SyCoCa: Symmetrizing Contrastive Captioners with Attentive Masking for Multimodal Alignment

Ziping Ma [1] Furong Xu [1] Jian Liu [1] Ming Yang [1] Qingpei Guo [1]

Abstract

Multimodal alignment between language and vision is the fundamental topic in current vision-language model research. Contrastive Captioners (CoCa), as a representative method, integrates Contrastive Language-Image Pretraining (CLIP) and Image Caption (IC) into a unified framework, resulting in impressive results. CLIP imposes a bidirectional constraints on global representations of entire images and sentences. Although IC conducts an unidirectional image-to-text generation on local representation, it lacks any constraint on local text-to-image reconstruction, which limits the ability to understand images at a fine-grained level when aligned with texts. To achieve multimodal alignment from both global and local perspectives, this paper proposes Symmetrizing Contrastive Captioners (SyCoCa), which introduces bidirectional interactions on images and texts across the global and local representation levels. Specifically, we expand a Text-Guided Masked Image Modeling (TG-MIM) head based on ITC and IC heads. The improved SyCoCa further leverages textual cues to reconstruct contextual images and visual cues to predict textual contents. When implementing bidirectional local interactions, the local contents of images tend to be cluttered or unrelated to their textual descriptions. Thus, we employ an attentive masking strategy to select effective image patches for interaction. Extensive experiments on five vision-language tasks, including image-text retrieval, image-captioning, visual question answering, and zero-shot/finetuned image classification, validate the effectiveness of our proposed method.

1. Introduction

In recent years, the dramatic progress of multimodal alignment between vision and language has reshaped computer vision research in some sense. The availability of large-scale datasets and powerful computational resources has led to many seminal works and impressive results in this field. Most methods use a contrastive objective to constrain global representations between modalities as shown in Figure 1 (a), such as CLIP (Radford et al., 2021; Yang et al., 2022a), ALBEF (Li et al., 2021), mPLUG (Li et al., 2022a; Xu et al., 2023). BEiT-v3 (Wang et al., 2023a) treats image as a foreign language and uses a mask-then-predict strategy for pre-training. This pretraining method requires separate fine-tuning for each downstream task and the pretrained model cannot be directly used across tasks. CoCa (Yu et al., 2022a) combines Image-Text Contrastive (ITC) and Image Caption (IC), resulting in a pretrained model that can be directly applied for both retrieval and IC tasks. It is a promising architecture and works well on multiple vision-language tasks.

However, in terms of local interaction between image patches and text tokens, the IC head in CoCa only utilizes the visual cues to generate textual descriptions, yet disregarding the visual context reconstruction from textual cues. From vision pre-training works, such as SimMIM (Xie et al., 2022) and MAE (He et al., 2022), we learn that the image reconstruction can learn a strong content representation. Therefore, in the multimodal scenario, when the text cue is introduced into the image reconstruction task and local interaction is performed, the representation of texts and images can be unified into one space, thereby further enhancing multimodal alignment.

In this paper, we propose a novel framework called Symmetrizing Contrastive Captioners (SyCoCa) that incorporates both local image-to-text generation and text-to-image reconstruction in addition to the global constractive objective. In addition to ITC head and IC head, we introduce a text-guided masked image modeling (TG-MIM) head. The difference between CoCa and our method is illustrated in Figure 1. CoCa only achieved undirectional interaction between image and text as shown in Figure 1 (b). Further, in Figure 1 (c), our TG-MIM introduces text for image recon-

[1] Ant Group. Correspondence to: Qingpei Guo <qingpei.gqp@antgroup.com>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

