A. Details of Experimental Setup

Our implementation follows CLIP [32], OpenCLIP [15], FLIP [22] and CLIPA [20]. In this section, we present the details of our experimental setup.

Dataset Data set statistics are summarized in Tab. 7. For the majority of our experiments, we pre-train the models on Conceptual Captions 3M (CC3M) [36], Conceptual 12M (CC12M) [3]. These datasets have been used by OpenCLIP, DeCLIP, SLIP, A-CLIP and BLIP [15, 18, 21, 30, 43] To reproduce CLIPA, we use LAION-400M [34], which is used by CLIPA. LAION-400M is approximately 133 times larger than CC3M and 32 times larger than CC12M, indicating that it contains a broader range of concepts and offers greater diversity compared to CC3M and CC12M.

Note that the difference in caption length with have an impact on masking performance, although we do not directly measure the difference in this paper.

Architecture For the image encoder, we use ViT-B/16 (86M parameters) [9] architectures with global average pooling and learnable positional embeddings. For the text encoder, we use a Transformer-based model [40] and byte-pair encoding with a 49K token vocabulary. Additionally, we run one experiment with a larger image encoder (VIT-L/16) with 303M parameters and report the results in the Appendix C of supplementary material. We chose these encoders because they are the largest ones available to us, given our current resource constraints. The input image size is 224 for all datasets. When using full text for training, the maximum context length is 32. Zero-padding is applied to input text that is shorter than the maximum token length of the model.

We trained the models using 8 RTX A5000 GPUs with the same settings to ensure consistent conditions across all models. We experiment with different text token input sizes, namely, 32, 16, 8, 6, and 4 text tokens. As the input size gets smaller, we can increase the batch size, maximizing computational memory usage. For CC3M and CC12M, the batch sizes corresponding to the input sizes are: 664, 832, 896, 928, and 944. Note that CLIP (and therefore CLIPA and CLIPF) is trained using contrastive learning, which benefits from larger batch sizes. Note also that although we used different text masking with the same settings and batch sizes, syntax masking was slower than the other text masking strategies when conducting the experiments because POS processing for each word is time-consuming.

We pre-train the model using the InfoNCE loss [39] with a learnable temperature parameter τ [5, 32]. To classify images, we calculate the cosine similarity between the image and text embeddings.

Training Following CLIP, OpenCLIP [15, 32], we pre-train our model for 30 epochs on the CC3M and CC12M datasets. For the LAION-400M dataset, we pre-train the model for 16 epochs, extending CLIPA's experiments, which were conducted for 6 epochs on the same dataset. Details of the pre-training configuration are given in Tab. 8.

Dataset	Samples	Total words	Caption length
CC3M	2.72×10^6	2.80×10^7	10.30 ± 4.72
CC12M	9.30×10^6	2.06×10^8	22.15 ± 17.20
LIAON-400M	2.98×10^8	3.71×10^9	12.51 ± 15.82

Table 7. Dataset statistics for pre-training datasets. Caption length refers to the number of words in the text.

Config	**Pre-training**	**Fine-tuning**
optimizer	AdamW [27]	AdamW [27]
learning rate	1e-3	1e-5
weight decay	0.2	0.2
optimizer momentum	$\beta_1, \beta_2 = 0.9, 0.98$ [4]	$\beta_1, \beta_2 = 0.9, 0.98$ [4]
learning rate schedule	cosine decay [26]	cosine decay [26]
warmup steps	10k	10%
epoch	30	1
t	10^{-6}	—
τ	0.07	—
numerical precision	amp	amp
RandomResizedCrop	(50, 100)	(50, 100)

Table 8. Details of the pre-training and fine-tuning setups.

Image masking	Text masking	Image tokens	Text tokens	Total	Percentage
0%	0.00%	196	32	228	100%
75%	0.00%	49	32	81	35.53%
	50.00%	49	16	65	28.51%
	75.00%	49	8	57	25.00%
	81.25%	49	6	55	24.12%
	87.50%	49	4	53	23.25%

Table 9. The number of image and text tokens that are processed during pre-training (CC3M and CC12M).

Similar to FLIP, to speed up training, we apply 75% image masking to the image encoder as the baseline model. As a result, the speedup is about 4× compared to training without image masking, while the reduction in the zero-shot performance of ImageNet-1K classification remains within reasonable bounds.

