CLIPS: An Enhanced CLIP Framework for Learning with Synthetic Captions

Anonymous ACL submission

Abstract

Previous works show that noisy, web-crawled image-text pairs may limit vision-language pretraining like CLIP and propose learning with synthetic captions as a promising alternative. Our work continues this effort, introducing two simple yet effective designs to better leverage richly described synthetic captions. Firstly, by observing a strong inverse effect in learning with synthetic captions—the short synthetic captions can generally lead to MUCH higher performance than full-length ones—we therefore fed only partial synthetic captions to the text encoder. Secondly, we incorporate an autoregressive captioner to mimic the recaptioning process—by conditioning on the paired image input and web-crawled text description, the captioner learns to predict the full-length synthetic caption generated by advanced MLLMs. Experiments show that our framework significantly improves zero-shot performance in cross-modal retrieval tasks, setting new SOTA results on MSCOCO and Flickr30K. Moreover, such trained vision encoders can enhance the visual capability of LLaVA, showing strong improvements on a range of MLLM benchmarks.

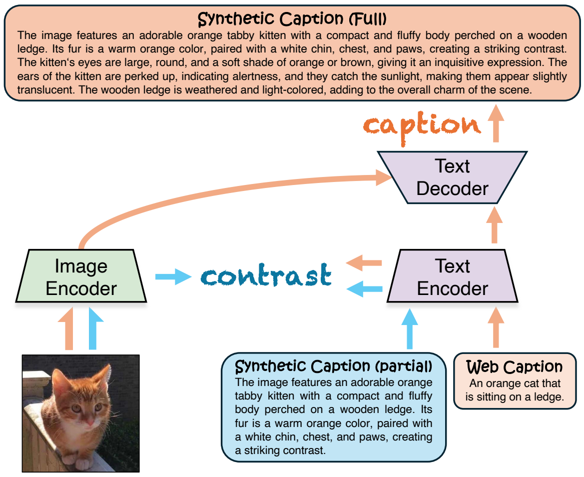


Figure 1: The pipeline of our proposed CLIPS. We introduce two simple yet effective designs—1) only a subpart of the synthetic caption is used in contrastive learning and 2) a captioner to predict the full synthetic caption based on the web-crawled caption and the image—to better leverage synthetic captions. Our method registers new SOTA results on MSCOCO, achieving 76.4% in text retrieval and 57.2% in image retrieval.

1 Introduction

The availability of large-scale image-text datasets, such as LAION (Schuhmann et al., 2022) and DataComp (Gadre et al., 2024), has been a key driver of the rapid development of vision-language models in recent years (Radford et al., 2021; Yu et al., 2022; Li et al., 2022; Bai et al., 2023; Chen et al., 2024c; Liu et al., 2024a). Nonetheless, these web-crawled datasets are generally noisy and not-high-quality (*e.g.*, image-text pairs could be mismatched), which potentially limits further performance improvements (Jia et al., 2021). Consequently, many works seek to improve the dataset quality by re-generating paired textual descriptions using multimodal large language models (MLLM) and incorporating these synthetic captions into training (Fan et al., 2024; Nguyen et al., 2024; Lai et al., 2025).

A straightforward approach to learning with synthetic captions is to simply replace the raw, web-crawled captions with rewritten ones (Nguyen et al., 2024; Li et al., 2024a; Lai et al., 2025). As demonstrated in VeCLIP (Lai et al., 2025) and Recap-DataComp-1B (Li et al., 2024a), (partially) substituting the original captions with those generated by advanced MLLMs during CLIP training can substantially enhance the models' capabilities, especially in cross-modal retrieval tasks. Building upon this line of research, our work continues the exploration of training with synthetic captions but focuses on enhancing the vision-language pretraining framework to better leverage these captions, similar to prior efforts (Fan et al., 2024; Lai et al., 2025; Zheng et al., 2025).

Because synthetic captions are typically highly descriptive—much longer and containing more detailed information than web-crawled captions—we introduce two simple yet effective designs to better leverage them in CLIP training. Our first design, inspired by the inverse scaling law of CLIP training revealed in (Li et al., 2024b), involves randomly sampling a portion of the synthetic caption to serve as input to the text encoder. Interestingly, we observe that transitioning from web-crawled captions to synthetic captions leads to a tipping point in these inverse effects—rather than hurting performance (with larger models being less affected with shorter captions) as noted in (Li et al., 2024b, 2023b), dropping parts of the synthetic caption surprisingly leads to performance improvements. The strongest performance is achieved when we randomly select a single sentence from the synthetic caption as the paired text description for contrastive learning, discarding the rest. Moreover, as shown in (Li et al., 2024b), learning with shorter text reduces overall computation, providing additional benefits.

Since the synthetic caption is only partially used in contrastive learning, our second design aims to incorporate their *full* use in an auxiliary task. Specifically, we follow CoCa (Yu et al., 2022) by incorporating an autoregressive decoder to predict captions. However, unlike the symmetric design in CoCa, where the input text and the output text are the same (*i.e.*, the web-crawled caption), we introduce an asymmetric design: the input to the text decoder is the web-crawled caption, and the prediction target is the full-length synthetic caption. This learning behavior mimics the recaptioning process performed by MLLMs and ensures full utilization of the knowledge within the complete synthetic captions.

We termed this resulting training framework as CLIPS. Experimental results show that CLIPS significantly enhances zero-shot performance in cross-modal retrieval. For example, with a ViT-L backbone, our CLIPS substantially outperforms SigLip (Zhai et al., 2023) by 4.7 (from 70.8 to 75.5) on MSCOCO's R@1 text retrieval, and by 3.3 (from 52.3 to 55.6) on MSCOCO's R@1 image retrieval. With increased computational resources and scaling, our best model further achieves 76.4% and 96.6% R@1 text retrieval performance on MSCOCO and Flickr30K respectively, and 57.2% and 83.9% R@1 image retrieval performance on the same datasets, setting new state-of-the-art (SOTA) results. Moreover, our CLIPS framework contributes to building stronger MLLMs—replacing the visual encoder from OpenAI-CLIP (Radford et al., 2021) with our CLIPS in LLaVA (Liu et al., 2024a,b) leads to strong performance gains across a range of MLLM benchmarks.