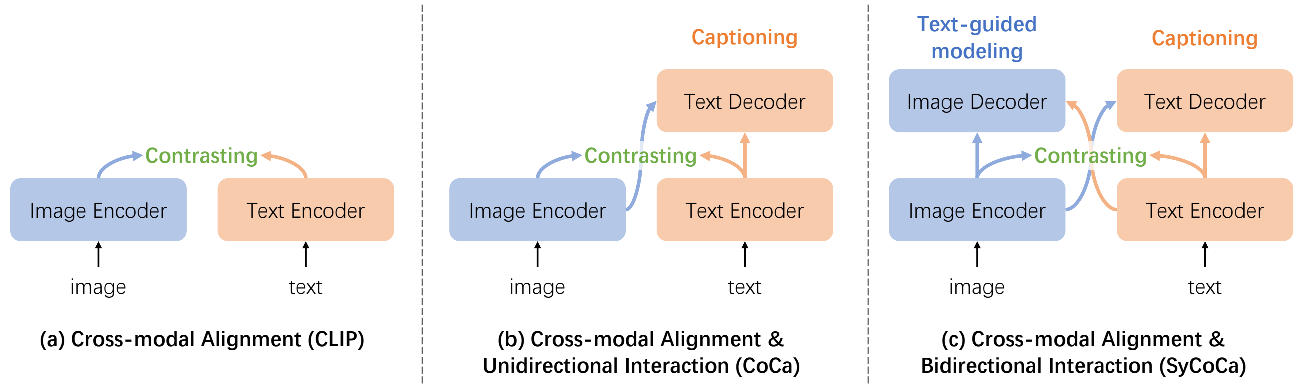


Figure 1. Comparison of the pipelines in vision-language pretraining frameworks. (a) CLIP: aligning global features across modalities through contrastive learning. (b) CoCa: introducing image captioning to construct unidirectional fine-grained interaction. (c) Our SyCoCa: bidirectional local interation with attentive masking to enhance comprehensive cross-modal understanding.

struction, which enhances the fine-grained representation ability of images. In this way, SyCoCa has the bidirectional global and local interactions between modalities.

During the actual multimodal alignment process, although an image is worth one thousand words, people seldom write one thousand words to describe the image content. Instead, text descriptions or image captions are often highly abstract just focusing on the main character, object, or event in an image. For example, the caption of a picture about the family dinner of Thanksgiving would rarely delineate the furniture in the room in details. Moreover, even short text descriptions may imply rich and abundant image contents. *e.g.,*. *a football match* may well imply a vivid scene of a grass field, multiple palyers, and croweded audience, etc. Therefore, it is necessary to choose appropriate local regions and representation for alignment. To select effective image patches for text, we employ an attentive masking strategy. Specifically, we calculate the similarity between image tokens and text tokens to determine the relevance of image patches. For the IC task, the most pertinent image patches are selected to aid in generating textual descriptions by considering their semantic similarity with text descriptions. In contrast, we choose the *least* relevant image patches with the text on the TG-MIM task, aiming to leverage the text tokens to assist in recovering image content. Extensive experiments demonstrate the effectiveness of our proposed method on five downstream tasks.

In summary, the main contributions of our work are listed as follows:

- We first propose symmetrizing contrastive captioners for multimodal alignment, which improves the understanding between images and sentences from both global and local perspectives.
- To promote bidirectional local interactions, we adopt an attentive masking strategy to select appropriate image patches for IC and TG-MIM heads respectively.
- Thorough experiments validate that our proposed *SyCoCa* outperforms CoCa on several downstream tasks, *e.g.,* image-text retrieval, image-captioning, visual question answering and image classification. For example, we obtained +5.1%/3.7% in mTR/mIR on Flicker-30k compared to CoCa in image-text retrieval tasks.

2. Related Work

Vision-Language Pretraining. In recent years, there has been tremendous progress in multimodal alignment, especially between vision and language. Extensive researchers have dedicated their efforts to exploring vision-language pretraining. Early works (Tan & Bansal, 2019; Chen et al., 2020b; Zhang et al., 2021) prefixed a pretrained object detection modules to extract visual representations, which were then aligned with textual representations to achieve multimodal alignment. Later efforts focused on training multimodal transformers from scratch, such as ViLT (Kim et al., 2021) and VLMo (Bao et al., 2022). Pre-training of a foundation model on gigantic data with tremendous computation led to the breakthrough of multimodal alignment of images and texts. CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) trained a dual-encoder on large-scale noisy image-text pairs using contrastive loss, obtaining generic image and text representations for crossmodal alignment and zero-shot image classification tasks. Florence (Yuan et al., 2021) used a unified contrastive learning for various vision and image-text benchmarks. BLIP-2 (Li et al., 2023a) trained a Q-former to align a frozen vision encoder and language encoder. To enhance local interaction of images and texts, CoCa (Yu et al., 2022a) introduced a decoder for image caption based on CLIP.