During text masking, we reduce the number of tokens from 32 to 16, 8, 6, and 4. To measure the training speed of CLIP, CLIPA, and CLIPF, we compare the number of text tokens processed by each model. As shown in Tab. 9, the number of text tokens processed by each model during pre-training is calculated based on different image and text masking ratios. When we pre-trained the model using image masking, the total number of tokens is 81 and the percentage of text tokens is 39.5%. When we apply 50% text masking, the total number of tokens is 65. Compared to training without text masking, this results in a speed increase of approximately 20%. Moreover, when we apply 87.5% text masking, the total number of tokens is 53, resulting in a speed increase of approximately 34% compared

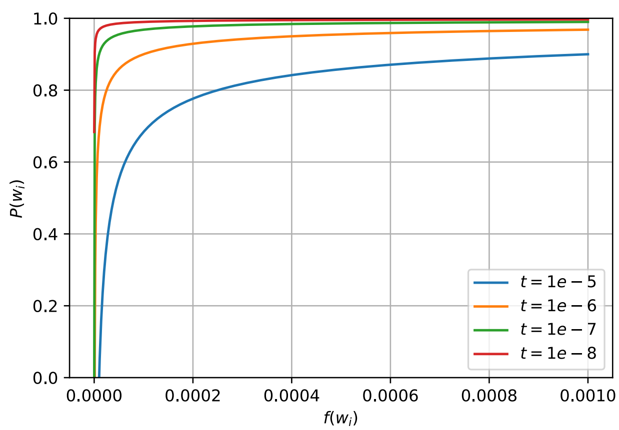

Figure 5. The curve of Equation 2. The x-axis is the word frequency $f(w_i)$, and the y-axis is the $P(w_i)$. The value of t of Equation 2 is set to 10^{-5}, 10^{-6}, 10^{-7}, 10^{-8}.

Method	Resulting Text
original	Walk of the happy young couple and Siberian dog. The handsome man is hugging the smiling red head girl
truncation	walk of the happy young couple
random	the happy and the is the
block	couple and siberian dog . the
syntax	walk dog man smiling head girl
CLIPF	siberian handsome man hugging smiling red

Table 10. Example from the CC12M dataset illustrating the effect of various text masking strategies, reducing text length to 6 words. The original caption is in the first row, followed by masked variants. Parameter t in Equation 2 is set to 10^{-6}.

to training without text masking.

Fine-tuning Following FLIP, and CLIPA, in order to bridge the distribution gap between pre-training and evaluation, we fine-tuned the model without images and text masking. Note that, fine-tuning the models without masking in masking work is focused on reducing the distribution gap and applied sparingly to avoid undoing the tradeoff. The details of the fine-tuning configuration are provided in Tab. 8.

Evaluation setup Following CLIP [32], FLIP [22], and CLIPA [20], several classification benchmarks were used. Among these benchmarks is ImageNet-1K, a widely recognized dataset in computer vision. It is frequently used for image classification and VLM tasks and comprises 50K validation samples across 1K different classes. We filled each class into the templates provided by CLIP [32] to calculate the average of the text embeddings. We use the same evaluation settings as CLIP [32] to evaluate the other downstream datasets.

Word	Probability
walk	0.926171
of	0.992838
the	0.995064
happy	0.951695
young	0.957311
couple	0.941174
and	0.991695
siberian	0.750920
dog	0.960531
.	0.993678

Table 11. The probability of masking words in the text is calculated using the formula provided in Equation 2. The example is selected from CC12M [3]. The value of t of Equation 2 is set to 10^{-6}.

B. Text Masking Analysis

B.1. Text Masking Cases

As illustrated in Fig. 5, there is a clear relationship between word masking probability and word frequency. Frequent words have a higher masking probability compared to infrequent ones. Additionally, a smaller threshold t leads to a smaller difference between the masking probabilities of frequent and infrequent words. Therefore, it was necessary to choose a relatively larger threshold to ensure that both frequent and infrequent words are effectively masked.

Tab. 10 presents the potential text resulting from the text masking technique. Since truncation is fixed in each epoch, it may result in the loss of important information at the end of the text. Certain words tend to appear in specific positions within the text; for example, “the” and “a” are most likely to be the first words of a sentence. As shown in Fig. 6, truncation retains more occurrences of “the” and “a” than other text masking strategies. In contrast, random and block strategies can generate different texts in each epoch for text data augmentation, but they may retain some words with a high frequency. Syntax masking retains the nouns in the text and remains the same in each epoch. However, this strategy may cause the model to overfit on the frequency of noun words. In contrast, CLIPF varies the text in each epoch, as words may be retained or removed according to their frequency. This strategy serves two primary purposes: it enhances text diversity and reduces the risk of overfitting to frequent words. Another advantage of CLIPF is that it can remove frequent prepositions that are less directly relevant to the objects in the image, such as “a,” “in,” “of,” and others. This helps the model focus on the most helpful aspects of the content. Tab. 11 shows the masking probabil-