2 Related Works

Vision-language Pre-training Vision-language pre-training aims to align vision and language modalities to create a unified representation for various image-text understanding tasks. Existing frameworks predominantly adopt either single-stream architecture (Li et al., 2019; Chen et al., 2020; Kim et al., 2021), which jointly represents different modalities using a shared encoder, or two-stream architectures (Lu et al., 2019; Li et al., 2021; Radford et al., 2021; Jia et al., 2021), which employ two independent encoders to process visual and textual inputs separately. Our work builds upon the latter approach, specifically focusing on CLIP, which leverages contrastive loss to align image and text representations.

The further enhancements to CLIP have generally pursued two main directions. The first direction involves extending CLIP's capabilities into generative tasks, such as image captioning, visual question answering, and image grounding. Notable works in this vein include CoCa (Yu et al., 2022) and BLIP (Li et al., 2022), which enables the unification of image-text understanding and generation tasks by transitioning from an encoder-only architecture to an encoder-decoder architecture. The second direction focuses on optimizing vision-language contrastive learning. FILIP (Yao et al., 2021) mitigates the fine-grained alignment issue in CLIP by modifying the contrastive loss. SigLip (Zhai et al., 2023) replaces contrastive loss with sigmoid loss to optimize computation efficiency. Llip (Lavoie et al., 2024) models captions diversity by associating multiple captions with a single image representation. CLOC (Chen et al., 2024a) enhances regional localization by introducing a region-text contrastive loss, improving the model's ability to focus on specific image regions corresponding to textual inputs. Our work also aims to improve CLIP but focuses on enhancing the leverage of richly described synthetic captions in training.

Learning from Synthetic Captions Recognizing that web-crawled image-text datasets are often noisy and contain mismatched pairs, recent works seek to improve dataset quality by rewriting captions. ALip (Yang et al., 2023) uses the OFA (Wang et al., 2022) model to generate synthetic captions and introduces a bi-path model to integrate supervision from two types of text. LaCLIP (Fan et al., 2024) rewrites captions using LLMs such as ChatGPT (OpenAI, 2022) and randomly selects one of these rephrasings during training. Nguyen *et al.* (Nguyen et al., 2024) use BLIP-2 (Li et al., 2023a) to rewrite captions for image-text pairs with low matching degrees in the original dataset. VeCLIP (Lai et al., 2025) first uses LLaVA (Liu et al., 2024b) to generate synthetic captions with rich visual details, then uses an LLM to fuse the raw and synthetic captions. Liu *et al.* (Liu et al., 2023) propose using multiple MLLMs to rewrite captions and apply text shearing to improve caption diversity. ShareGPT4V (Chen et al., 2023) feeds carefully designed prompts and images to GPT-4V (Achiam et al., 2023) to create a high-quality dataset, which has been widely adopted in subsequent works (Lin et al., 2024; Chen et al., 2024b; Chu et al., 2024). Li *et al.* (Li et al., 2024a) rewrites captions using a more advanced LLaMA-3-based (Meta, 2024) LLaVA in the much larger-scale DataComp-1B dataset (Gadre et al., 2024). SynthCLIP (Hammoud et al., 2024) trains entirely on synthetic datasets by generating image-text pairs using text-to-image models and LLMs. DreamLip (Zheng et al., 2025) builds a short caption set for each image and computes multi-positive contrastive loss and sub-caption specific grouping loss. Our first design is closely related to DreamLip, but presents a more general message about the inverse effect of learning with synthetic captions (*i.e.*, short ones are highly preferred). Additionally, to fully leverage the information in the complete synthetic captions, our second design incorporates an CoCa-like but asymmetric decoder (*i.e.*, web-crawled caption as the input, full synthetic caption as the output).

3 Method

This section first covers related preliminaries, including the contrastive loss in CLIP and the generative loss in CoCa. We then present the observed inverse effect of learning with synthetic captions. Lastly, we introduce our simplified multi-positive contrastive loss for the encoder and an asymmetric decoder design for predicting full-length synthetic captions.

3.1 Preliminaries

CLIP Let $\{(x_i, y_i)\}_{i=1}^N$ denote a set of image-text pairs, where x_i represents an image and y_i represents the corresponding text description. Images and texts are encoded using a vision encoder $f(\cdot)$ and a text encoder $g(\cdot)$, respectively. For the image loss L_I, we first calculate the similarity of each image in the local batch with all texts. These similarity values are then used to compute the InfoNCE loss (Oord et al., 2018), with correctly matched image-text pairs treated as positive samples and non-matching pairs as negative samples. The formula can be expressed as:

$$L_\text{I} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\left(\text{sim}(f(x_i), g(y_i))/\tau\right)}{\sum_{j=1}^{N}\exp\left(\text{sim}(f(x_i), g(y_j))/\tau\right)} \tag{1}$$

where $\text{sim}(f(x_i), g(y_j))$ denotes the similarity function (*e.g.*, cosine similarity) and τ is a temperature parameter.

For the text loss L_text, the formula is similar, but the roles of images and texts are swapped. The total loss L_contrast is the average of the image loss and text loss:

$$L_\text{contrast} = \frac{1}{2}\left(L_\text{image} + L_\text{text}\right) \tag{2}$$

CoCa The generative loss here is instantiated as autoregressively predicting the next token in a target sequence conditioned on image features and previously generated tokens. Specifically, the text branch follows an encoder-decoder architecture where the unimodal text encoder encodes text, and the multimodal text decoder generates the output sequence. This process can be formalized as:

$$L_\text{gen} = -\sum_{t=1}^{T}\log P(y_t \mid y_{<t}, E(x)) \tag{3}$$

where y_t represents the target token at time step t, $y_{<t}$ denotes all tokens preceding y_t, $E(x)$ represents the encoded features of the input x, and $P(y_t \mid y_{<t}, E(x))$ is the conditional probability of the target token y_t given the previous tokens and encoded features.