In this paper, we symmetrize CoCa with an attentive TG-MIM task to further achieve bidirectional local interactions and enhances fine-grained alignment capabilities between modalities. The major difference between SyCoCa against bidirectional generation methods (Kim et al., 2022; You et al., 2023), which employ discrete auto-encoders to generate images from text, lies in our model's attentive TG-MIM task. This task enforces multimodal alignment constraints on patches that are paired with precise descriptions, since the abstract nature of image captions.

Masked Image Modeling. Beyond contrastive learning, masked image modeling (MIM) (Chen et al., 2020a; Doersch et al., 2015; Pathak et al., 2016; Trinh et al., 2019) emerges as another promising technique in vision pre-training. Recently, iGPT (Chen et al., 2020a), ViT (Dosovitskiy et al., 2020) and BEiT (Bao et al., 2021) recalled this learning approach on the modern vision transformers, which show great potential in representation learning by introducing special designs, such as clustering on pixels (Chen et al., 2020a), prediction of mean color (Dosovitskiy et al., 2020), and tokenization via an additional dVAE network with a block-wise masking strategy (Bao et al., 2021). SimMIM (Xie et al., 2022) predicted RGB values of raw pixels by direct regression. MAE (He et al., 2022) used an asymmetric encoder-decoder architecture, where the encoder operates only on the visible subset of patches, then a lightweight decoder reconstructs the original image from the latent representation and mask tokens. To enrich the joint representation learning in a multimodal context, BEiT-v3 (Wang et al., 2023a), EVE (Chen et al., 2023), MAMO (Zhao et al., 2023) and MaskVLM (Kwon et al., 2023) adopted the MIM task in multimodal pre-training and trained models to predict randomly masked image patches or text tokens.

Inspired by MAE, we propose a TG-MIM head to improve fine-grained multimodal alignment. The main distinction between SyCoCa against the masked multimodal modeling methods (e.g., BEiT-v3) encompasses: 1) we adopt IC to amplify the prominence of visual clues within the cross-modal textual representation; 2) we augment IC with the attentive masking mechanism refined the model's ability of comprehending the image from key perspectives; 3) we also apply attentive masking with TG-MIM to focus reconstruction efforts on appropriate patches that are contextually aligned with the description.

3. Method

We are interested in leveraging bidirectional cross-modal interactions to learn fine-grained visual and textual representations within an aligned latent space. Contrastive Captioner (CoCa) proposes to incorporate an image captioning task alongside the contrastive task, which learns fine-grained representation and establishes a stronger alignment across modalities than CLIP. It is worth noting that the modality interaction in CoCa is unidirectional and asymmetric, focusing solely on genetating text from images.

To enhance vision-language alignment, we introduce a novel framework called SyCoCa, which expands the existing CoCa model by incorporating bidirectional local interactions with attentive masking. Our framework comprises of three objectives: (i) image-text contrasting (ITC), (ii) image captioning (IC), and (iii) text-guided masked image modeling (TG-MIM). Figure 2 shows the overall framework of SyCoCa.

The goal of SyCoCa is to further explore the potential of cross-modal interactive prediction tasks. Building upon CoCa, SyCoCa incorporates TG-MIM to establish predictions from text to image, aiming to compensate for the inherent asymmetry in CoCa. This establishes bidirectional prediction between modalities, promoting fine-grained vision-language understanding. Furthermore, we design a novel attentive masking procedure that enables the bidirectional cross-modal interactions to focus on different regions of images, guided by the accompanying texts.

3.1. Model Architecture

As shown in Figure 2, the overall model architecture consists of four key components: the image encoder, text encoder, (text-to-)image decoder, and (image-to-)text decoder. Here, we provide a detailed explanation to each one.

Image Encoder. We employ vision transformer as the image encoder to model an input image. The image encoder takes image patches as input and encodes them into a sequence of embeddings $\{v^{cls}, v^1, ..., v^P\}$ where each embedding corresponds to a specific image patch. Additionally, an extra [CLS] embedding is included to provide a global representation of the image.