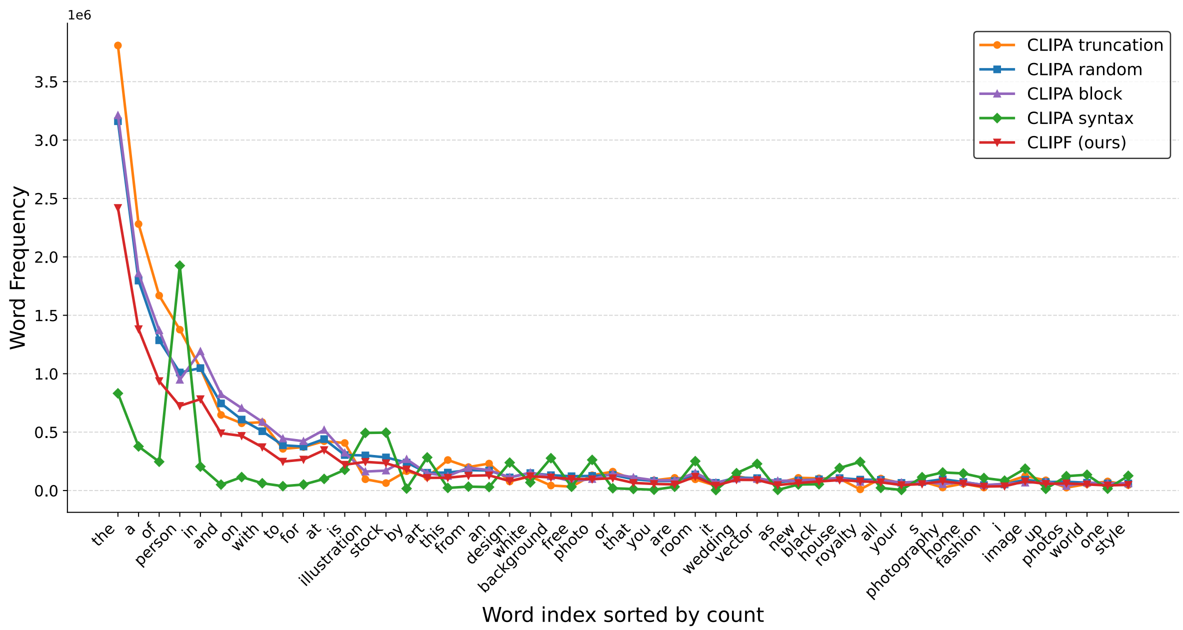

Figure 6. **The complete x-axis label for Fig. 3. in Sec. 3** We set the text length after text masking to 6. The x-axis represents the word index, which is sorted by counts of the original data, and the y-axis shows the word frequency. The dataset used is CC12M and the value of t of Equation 2 is set to 10^{-6}. We remove special characters from the vocabulary.

Masking	**NN Count**	**JJ**	**VB**	**OTHER**	**NN (%)**	**JJ (%)**	**VB (%)**	**OTHER (%)**	**Total**
Before masking	103,469,117	10,245,828	10,649,943	81,351,966	50.30%	4.98%	5.18%	39.55%	205,716,854
Truncation	28,628,129	2,859,462	3,272,401	20,585,364	51.72%	5.17%	5.91%	37.20%	55,345,356
Random	28,334,735	2,666,847	2,790,600	21,550,517	51.21%	4.82%	5.04%	38.93%	55,342,699
Block	27,314,723	2,624,594	2,897,434	22,510,031	49.37%	4.74%	5.24%	40.66%	55,346,782
Syntax	48,989,317	1,483,666	1,618,088	3,245,892	88.53%	2.69%	2.92%	5.87%	55,336,963
SW-CLIP	34,645,367	3,346,676	4,073,825	6,062,345	71.98%	6.95%	8.47%	12.60%	48,128,213
CLIPF	33,516,439	2,803,409	3,473,119	15,666,310	60.43%	5.06%	6.20%	28.24%	55,459,277

Table 12. Distribution of syntax counts and percentages before and after applying text masking. The dataset is CC12M and we retain 6 words for each text after applying text masking.

ities for certain words. High-frequency words such as "of" "the" "and" and "." are highly likely to be masked from the text.

B.2. Word Category Distribution Across Different Text Masking Strategies

As shown in Tab. 12, we present the word type counts corresponding to Tab. 1 in the main paper. After applying text masking strategies such as truncation, random, block, syntax, and CLIPF, the word counts remain similar. However, SW-CLIP does not maximize the use of the input slots, utilizing only 85% of the words compared to other text masking strategies. Additionally, SW-CLIP masks a high percentage of other types of words, which may impact the model's zero-shot ability.