3.2 Inverse Effect with Synthetic Captions

Inspired by the prior work (Li et al., 2024b), we first check how CLIP models behave when learning with *shorter synthetic captions*. As illustrated in

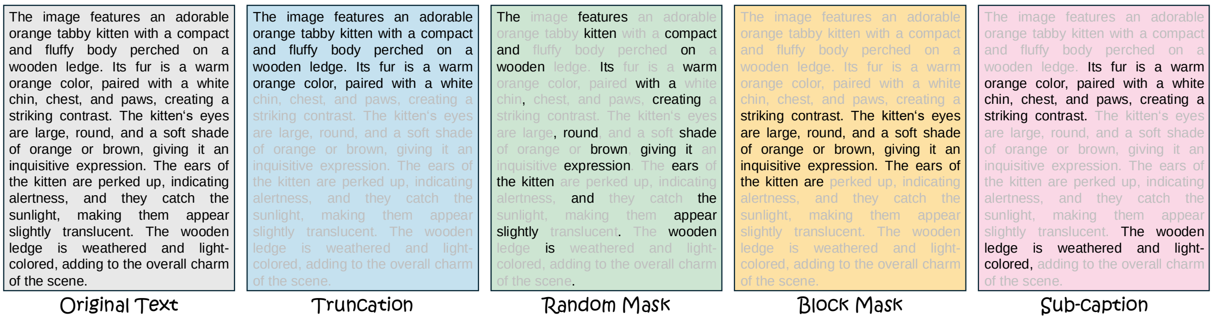


Figure 2: Visualization of four different token reduction strategies. These strategies can improve the model's learning efficiency on synthetic captions to varying degrees. Among these strategies, the sub-caption and block mask perform best.

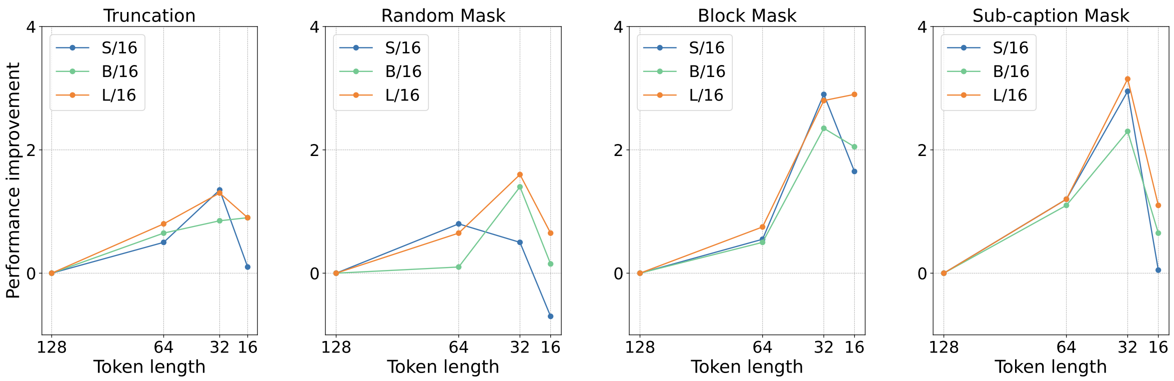


Figure 3: The inverse scaling effect of synthetic captions. Unlike the performance drop from reducing token length in original captions, shortening the token length of synthetic captions consistently improves model performance.

Figure 2, we employ four strategies to reduce the token length of synthetic captions: truncation, random mask, block mask, and sub-caption mask. The truncation, random mask, and block mask strategies are based on the approaches described in (Li et al., 2024b). Notably, we omit the syntax mask from (Li et al., 2024b), which performed best when learning with web-crawled captions but yielded very non-competitive results in our synthetic caption experiments.[1] The sub-caption mask strategy is adapted from the sub-caption extraction method proposed in (Zheng et al., 2025).

Specifically, given a synthetic caption sequence of length K and a target token length L, the truncation directly selects the first L tokens; The random mask obtains L tokens through random sampling; The block mask randomly selects a starting point in the original sequence and take the subsequent L tokens. For the sub-caption mask, we split the synthetic caption at periods, resulting in segments $\{S_1, S_2, \ldots, S_n\}$, where n is the number of sub-captions. We randomly select a sub-caption and check its length: If the length meets and exceeds the predefined limit, we truncate it; Otherwise, we randomly select another sub-caption from the remaining ones, concatenate it with the previous segment, and then check the length again. This process can be formalized as:

$$\text{Subcaption}(S, L) = \begin{cases} \text{Truncate}(S_i), & \text{if } |S_i| > L \\ \text{Concat}(\{S_i, S_j, \ldots\}), & \text{if } |S_i| < L \end{cases} \tag{4}$$

where S_i and S_j denotes randomly selected sub-captions from the set $\{S_1, S_2, \ldots, S_n\}$.

We follow the training setup described in (Li et al., 2024b) to train all models. Additionally, similar to (Zheng et al., 2025), we include both synthetic captions and the original web-crawled captions in the contrastive learning process. We conduct experiments using ViT (Dosovitskiy et al., 2020) models of sizes S/16, B/16, and L/16 for each token reduction strategy, progressively reducing the token length from 128 to 64, 32, and 16. We evaluate the models' average R@1 retrieval performance on the MSCOCO dataset and present the results in Figure 3.

[1]We tested the syntax mask with the ViT-B/16 model and a token length of 32, finding that it achieved an average retrieval performance of 52.8, slightly below the performance without any mask. In contrast, under the same setup, the truncation, random mask, and block mask strategies led to noticeable performance improvements, as shown in Figure 3.

Main observations. Firstly, these results confirm that the inverse effects also exists when training with synthetic captions; that is, learning with reduced token lengths is generally preferred. But we stress that this inverse effect becomes significantly more pronounced when transitioning from web-crawled captions to synthetic captions—while the inverse effect in (Li et al., 2024b) refers to that larger models being less *adversely* affected when learning with reduced-length web-crawled captions, our experiments hereby reveals that, when using synthetic captions, reducing the token length consistently yields *noticeable performance gains* across all model sizes. In other words, learning with synthetic captions at a reduced length is a more "optimal" way to train CLIP, enjoying the benefits of both higher performance and higher efficiency (due to less text token used). As a side note, this observation also corroborates the prior works on showing text shearing (Liu et al., 2023) and sub-caption extraction (Zheng et al., 2025) are effective strategies to process long synthetic captions for enhancing performance.

Moreover, by taking a closer look at Figure 3, we note that the sub-caption mask and block mask perform the best, and the truncation and random mask are slightly less effective. We conjecture that, in addition to sequence length being a critical factor in contrastive learning, the diversity and coherence of the reduced text segments also play significant roles. Furthermore, although we do not observe a clear scaling trend between the S/16 and B/16 models, it is evident that the larger L/16 model achieves the most substantial performance gains compared to the smaller models.

