 2), we follow the standard ViT training setup used in [56], including AdamW optimization, cosine learning-rate decay, linear warmup, RandAugment, Mixup, CutMix, and label smoothing. Across all datasets, we use a base learning rate of $2e^{-3}$ and train for 300 epochs with 50 warmup epochs. We use a batch size of 512 for all models, except for ViT-B on IMAGENET-1K, where we use a batch size of 4096.

For the experiments in Section 4.2 (Table 3), we further fine-tune the IMAGENET-1K ViT-B checkpoints on the smaller datasets. We use the same training setup as above (with batch size 512), but train for 50 epochs with 5 warmup epochs and a base learning rate of $5e^{-4}$.

C. Additional Results

Method	TINY-IMAGENET	FOOD-101	CIFAR-10	CIFAR-100	STL-10
Default init.	55.42	74.52	91.29	68.52	60.52
FractalDB warm-up	55.17	74.25	88.98	64.61	58.62
▶ FractalDB warm-up + emb.	55.64	75.99	90.44	67.35	65.62
Procedural warm-up (ours)	**58.20**	**79.47**	**92.81**	**71.98**	**66.48**

Table 10. Comparison with warm-up on FractalDB when retaining pretrained embeddings. Unlike the main setting where embeddings are reset for consistency across methods, this variant preserves FractalDB's patch embeddings. Our method still outperforms this baseline.

	k=16	k=32	k=64	k=80
Procedural warm-up	69.05	69.78	**71.98**	71.83

Table 11. CIFAR-100 top-1 accuracy (%) with varying vocabulary size (k). All settings outperform random initialization, with peak performance at an intermediate vocabulary size.

Method	TINY-IMAGENET	C100
Default init.	55.42 ± 0.66	68.52 ± 0.27
Mimetic init.	57.20 ± 0.62	70.72 ± 0.39
FractalDB warm-up	55.17 ± 0.33	64.61 ± 0.51
Procedural warm-up (ours)	$\mathbf{58.20 \pm 0.22}$	$\mathbf{71.98 \pm 0.74}$

Table 12. Mean $\pm$ standard deviation over 3 random seeds on representative datasets of differing scale, showing consistent gains.