B.3. Word Distribution Analysis for CLIPF

In addition, we analyzed frequency text masking strategies that varied according to the number of words used, as shown in Fig. 7. As the number of words decreased from 16 to 8 and then to 6, more frequent words were masked. However, reducing the number to 4 words led to a smaller vocabulary size, resulting in the loss of some important information. Consequently, the performance of the model pre-trained with 4 words was substantially lower compared to the model pre-trained with 6 words. We recommend setting the number of text words in a frequency-based text masking strategy to strike a balance between frequent and infrequent words and to maintain a larger vocabulary. Based on our experiments, the optimal number of text masking was found to be approximately 40-60% of the average length of the original text. This configuration helps achieve a balanced word

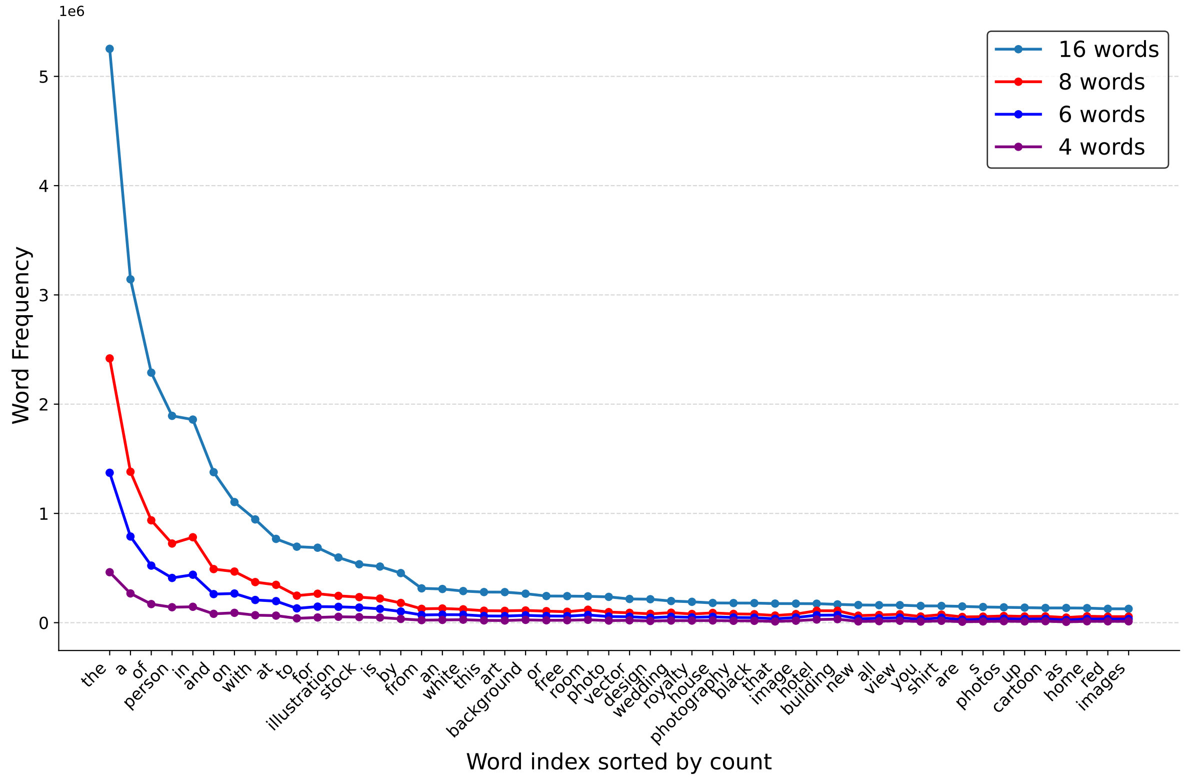

Figure 7. Frequency-based text masking strategies vary according to the number of text words used during pre-training. The dataset used is CC12M and the value of t of Equation 2 is set to 10^{-6}.

Models	Masking	Image Tokens	Text Tokens	ViT-B/16					
				IN-A	IN-O	IN-R	IN-S	IN-V2	ON
CLIP	✗	197	32	8.97	**37.85**	49.11	25.70	31.48	24.20
CLIPF	frequency	98	8	**10.37**	37.75	**52.33**	**28.52**	**35.62**	24.00

Table 13. Zero-shot robustness evaluation. Comparison of the zero-shot accuracy performance of CLIP and CLIPF on various datasets when using more image tokens. The models are pre-trained on **CC12M** [3] for 30 epochs with image masking (**50%**) to speed up training and fine-tune the model an additional epoch without image and text masking.

Models	Masking	Image Tokens	Text Tokens	Text Retrieval		Image Retrieval	
				Flickr30k R@1	COCO R@1	Flickr30k R@1	COCO R@1
CLIP	✗	197	32	62.62	35.54	45.42	24.22
CLIPF	frequency	98	8	**63.11**	**37.14**	**46.84**	**24.89**

Table 14. Zero-shot Image-Text retrieval evaluation. We evaluated CLIP, CLIPF image-text retrieval performance on the COCO and Flickr30k datasets when using more image tokens. The models are pre-trained on **CC12M** [3] for 30 epochs with image masking (**50%**) to speed up training and fine-tune the model an additional epoch without image and text masking.

distribution, which is beneficial for pre-training VLMs.