3.3 Encoder: Learning with Short Captions

Based on the observation above, we adopt the sub-caption strategy as the default preprocessing method for synthetic captions in our encoder design. Specifically, considering that all ViTs generally achieve the strongest performance at an input token length of 32, roughly corresponding to one or two sentences from the full synthetic caption, we thereby implement the sub-caption strategy by sampling a single random sentence from each full synthetic caption. This sampled sentence, along with the original web-crawled caption, is then input into the text encoder, and a multiple-positive contrastive loss will be applied. This can be expressed as:

$$L_{\text{contrast}} = -\frac{1}{2N}\sum_{i=1}^{N}\left(L^i_{\text{orig}} + L^i_{\text{syn-short}}\right) \qquad (5)$$

where

$$L^i_{\text{orig}} = \log\frac{\exp(S^i_{i,\text{orig}})}{\sum_{j=1}^{N}\exp(S^k_{i,\text{orig}})}, \qquad (6)$$

$$L^i_{\text{syn-short}} = \log\frac{\exp(S^i_{i,\text{syn-short}})}{\sum_{j=1}^{N}\exp(S^k_{i,\text{syn-short}})} \qquad (7)$$

where $S^i_{i,\text{orig}}$ and $S^i_{i,\text{syn-short}}$ represent the scaled similarity for the original and short synthetic captions, respectively.

Note this training process is closely related to the prior work DreamLip (Zheng et al., 2025)—which uses the sub-caption strategy to preprocess synthetic data—but in a much simpler format. Specifically, in DreamLip, an average contrastive loss is calculated over all captions (*i.e.*, the original web-crawled caption + a set of diverse synthetic captions) for image-text alignment. In comparison, ours directly trains with a single short synthetic caption rather than a set of them (*e.g.*, typically this set contains more than 6 elements). As empirically shown in Section 4, our simplified strategy is sufficient to achieve strong performance, with our best model setting new state-of-the-art results in cross-modal retrieval tasks. In addition to this training simplification, our experiments are conducted at a significantly larger computational scale, *i.e.*, our training dataset is ~33× larger (*i.e.*, Merged-30M (Zheng et al., 2025) *vs.* DataComp-1B) and the explored model size is up to ViT-H (while only up to ViT-B in DreamLip).

3.4 Decoder: Predicting Full Synthetic Caption

In our encoder design, we utilize only a portion of the synthetic captions during contrastive learning to achieve stronger performance and higher computation efficiency. However, this approach leaves substantial contextual information within the richly described synthetic captions unexploited. Furthermore, the preference for learning with shorter captions may suggest that the text encoder is unable to fully comprehend long sequences within the CLIP's contrastive learning framework. To this end, our second design introduces a generative task as an auxiliary objective to assist CLIP in capturing the full data distribution. Specifically, we adopt the

Table 1: **Zero-shot cross-modal retrieval results on MSCOCO and Flickr30K**, with CLIPA and CoCa results reproduced by us. Both methods are implemented with a mixture training, where the original caption accounts for 80% and the synthetic caption accounts for 20%.

Model	**Method**	**MSCOCO**						**Flickr30K**					
		Image→Text			Text→Image			Image→Text			Text→Image		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
S/16	CLIPA (Li et al., 2024a)	55.8	79.5	87.6	35.0	60.9	71.6	79.6	95.4	98.0	59.2	83.2	89.3
	CoCa (Yu et al., 2022)	55.5	79.2	86.8	35.7	61.3	71.6	79.4	94.1	97.3	59.3	83.4	89.7
	CLIPS	**61.8**	**83.7**	**90.0**	**39.4**	**65.2**	**75.2**	**85.9**	**96.7**	**98.5**	**66.4**	**87.2**	**92.7**
B/16	CLIPA (Li et al., 2024a)	62.8	85.0	91.0	42.7	67.6	77.5	86.7	97.9	98.7	67.9	88.8	92.8
	CoCa (Yu et al., 2022)	63.1	84.4	90.1	43.2	68.3	77.6	87.8	97.8	99.2	68.2	88.9	93.2
	CLIPS	**68.8**	**88.4**	**92.9**	**48.2**	**72.8**	**81.4**	**92.5**	**99.3**	**99.8**	**75.3**	**92.4**	**95.8**
L/16	CLIPA (Li et al., 2024a)	66.8	87.0	92.5	47.8	72.6	81.3	91.3	98.9	99.4	73.9	92.2	95.6
	CoCa (Yu et al., 2022)	67.0	87.1	92.4	48.0	72.3	80.9	89.7	98.7	99.6	74.1	92.0	95.5
	CLIPS	**73.6**	**90.6**	**94.5**	**53.6**	**76.6**	**84.4**	**94.2**	**99.2**	**99.9**	**80.6**	**95.0**	**97.2**

caption modeling strategy from CoCa (Yu et al., 2022) by incorporating an autoregressive decoder for caption generation. Importantly, unlike the symmetric decoder design in CoCa—where the decoder processes identical input and output text—we introduce an asymmetric learning structure: the text decoder takes the web-crawled caption as input and generates the full-length synthetic caption as the target. This learning approach mimics the recaptioning process performed by MLLMs, ensuring the full utilization of the knowledge contained in the synthetic captions. Additionally, it facilitates modeling the relationship between the two types of captions, with synthetic captions serving as a more semantically aligned and knowledge-rich representation of the web-crawled captions.

In our implementation, unlike the text branch in CoCa, we remove the causal mask from the original unimodal text decoder to implement a more effective text encoder, which will be further discussed in Sec. 4.2. Then, in the multimodal text decoder, since the bidirectional text encoder's output cannot be directly used as input, we replace the sequential input text with a set of randomly initialized learnable tokens and use images and web-crawled captions as conditioning information. Formally, given the image features $\mathbf{I}$ and learned tokens $\mathbf{L}_{\text{learn}}$, predicting full-length synthetic captions from web-crawled captions can be expressed based on the Eq. (3) as:

$$L_{\text{gen}} = -\sum_{t=1}^{T} \log P(y_t \mid \mathbf{I}, \mathbf{C}_{\text{web}}, \mathbf{L}_{\text{learn},<t}) \quad (8)$$

where y_t represents the target token at time step t, $\mathbf{L}_{\text{learn},<t}$ denotes all learnable tokens preceding t, and $\mathbf{C}_{\text{web}}$ represents the web-crawled captions. The term $P(y_t \mid \mathbf{I}_{\text{image}}, \mathbf{C}_{\text{web}}, \mathbf{L}_{\text{learn},<t})$ is the conditional probability of generating the target token y_t given the image features, web-crawled captions, and preceding learnable tokens.

