A. Sensitive Study

We conduct the sensitive studies on the hyperparameters of GrowCLIP, including: (i) α, the model parameter coefficient in growth architecture selection, (ii) β, the proportion of old model in parameter inheriting, (iii) γ, the proportion of new model in parameter inheriting. Note that all experiments are evaluated on zero-shot image classification of ImageNet under the growth step 2. As shown in Table I, if the model parameter coefficient α is larger, the smaller model will be selected. And the best proportion of the old model and the new one in parameter inheriting is $\beta = 0.3$, $\gamma = 0.001$.

α	5	**0.5**	0.05
Para.	30.1M	116.6M	129.2M
Acc.	23.2	25.7	25.7
β	0.1	**0.3**	0.5
Para.	109.5M	116.6M	37.2M
Acc.	21.7	25.7	22.7
γ	10^{-6}	$\mathbf{10^{-3}}$	1
Para.	116.6M	116.6M	116.6M
Acc.	25.3	25.7	25.5

Table I. Sensitive Study on hyperparameters. (Para.: parameter of the model. Acc.: top-1 accuracy (%) of zero-shot image classification on ImageNet.)

We conduct sensitive studies on the number of growth steps and the result is shown in Figure A. We split CC12M into 6 steps. Note that the size of the marker represents the size of the model. Our model grows adaptively as well when the data grows.

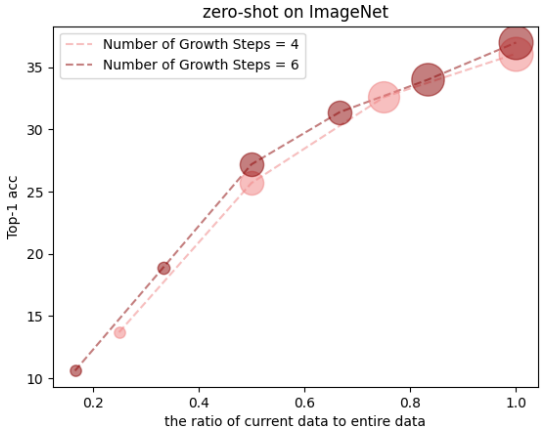

Figure A. The performance of model with CC12M split into different subsets. The size of the marker represents the size of the model.

B. Analysis

Is Growth Architecture Selection (GAS) effective? Different from One-shot NAS methods, GrowCLIP only takes 2 epochs to train the supernet and selects the best architecture, which benefits from our parameter inheriting with momentum (PIM). To test the sensitivity of the number of supernet training epochs, we calculate the rank of defined metric (Eq. (7) of main paper) of selected subnet for illustration.

Epochs	**2**	6	10	14	18	22	26
Rank	1	2	1	1	1	1	1

Table II. The rank of the grown model we chose with different numbers of supernet training epochs.

As shown in Table II, the architecture chosen after 2 epochs supernet training is almost the same as those chosen after longer supernet training, demonstrating that 2 epochs are enough to determine the optimal subnet in our scenario. (Rank 1 for epoch 26 means that the selected architecture also has the best performance among the candidate architectures after 26 epochs supernet training.)

Model	Para.	Acc.
GrowCLIP-S2	116.6M	**25.7**
Smallest model in growth space	30.0M	21.3
Biggest model in growth space	129.2M	**25.7**
Ramdom model 1 in growth space	109.5M	23.3
Ramdom model 2 in growth space	80.3M	25.3
Ramdom model 3 in growth space	129.1M	23.7

Table III. Effectiveness of Growth Architecture Selection (GAS). (Acc.: top-1 accuracy (%) of zero-shot image classification on ImageNet.)

To further prove the effectiveness of GAS, we compared the model selected by GAS and others in the growth space in Table III. The result shows that the selected model with GrowCLIP obtains the best performance.

	Para.	Acc.
CLIP-ViT-B/16*-S	127.2M	28.2
CLIP-ViT-B/16*	212.2M	29.8
CLIP-ViT-B/16*-0.5D	127.2M	27.6

Table IV. Analysis of shared encoder. CLIP-ViT-B/16* sets the width of text transformer as 768 to unify the dimension of both encoders. CLIP-ViT-B/16*-S only uses shared encoder. CLIP-ViT-B/16*-0.5D is half as deep as CLIP-ViT-B/16*

	Image encoder			Text encoder		Shared encoder
	Transformer blocks	Transformer heads	Convolutional layers	Transformer blocks	Transformer heads	Transformer blocks
GrowCLIP-S1	6	6	0	6	4	0
GrowCLIP-S2	10	10	2	6	8	4
GrowCLIP-S3	14	10	2	10	8	8
GrowCLIP-S4	18	10	4	10	8	8

Table V. The selected architecture of GrowCLIP in each growth step.

What can shared encoder bring? The shared encoder can bring better trade off between the performance and the model size. As shown in Table IV, compared with CLIP-ViT-B/16*, CLIP-ViT-B/16*-S has worse performance but uses fewer parameters. Compared with CLIP-ViT-B/16*-0.5D, which has the same parameters as CLIP-ViT-B/16*-S and the same model structure as CLIP-ViT-B/16*, CLIP-ViT-B/16*-S has better performance.

C. The Architecture of GrowCLIP in each Growth Step

The selected architecture in each step of GrowCLIP is shown in Table V. We can observe that the image encoder, text encoder, and shared encoder all are enlarged with the incoming data at the growth step 2, since the architecture is quite small relative to the size of the data. And the image encoder is enlarged, while the architecture of the text encoder and shared encoder are unchanged at the growth step 4. The reason behind the architecture bias may be that the bottleneck of performance is the image encoder part at the current setting. In other words, the scale of the text encoder and shared encoder are enough for current data size and the algorithm prefers to increase the size of the image encoder given a certain amount quota of parameters.