C. More Results

More detailed results are presented in this section, including the results and learn curve of image-text retrieval.

Models	Masking	Image Tokens	Text Tokens	Text Retrieval						Image Retrieval					
				Flickr30k			COCO			Flickr30k			COCO		
				R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
CLIP	✗	197	32	**62.62**	**86.00**	**91.81**	**35.54**	**62.38**	**74.08**	**45.42**	**72.56**	**81.50**	**24.22**	**48.42**	**60.42**
FLIP	✗	49	32	54.73	80.37	87.97	29.34	56.08	67.00	38.62	66.09	75.40	20.88	43.22	54.78
FLIP	truncation		16	44.67	73.08	81.85	25.54	51.90	64.64	34.99	61.05	70.85	19.64	41.91	53.51
	random			**58.48**	**84.62**	**90.53**	**32.36**	**58.76**	69.98	43.61	70.67	80.04	**22.81**	46.00	57.71
CLIPA	block			56.51	81.26	89.05	30.82	58.32	**70.38**	**44.06**	**71.20**	**80.10**	22.66	**46.24**	**58.06**
	syntax			54.54	81.07	88.36	29.60	56.52	68.54	41.16	68.32	77.51	21.40	44.82	56.56
CLIPF	frequency			57.89	**84.62**	90.04	31.52	58.38	70.30	42.72	69.61	78.60	22.57	46.15	57.95
FLIP	truncation		8	30.47	59.47	70.51	16.92	38.62	51.34	23.96	47.29	57.85	12.78	31.66	42.92
	random			58.58	84.52	91.81	**36.24**	62.16	72.90	43.79	70.89	80.14	23.16	46.74	58.86
CLIPA	block			**60.06**	85.01	**92.11**	35.88	**62.58**	**73.74**	**45.98**	**73.23**	**82.49**	**24.65**	**48.81**	**60.70**
	syntax			50.30	78.30	86.98	29.64	55.42	67.18	38.46	66.77	77.36	20.28	43.78	55.40
CLIPF	frequency	49		58.68	**85.10**	91.51	34.74	61.38	71.88	44.57	72.58	81.92	23.30	47.50	59.24
FLIP	truncation		6	32.25	61.05	72.98	16.18	39.52	52.48	22.39	47.46	59.39	11.93	30.03	40.86
	random			55.23	**82.54**	**89.25**	32.58	57.54	68.68	**42.47**	**70.12**	**79.43**	21.34	44.79	56.26
CLIPA	block			54.44	79.68	88.66	**33.12**	**58.96**	**70.64**	41.68	69.13	79.33	**21.73**	**45.23**	**57.29**
	syntax			46.15	76.04	84.81	26.78	52.56	64.62	34.08	61.76	73.25	17.99	40.22	52.05
CLIPF	frequency			**56.02**	82.05	88.56	32.32	58.58	70.00	41.05	69.17	78.88	21.28	44.55	56.05
FLIP	truncation		4	30.97	56.51	67.06	17.44	38.48	49.76	21.09	43.77	55.70	11.07	27.98	38.54
	random			41.62	69.53	80.18	24.14	48.32	60.14	30.26	56.92	68.42	15.63	35.72	47.30
CLIPA	block			40.83	69.03	79.09	23.56	47.94	59.20	30.77	57.85	69.29	14.98	35.64	47.35
	syntax			38.66	65.88	75.74	21.68	44.38	56.60	26.08	52.80	64.32	14.14	33.71	44.98
CLIPF	frequency			**45.07**	**72.49**	**81.95**	**25.60**	**49.98**	**61.96**	**31.72**	**59.35**	**71.66**	**15.96**	**36.58**	**48.32**

Table 15. **Zero-shot Image-Text Retrieval,** we evaluated CLIP, FLIP, and CLIPF image-text retrieval performance on COCO and Flickr30k datasets. The backbone of the image encoder is ViT-B/16, and the model is pre-trained on CC12M for 30 epochs with image masking (75%) to speed up training and fine-tune the model additional epoch without image and text masking.

Models	Datasets	Image Tokens	Text Tokens	ViT-B/16	
				pre-train	fine-tune
SW-CLIP	CC3M	49	6	**14.9**	16.8
CLIPF				14.4	**18.2**
SW-CLIP	CC12M	49	8	35.9	38.5
CLIPF				**36.6**	**39.3**

Table 16. Comparison of the zero-shot accuracy performance of SW-CLIP and CLIPF on ImageNet-1k. The models are pre-trained on **CC12M** [3] for 30 epochs with image masking (75%) to speed up training and fine-tune the model an additional epoch without image and text masking.

C.1. More image tokens

As shown in Tab. 13 and Tab. 14, CLIPF achieves better zero-shot robustness in 4 out of 6 datasets and image–text retrieval performance in both Flick30k and COCO datasets than the original CLIP, even though it only uses 50% of the image tokens and 25% of the text tokens.