Moreover, our experiments show that simply concatenating information from different modalities yields stronger results than employing modality fusion with cross-attention mechanisms. Therefore, our decoder utilizes a self-attention mechanism accompanied by a specially designed combination mask. Specifically, to construct the input to the decoder, we concatenate the image features and the web-crawled caption to form the conditioning sequence, then concatenate this condition with the learnable tokens to form the complete input sequence. We design the combination mask M to ensure that tokens within the condition can attend to each other, while the learnable tokens follow an autoregressive prediction pattern. The mask M is defined as:

$$M[i,j] = \begin{cases} 1 & \text{if } i, j \le L_{\text{cond}} \\ 1 & \text{if } i > L_{\text{cond}} \text{ and } j \le L_{\text{cond}} \\ 1 & \text{if } i, j > L_{\text{cond}} \text{ and } i \ge j \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

where L_{cond} represents the total length of the condition tokens, which are formed by concatenating the image tokens and the web-crawled caption tokens.

The combined input sequence passes through a self-attention mechanism guided by a mask M to enable appropriate token interactions. To align the generated sequence with the full-length synthetic captions $\mathbf{T}_{\text{synthetic, full}}$, we compute the loss based on the likelihood of each token in the target generated sequence. Accordingly, the final generative loss L_{gen} is defined as:

$$L_{\text{gen}} = -\sum_{t=1}^{T} \log P_\theta \left(y_t \mid \mathbf{I} \oplus \mathbf{C}_{\text{web}} \oplus \mathbf{L}_{\text{learn},<t}, M\right) \tag{10}$$

where y_t represents the t-th token in the full-length synthetic caption $\mathbf{T}_{\text{synthetic, full}}$. The overall optimization objective of the entire model can be formalized as:

$$L_{\text{total}} = \alpha \cdot L_{\text{contrast}} + \beta \cdot L_{\text{gen}} \tag{11}$$

where α and β are weighting factors that balance the contributions of the contrastive loss and the generative loss.

4 Experiments

4.1 Pre-training Details

We conduct our experiments using the Recap-DataComp-1B dataset (Li et al., 2024a) for pre-training. Following the efficient training recipe introduced in (Li et al., 2024b), all main experiments involve pre-training models for 2,000 ImageNet-eq. epochs (~2.6B samples seen) with images at a low resolution (*e.g.*, resizing to 112 × 112 by default), followed by fine-tuning for 100 ImageNet-eq. epochs at the resolution of 224 × 224. For the text branch, the input token length is 80,[2] and the output token length is 128, matching the number of learnable tokens in the decoder. The batch size is 32,768 for pre-training and 16,384 for fine-tuning. In the decoder's autoregressive generation, image tokens and text tokens are concatenated to fuse information from different modalities. The weights for the contrastive loss (α) and generative loss (β) are set to 1 and 2, respectively.

To further enhance training, in our experiment about comparison with state-of-the-art (SOTA) methods, we increase the pre-training epochs to 10,000 ImageNet-equivalent epochs (~13B samples seen) and fine-tune for 400 ImageNet-equivalent epochs. The image resolution during pre-training is set to 84 for accelerating training, and remains at 224 during fine-tuning. Other settings remain consistent with those in the main experiments. More Detailed training parameters are provided in the appendix.

[2]Note that, despite setting the max text length as 80, we still only sample a single sentence from the synthetic caption as described in Section 3.3 and pad the remaining positions. Empirically, we observe that this strategy yields an additional ~1% performance improvement compared to setting the maximum text length to 32.

4.2 Evaluation

Zero-Shot cross-modal retrieval. We evaluate the zero-shot cross-modal retrieval performance across different model sizes on the MSCOCO (Lin et al., 2014) and Flickr30K (Plummer et al., 2015) datasets. We use CLIPA and CoCa as baselines and follow (Li et al., 2024a) to train with a mixture of web-crawled captions (80%) and synthetic captions (20%).

As reported in Table 1, our method consistently achieves superior performance across all benchmarks and model sizes, yielding significant improvements over the baselines. Notably, these substantial gains enable our smaller models to match or even surpass the performance of larger models from previous works. For instance, our S/16 model attains results comparable to the B/16 models of CLIPA and CoCa, and our B/16 model reaches the performance level of their L/16 models. These findings demonstrate the efficacy of our approach in enhancing cross-modal representation learning.

Comparision with SOTA methods. In Table 2, we further compare our method with state-of-the-art vision-language pre-training approaches, reporting top-1 accuracy on ImageNet-1K and zero-shot recall rates for image and text retrieval tasks on MSCOCO and Flickr30K. At the model size of L/14, our CLIPS achieves recall@1 scores of 75.5 and 96.5 for text retrieval, and 55.6 and 82.3 for image retrieval on MSCOCO and Flickr30K, respectively. These results are significantly higher than those of the prior art SigLIP (Zhai et al., 2023), and they also surpass the concurrent work CLOC (Chen et al., 2024a), which pretrains CLIP with a more fine-grained region-text contrastive loss. This performance trend remains consistent at the huge model size—*i.e.*, with ViT-H/14, our CLIPS strongly attains recall@1 scores of 76.4 and 96.6 for text retrieval, and 57.2 and 83.9 for image retrieval, setting new SOTA records.

Despite the strong cross-modal retrieval performance, we note that our CLIPS is less competitive on ImageNet zero-shot classification accuracy. We conjecture that this is mainly due to two factors: (1) training with synthetic caption is expected to yield lower ImageNet accuracy, as empirically shown in (Li et al., 2024a); and (2) our training dataset, DataComp-1B, is expected to yield lower ImageNet accuracy than the higher-quality DFN dataset (Fang et al., 2023) and possibly Google's in-house WebLI dataset (Chen et al., 2022).