C.2. The details of Image-text Retrieval

In Tab. 15, we show more details of Zero-shot Image-Text Retrieval on COCO and Flickr30k datasets.

C.3. Learn Curved for DTD and OxfordPets Datasets

The learning curves for the DTD [6] and OxfordPets [31] datasets are presented because not all class names in these datasets are nouns. As shown in Fig. 8, the performance of syntax masking on the DTD dataset is consistently lower than that of random, block, and frequency masking across all epochs, rather than initially outperforming them and declining at later stages in the ImageNet dataset, as shown in Fig. 9. Moreover, the performance of syntax masking is very poor and almost close to truncation masking. OxfordPets is a dataset containing incomplete noun class names. As shown in Fig. 10, in the early training epochs, syntax masking performs similarly to frequency masking and still outperforms truncation, block, and random masking. However, in the later stages of training, syntax masking performs much worse than the other text masking strategies.

C.4. Comparison with SW-CLIP

In this section, we provide a comparison of CLIPF with SW-CLIP [23], which also uses frequency-based sampling but imposes a threshold on the frequency score. The effect of this threshold is that input tokens will go unused

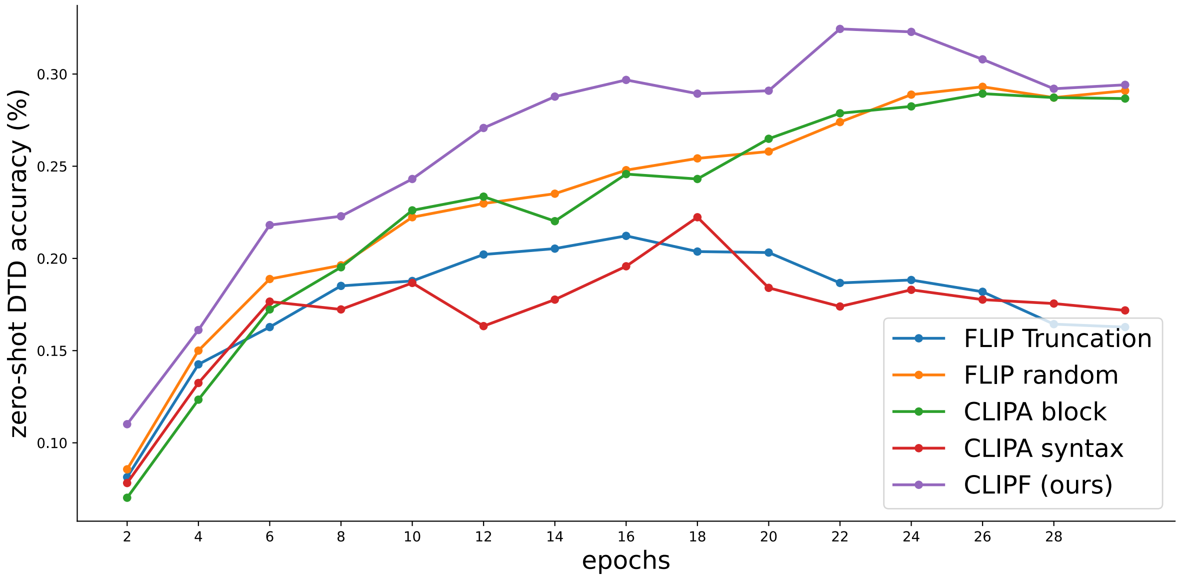

Figure 8. Zero-shot classification accuracy on DTD dataset over training epochs for CLIPF and CLIPA strategies. The backbone of the image encoder is ViT-B/16, and the model is pre-trained on CC12M for 30 epochs with image and text masking (75%) to speed up training and fine-tune the model additional epoch without image and text masking.