Table 2: **Comparison with other SOTA vision-language pre-training methods** trained on public or private dataset. We report top-1 ImageNet-1K (Deng et al., 2009) classification accuracy and zero-shot recall of image and text retrieval on MSCOCO and Flickr30K.

method	model size	# patches	dataset	public	IN-1K	MSCOCO R@1		Flickr30K R@1	
					val.	I→T	T→I	I→T	T→I
CLIP (Radford et al., 2021)	Large	256	WIT-400M (Radford et al., 2021)	✗	75.5	56.3	36.5	85.2	65.0
CoCa (Yu et al., 2022)	Large	256	LAION-2B (Schuhmann et al., 2022)	✗	75.6	62.9	45.7	88.4	74.3
CLIP (Gadre et al., 2024)	Large	256	DataComp-1B (Gadre et al., 2024)	✓	79.2	63.3	45.7	89.0	73.4
OpenCLIP (Ilharco et al., 2021)	Large	256	LAION-2B (Schuhmann et al., 2022)	✓	75.5	63.4	46.5	89.5	75.5
SigLIP (Zhai et al., 2023)	Large	256	WebLI-5B (Chen et al., 2022)	✗	**80.5**	70.8	52.3	91.8	79.0
CLOC (Chen et al., 2024a)	Large	576	WiT (Wu et al., 2023)+DFN-5B (Fang et al., 2023)	✗	80.1	74.8	54.4	-	-
CLIPS	Large	256	Recap-DataComp-1B (Li et al., 2024a)	✓	78.5	**75.5**	**55.6**	**96.5**	**82.3**
CLIP (Fang et al., 2023)	Huge	729	DFN-5B (Fang et al., 2023)	✗	**84.4**	71.9	55.6	94.0	82.0
SigLIP (Zhai et al., 2023)	SO(400M)	729	WebLI-5B (Chen et al., 2022)	✗	83.1	72.4	54.2	94.3	83.0
CLOC (Chen et al., 2024a)	Huge	576	WiT (Wu et al., 2023)+DFN-5B (Fang et al., 2023)	✗	81.3	75.7	55.1	-	-
CLIPS	Huge	256	Recap-DataComp-1B (Li et al., 2024a)	✓	79.9	**76.4**	**57.2**	**96.6**	**83.9**

Table 3: Comparison of LLaVA-1.5 performance trained with our visual encoder versus CLIP's visual encoder across multiple MLLM benchmarks. The visual encoder size is L/14, the LLM used is LLaMA-3, and all results are reproduced by us.

ViT	MME-Cognition	MME-Perception	MMMU	MM-VET	GQA	ChartQA	POPE	NoCaps	TextVQA
OpenAI-CLIP	295.4	**1433.5**	37.8	34.7	57.5	13.2	83.8	92.1	56.8
CLIPS	**326.4**	1416.6	**38.2**	**35.8**	**60.3**	**14.2**	**86.7**	**102.5**	**57.6**

CLIPS in LLaVA. To assess the visual representation capabilities of our pre-trained model, we integrate the CLIPS visual encoder into LLaVA-1.5 (Liu et al., 2024a) and evaluate its performance. Specifically, we replace the original OpenAI-CLIP-L/14 visual encoder with our CLIPS-L/14 and utilized LLaMA-3 (Dubey et al., 2024) as the language model. Since our pre-training employs images at a smaller resolution, we fine-tune CLIPS-L/14 at a resolution of 336×336 to match the configuration of OpenAI-CLIP-L/14. We then evaluate LLaVA's performance on multiple multimodal large language model (MLLM) benchmarks, including MME (Fu et al., 2023), MMMU (Yue et al., 2024), GQA (Hudson and Manning, 2019), ChartQA (Masry et al., 2022), POPE (Li et al., 2023c), NoCaps (Agrawal et al., 2019), and TextVQA (Singh et al., 2019).

The results summarized in Table 3 demonstrate that integrating CLIPS significantly enhances LLaVA's performance across multiple metrics compared to using the original OpenAI-CLIP visual encoder. Specifically, out of the nine metrics evaluated, substituting OpenAI-CLIP with our CLIPS leads to performance gains in eight metrics. Notably, the most substantial improvements are observed on the NoCaps task, where performance increases by 10.4 points (from 92.1 to 102.5), and on the MME-Cognition task, with a boost of 31.0 points (from 295.4 to 326.4). Collectively, these results confirm that the strong visual capabilities of our CLIPS effectively transfer to the multimodal setting, enabling MLLMs to achieve a deeper understanding of visual content. This highlights the potential of CLIPS as a general-purpose vision encoder for multimodal applications.

5 Conclusion

This work introduces CLIPS, with two simple and effective changes to enhance vision-language pre-training with synthetic captions. Our first design in on the text encoder—by observing the strong inverse effect in learning with short synthetic caption, we therefore feed only a portion of synthetic caption for contrastive learning. Then, to fully leverage the full synthetic caption, our second design incorporates an autoregressive captioner to mimic the recaptioning process—conditioning on the paired image input and web-crawled text description, the captioner learns to predict the full-length synthetic caption generated by advanced MLLMs. Experimental results show that CLIPS significantly improves zero-shot performance in cross-modal retrieval, reaching 76.4% and 96.6% for text retrieval and 57.2% and 83.9% for image retrieval on MSCOCO and Flickr30K, respectively, setting new SOTA results. Moreover, the visual encoder we trained strongly improves the visual capabilities of LLaVA, achieving notable gains across multiple MLLM benchmarks.

6 Limitations

Although our method significantly advances visual-language representation learning, it still lacks fine-grained alignment at the token level. Additionally, while long texts play a role in generative learning, their direct application to contrastive learning remains underexplored. In the future, we will explore implementing a fine-grained alignment framework, making fuller use of synthetic captions, and employing visual encoders with improved representations to achieve more powerful MLLMs.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. 2019. nocaps: novel object captioning at scale. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8948–8957.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Hong-You Chen, Zhengfeng Lai, Haotian Zhang, Xinze Wang, Marcin Eichner, Keen You, Meng Cao, Bowen Zhang, Yinfei Yang, and Zhe Gan. 2024a. Contrastive localized language-image pre-training. *arXiv preprint arXiv:2410.02746*.

Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. 2024b. Pixart-\sigma: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. *arXiv preprint arXiv:2403.04692*.

Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2023. Sharegpt4v: Improving large multimodal models with better captions. *arXiv preprint arXiv:2311.12793*.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, and 1 others. 2022. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024c. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198.

Xiangxiang Chu, Limeng Qiao, Xinyu Zhang, Shuang Xu, Fei Wei, Yang Yang, Xiaofei Sun, Yiming Hu, Xinyang Lin, Bo Zhang, and 1 others. 2024. Mobilevlm v2: Faster and stronger baseline for vision language model. *arXiv preprint arXiv:2402.03766*.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, and 1 others. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. 2024. Improving clip training with language rewrites. *Advances in Neural Information Processing Systems*, 36.

Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. 2023. Data filtering networks. *arXiv preprint arXiv:2309.17425*.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and 1 others. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, and 1 others. 2024. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36.