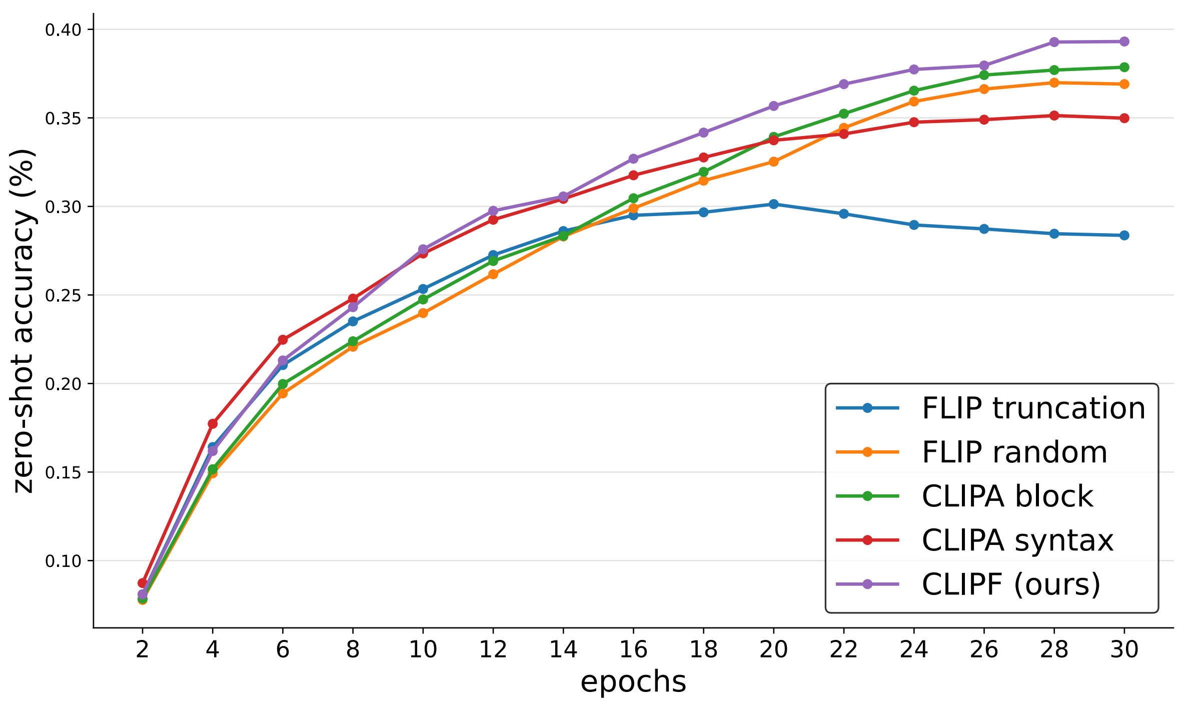

Figure 9. Zero-shot classification accuracy on ImageNet-1K dataset over training epochs for CLIPF and CLIPA strategies **after fine-tuning**. The backbone of the image encoder is ViT-B/16, and the model is pre-trained on CC12M for 30 epochs with image and text masking (75%) to speed up training and fine-tune the model additional epoch without image and text masking.

if not enough words in a training text surpass the threshold. In contrast, CLIPF calculates the word frequency by the threshold to prioritize words and maximize the input slots. For instance, using CC3M as an example, the average length of text before masking is 10.31 ± 4.7, and after SW-CLIPF masking, it is 4.26 ± 2.6 [23]. It is clear that the use

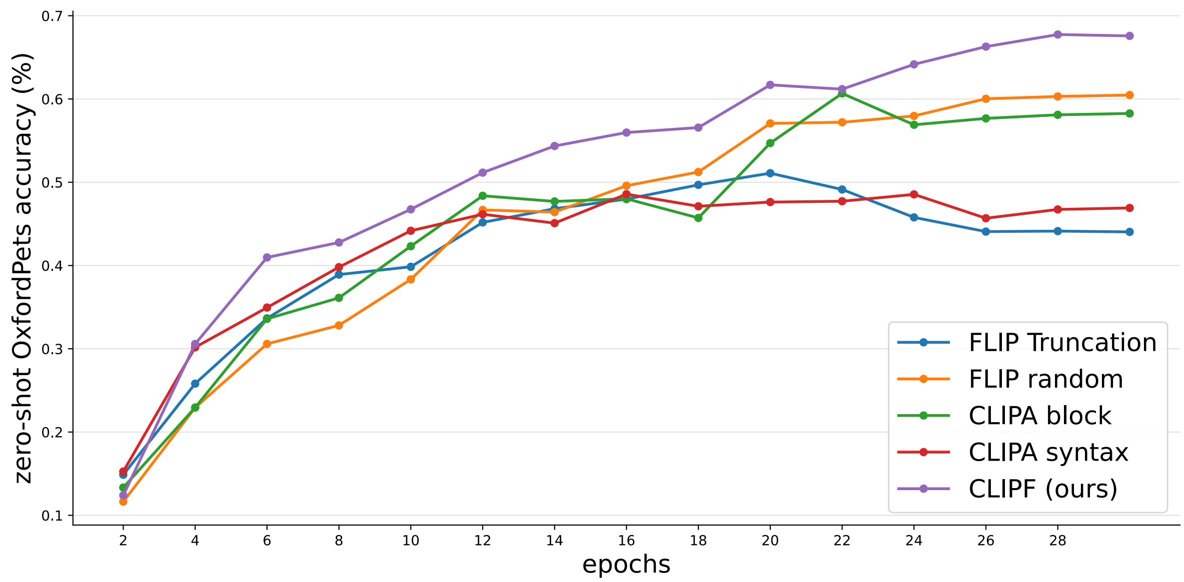

Figure 10. Zero-shot classification accuracy on OxfordPets dataset over training epochs for CLIPF and CLIPA strategies. The backbone of the image encoder is ViT-B/16, and the model is pre-trained on CC12M for 30 epochs with image and text masking (75%) to speed up training and fine-tune the model additional epoch without image and text masking.

Models	Masking	Image Tokens	Text Tokens	ViT-L/16 pre-train	ViT-L/16 fine-tune
FLIP	truncation	49	8	28.4	31.7
	random			36.0	38.3
CLIPA	block			36.8	39.5
	syntax			32.4	37.5
CLIPF	frequency			**38.0**	**40.2**

Table 17. Comparison of the ImageNet-1k zero-shot accuracy performance of **ViT-L/16** with different text masking strategies. The models are pre-trained on **CC12M** [3] for 30 epochs with image masking (75%) to speed up training and fine-tune the model an additional epoch without image and text masking.

of the input slots is not maximized when the input length of text tokens is set to 6 or longer. For a fair comparison with CLIPF, we pre-train SW-CLIP using the same setup: 75% image masking and 81.25% text masking. Subsequently, we fine-tuned both models with an additional epoch without any image or text masking. The results in Tab. 16 show that after fine-tuning CLIPF outperforms SW-CLIP.

C.5. Large Image Encoder Architecture

We carried out an extra experiment which is apply different text masking strategies on larger architectures which is ViT-L/16. The behavior of different architectures is very consistent when applying different text masking strategies, as shown in Tab. 17.

C.6. Applying text masking on SigLIP

We also applied the text masking strategies to SigLIP; the results are presented in Tab. 18. CLIPF still achieves superior performance compared to other text masking strategies and without text masking. This further supports our conclusion that frequency-based masking is the more effective strategy.

D. Ablation

D.1. Threshold Analysis

To investigate the impact of the threshold t in Equation 2 on model performance, we pre-trained models with varying thresholds. As shown in Tab. 19, thresholds of $1e-5$, $1e-6$, and $1e-7$ yield comparable results, all outperforming the other text masking strategies. This indicates that CLIPF is relatively insensitive to small variations in

Models	Masking	Image Tokens	Text Tokens	ViT-B/16 pre-train	ViT-B/16 fine-tune
SigLIP	✗	197	32	39.3	✗
SigLIP	✗	49	32	28.4	29.6
SigLIP	truncation	49	8	20.4	27.1
	random			29.7	31.7
	block			30.5	32.7
	syntax			25.9	31.6
SigLIP	frequency			32.4	34.8

Table 18. **Comparison of the ImageNet-1k zero-shot accuracy performance of SigLIP with different text masking strategies.** The models are pre-trained on **CC12M** [3] for 30 epochs with image masking (75%) to speed up training and fine-tune the model an additional epoch without image and text masking.

model	Image tokens	Text tokens	Threshold	ViT-B/16 pre-train	ViT-B/16 fine-tune
CLIPF	49	8	1e-5	35.6	38.4
			1e-6	**36.6**	**39.3**
			1e-7	36.1	38.6

Table 19. We pre-train CLIPF on the CC12M dataset across different thresholds in Equation 2. The models are pre-trained on **CC12M** [3] for 30 epochs with image masking (75%) to speed up training and fine-tune the model an additional epoch without image and text masking.

Models	Masking	Image Tokens	Text Tokens	ViT-B/16 pre-train	ViT-B/16 fine-tune
CLIPF	frequency-token	49	8	36.0	38.0
	frequency-word	49	8	36.6	**39.3**

Table 20. **Comparison of CLIPF pre-trained with token masking and word masking on ImageNet-1K for zero-shot classification.** The models are pre-trained on **CC12M** [3] for 30 epochs with image masking (75%) to speed up training and fine-tune the model an additional epoch without image and text masking.

Model	Masking	Time (minutes)
FLIP	trancation	39.81
	random	48.43
CLIPA	block	40.81
	syntax	217.76
CLIPF	frequency	84.64

Table 21. Processing times for the CC12M dataset by using different masking strategies.

thresholds, which are intended to maintain the word masking probability within the range of 0 to 1. However, using too small a threshold reduces the differences in text masking probabilities; therefore, we do not recommend using an excessively low threshold.

D.2. Word and Token Analysis

Since Open_CLIP encodes text using byte-pair encoding (BPE) [15], some words are represented by multiple tokens. Therefore, in this experiment, we calculate the masking probability based on token frequency rather than whole words. Tab. 20 compares the performance of both approaches. Although token frequency masking may disrupt the original word structure and thus achieve lower performance compared to word frequency masking, it still slightly outperforms other text masking strategies.

E. Computational overhead

We carried out a simple measurement of how long it takes to tokenize CC12M using one thread. As shown in Tab. 21, syntax masking required 2.5 times the number of minutes of CLIPF (218 vs. 85) and about 5.4 times that of truncation, random, and block masking.