Hasan Abed Al Kader Hammoud, Hani Itani, Fabio Pizzati, Philip Torr, Adel Bibi, and Bernard Ghanem. 2024. Synthclip: Are we ready for a fully synthetic clip training? *arXiv preprint arXiv:2402.01832*.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. 2021. Openclip. *github*.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, pages 5583–5594. PMLR.

Zhengfeng Lai, Haotian Zhang, Bowen Zhang, Wentao Wu, Haoping Bai, Aleksei Timofeev, Xianzhi Du, Zhe Gan, Jiulong Shan, Chen-Nee Chuah, and 1 others. 2025. Veclip: Improving clip training via visual-enriched captions. In *European Conference on Computer Vision*, pages 111–127. Springer.

Samuel Lavoie, Polina Kirichenko, Mark Ibrahim, Mahmoud Assran, Andrew Gordon Wildon, Aaron Courville, and Nicolas Ballas. 2024. Modeling caption diversity in contrastive vision-language pretraining. *arXiv preprint arXiv:2405.00740*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pretraining with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pretraining for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Xianhang Li, Haoqin Tu, Mude Hui, Zeyu Wang, Bingchen Zhao, Junfei Xiao, Sucheng Ren, Jieru Mei, Qing Liu, Huangjie Zheng, and 1 others. 2024a. What if we recaption billions of web images with llama-3? *arXiv preprint arXiv:2406.08478*.

Xianhang Li, Zeyu Wang, and Cihang Xie. 2023b. Clipa-v2: Scaling clip training with 81.1% zero-shot imagenet accuracy within a $10,000 budget; an extra $4,000 unlocks 81.8% accuracy. *arXiv preprint arXiv:2306.15658*.

Xianhang Li, Zeyu Wang, and Cihang Xie. 2024b. An inverse scaling law for clip training. *Advances in Neural Information Processing Systems*, 36.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023c. Evaluating object hallucination in large vision-language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Bin Lin, Zhenyu Tang, Yang Ye, Jiaxi Cui, Bin Zhu, Peng Jin, Junwu Zhang, Munan Ning, and Li Yuan. 2024. Moe-llava: Mixture of experts for large vision-language models. *arXiv preprint arXiv:2401.15947*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024b. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Yanqing Liu, Kai Wang, Wenqi Shao, Ping Luo, Yu Qiao, Mike Zheng Shou, Kaipeng Zhang, and Yang You. 2023. Mllms-augmented visual-language representation learning. *arXiv preprint arXiv:2311.18765*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Ahmed Masry, Do Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland. Association for Computational Linguistics.

AI Meta. 2024. Introducing meta llama 3: The most capable openly available llm to date. *Meta AI*.

Thao Nguyen, Samir Yitzhak Gadre, Gabriel Ilharco, Sewoong Oh, and Ludwig Schmidt. 2024. Improving multimodal datasets with image captioning. *Advances in Neural Information Processing Systems*, 36.

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

OpenAI. 2022. Introducing chatgpt. *https://openai.com/blog/chatgpt*.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, and 1 others. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294.

Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8317–8326.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International conference on machine learning*, pages 23318–23340. PMLR.

Wentao Wu, Aleksei Timofeev, Chen Chen, Bowen Zhang, Kun Duan, Shuangning Liu, Yantao Zheng, Jonathon Shlens, Xianzhi Du, Zhe Gan, and 1 others. 2023. Mofi: Learning image representations from noisy entity annotated images. *arXiv preprint arXiv:2306.07952*.

Kaicheng Yang, Jiankang Deng, Xiang An, Jiawei Li, Ziyong Feng, Jia Guo, Jing Yang, and Tongliang Liu. 2023. Alip: Adaptive language-image pre-training with synthetic caption. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2922–2931.

Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. 2021. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, and 3 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of CVPR*.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986.

Kecheng Zheng, Yifei Zhang, Wei Wu, Fan Lu, Shuailei Ma, Xin Jin, Wei Chen, and Yujun Shen. 2025. Dreamlip: Language-image pre-training with long captions. In *European Conference on Computer Vision*, pages 73–90. Springer.

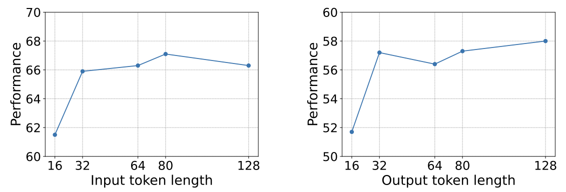
A Additional Ablation Studies

A.1 Components.

We hereby ablate the effects of different design components—sub-caption strategy, multi-positive contrastive loss, and generative loss—in CLIPS. We use the Recap-CLIP-B/16 model with a mixed ratio of 0.6 (Li et al., 2024a) as our baseline and examine how these components contribute to improvements in cross-modal retrieval performance on the MSCOCO dataset and classification accuracy on ImageNet. We first introduce sub-caption extraction for the remaining 40% of synthetic captions under the given setting. As reported in the second row of Table 4, this strategy enhances the model's cross-modal retrieval performance; but it also leads to a slight accuracy drop on ImageNet, possibly due to the reduced informational content in the synthetic sub-captions. Subsequently, we incorporate the multi-positive contrastive loss. By engaging in contrastive learning with both the original captions and the synthetic sub-captions, this approach facilitates the establishment of a more robust one-to-many relationship between images and captions. As a result, we observe substantial and consistent improvements in both cross-modal retrieval performance and ImageNet accuracy. Finally, we introduce the generative loss, which involves reconstructing the full synthetic captions based on the images and the original web-crawled captions. This addition consistently benefits all

Table 4: Ablation study on components (Zero-shot Performance): SC = Sub-caption, MP = Multi-positive Contrastive Loss, GL = Generative Loss. I→T = Image-to-Text Retrieval, T→I = Text-to-Image Retrieval, both are MSCOCO's R@1.

Component	**I→T**	**T→I**	**IN1K**
Baseline	63.1	43.1	69.0
+ SC	64.5$_{+1.4\%}$	44.7$_{+1.6\%}$	68.4$_{-0.6\%}$
+ SC & MP	67.1$_{+4.0\%}$	46.0$_{+2.9\%}$	69.4$_{+0.4\%}$
+ SC & MP & GL	68.8$_{+5.7\%}$	48.2$_{+5.1\%}$	70.2$_{+1.2\%}$



(a) Input Token Length vs. Model Performance.

(b) Output Token Length vs. Model Performance.

Figure 4: Ablation study on input and output token lengths. (a) pads a single sub-caption to different input lengths. (b) keeps the input valid token length constant and varies the target output token length. Performance is measured by R@1 of I→T on MSCOCO.

evaluated metrics. Specifically, when compared to the original baseline, this final configuration yields an improvement of +5.7% in image-to-text retrieval, +5.1% in text-to-image retrieval, and +1.2% in ImageNet accuracy.

A.2 Single sub-caption.

In Section 3, we observe that the sequence length of the input text is a key factor affecting contrastive learning performance, with semantic continuity also playing a role. As randomly sampling multiple sub-captions and then truncating them can introduce semantic discontinuity, we hereby explore how randomly sampling a single sub-caption and padding it to different lengths affects model performance. As shown in Figure 4a, increasing the input sequence length leads to improvements in model performance up to a certain point. This suggests that contrastive learning benefits from longer effective input lengths and that moderate padding can enhance performance. However, when the number of input tokens reaches 128, performance begins to decline, indicating that excessive padding may interfere with the model's ability to learn effective representations.

Table 5: Ablation study on generation target (Zero-shot R@1 Retrieval Performance). Synthetic captions bring more significant performance improvements when used as generation targets.

Generation Target	**MSCOCO**		**Flickr30K**	
	I→T	T→I	I→T	T→I
Web captions	55.9	36.9	88.1	85.0
Synthetic captions	58.4$_{+2.5\%}$	38.8$_{+1.9\%}$	90.2$_{+2.1\%}$	87.3$_{+2.3\%}$

A.3 Generation target.

We explore how using text from different sources as generation targets affects model performance, as presented in Table 5. In these experiments, we maintain the use of web-crawled captions for contrastive learning but introduce an additional text batch for varying the generation target. Compared to the default setup in CoCa, where the web-crawled caption is set as the prediction target, our findings reveal that switching to synthetic captions as generation targets leads to significant performance improvements in cross-modal retrieval. This result underscores the rationale for selecting complete synthetic captions as prediction targets: their richer knowledge and smoother semantic distribution make them ideal for modeling, thereby achieving stronger performance gains.

A.4 Generated sequence length.

We further examine how varying the lengths of generated sequences impact model performance while keeping the effective input length fixed. Due to the short length of the original captions, we rely entirely on synthetic captions to train CoCa.To maintain consistency in contrastive learning, we sample a single sub-caption and pad it to different lengths to preserve the effective text length, while exploring the impact of output lengths by adjusting the autoregressive label length. The experimental results are presented in Figure 4b. We find that model performance tends to improve as the generated sequence length increases. This observation further confirms that long sequences are suitable for generative learning.

A.5 Fusion Type and Conditioning Strategy

In the context of generative learning, we hereby explore the impact of various conditioning modalities and fusion methods. Specifically, we employ concatenation and cross-attention as fusion types and examine the model's retrieval performance when conditioned on either the image alone or a combination of image and text. The experimental re-

Table 6: Ablation study on fusion types and conditions. Concat denotes concatenating condition tokens and learnable tokens as the input. Img&Txt denotes concatenating image tokens and global text tokens as conditions.

Fusion type		Condition		MSCOCO	
Concat	Cross_attn	Img	Img&Txt	I→T	T→I
✓		✓		67.8	47.1
✓			✓	**68.8**	**48.2**
	✓	✓		68.3	47.4
	✓		✓	68.3	48.2

sults are reported in Table 6. Our findings indicate that integrating image and text—whether by using self-attention after concatenation or by applying cross-attention directly—consistently enhances model performance. Regarding the fusion techniques, cross-attention performs better when using the image alone as the condition, whereas concatenation slightly outperforms cross-attention when both image and text are used as conditions. Based on these results, we therefore choose concatenation as the default fusion type in our main experiments, with both image and text as conditions, to achieve the best performance.

A.6 Effect of Causal Masking

Causal mask. In CoCa, the text encoder uses a causal mask to ensure that the model processes text sequentially. However, we find that the causal mask is not essential in our CLIPS; in fact, removing it can enhance model performance. Specifically, in Table 7, we explore the impact of the causal mask, learnable tokens, and input content on the model's efficacy. When utilizing the causal mask, we examine two settings: *first prediction* and *random prediction*. In the first prediction setting, the model uses the first sub-caption to predict the complete synthetic caption, whereas in the random prediction setting, it uses a randomly selected sub-caption. Our results indicate that, in both settings, the use of causal masks leads to a decrease in model performance. We hypothesize that this is because, in contrastive learning, capturing a strong global representation is more crucial than focusing on local features. This global representation also facilitates effective text reconstruction by the text decoder.

B Experiment Details

We provide detailed pre-training and fine-tuning hyperparameters in Table 8. Except for settings specific to our method, all other parameters follow

Table 7: Discussion about causal mask. L-tokens represents learnable tokens, and content represents the input text.

Causal mask		L-tokens		Content		MSCOCO	
w/	w/o	w/	w/o	first	random	I→T	T→I
	✓	✓		None	None	**68.8**	**48.2**
✓		✓		None	None	67.8	47.6
✓			✓	✓		66.8	46.2
✓			✓		✓	67.6	47.4

Recap-DataComp-1B (Li et al., 2024a). The hyperparameters for Large-14 and Huge-14 in the SOTA experiments are listed in Table 9.

Table 8: Hyperparameters for Pretraining and Finetuning

Hyperparameter	Pretraining	Finetuning
Learning rate	$8 \times 128 \times 10^{-6}$	$4 \times 64 \times 10^{-7}$
Batch size	32768	16384
Optimizer	AdamW ($\beta_1 = 0.9, \beta_2 = 0.95$)	
Weight decay	0.2	0.2
Number of epochs	2000	100
Warm-up steps	40	20
CLIP-loss-weight	1	1
Caption-loss-weight	2	2
Input token length	80	80
Output token length	128	128
Resolution	112	224
temperature-init	1/0.07	1/0.07

Table 9: Hyperparameters for SOTA Experiments

Hyperparameter	**Pretraining**	**Fine-tuning**
Learning rate	$8 \times 128 \times 10^{-6}$	$4 \times 64 \times 10^{-7}$
Batch size	32768	16384
Optimizer	AdamW $(\beta_1 = 0.9, \beta_2 = 0.95)$	
Weight decay	0.2	0.2
Number of epochs	10000	400
Warm-up steps	40	20
CLIP-loss-weight	1	1
Caption-loss-weight	2	2
Input token length	80	80
Output token length	128	128
Resolution	84	224
temperature-init	1/0.07	1/0.07