Text Encoder. As the text encoder, we adopt a causal masked transformer encoder to model text inputs. This encoder takes the simple BPE tokenized input text and converts it into a sequence vector represented as $\{w^1, ..., w^S, w^{cls}\}$, in which the embedding of the [CLS] token summarizes the global text feature. The purpose of adopting causal attention mask is to prevent any potential information leakage from future tokens to past tokens during the text encoding process.

Image / Text Decoder. To further capture the interaction between image-text pairs, we use an image decoder and a text decoder. Each decoder utilizes cross-attention transformer modules to deeply fuse image and text information, enabling

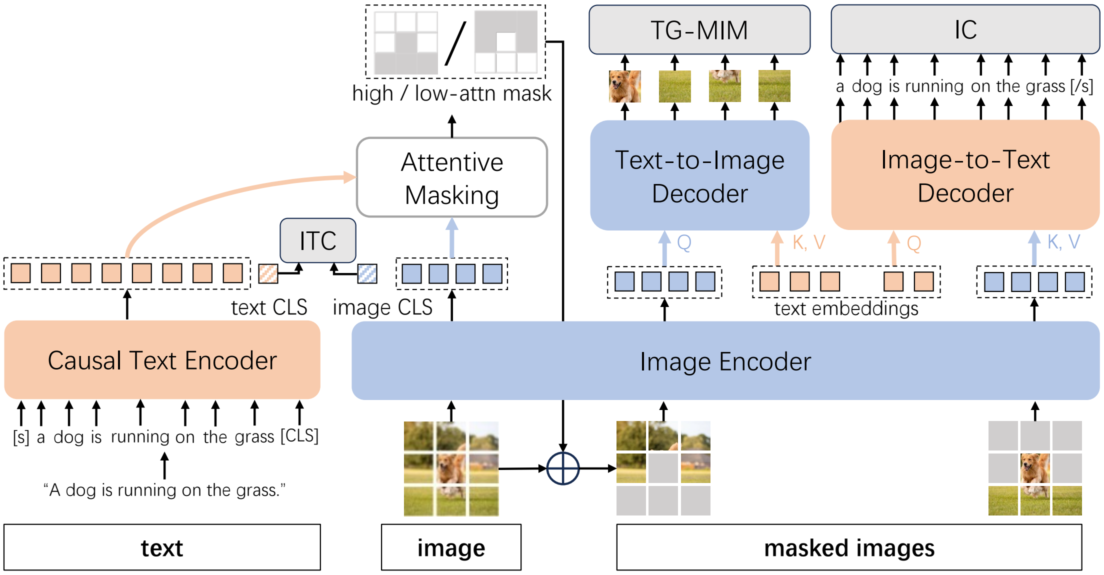


Figure 2. The detailed illustration of our proposed method. The framework of our method consist of four modules: an image encoder, a (causal) text encoder, a (text-to-)image decoder, a (image-to-)text decoder. Our method focuses on three pretraining objectives: image-text contrasting (ITC), text-guided masked image modeling (TG-MIM) and image captioning (IC).

bidirectional local interaction. The cross-modality multi-head attention module uses the representations of one modality as the query and the other modality's representations as the key and value. More specifically, the image decoder predicts the pixel values of masked image patches, leveraging cross-attention to incorporate relevant text. Similarly, the text decoder predicts the logits of the next token, utilizing cross-attention to integrate relevant image information. This bidirectional fusion mechanism ensures a comprehensive mutual understanding of the image-text pairs.

3.2. Training Objectives

We start from an image-text pair dataset consisting of pairs (I_i, T_i), where I_i represents an image and T_i is the corresponding text caption. The image encoder E_I and text encoder E_T are responsible for encoding the image and text, respectively, yielding the embeddings:

$$\{v_i^{cls}, v_i^1, ..., v_i^P\} = E_I(I_i) \quad \text{and} \tag{1}$$

$$\{w_i^1, ..., w_i^S, w_i^{cls}\} = E_T(T_i). \tag{2}$$

P refers to the number of patches, and S is the length of tokenized text sequence. Here, we present the applied training objectives in details.

Image Text Contrasting (ITC). In the training process, we consider a batch of N image-text pairs $\{I_i, T_i\}_{i=1}^N$ and their corresponding global representations $\{v_i^{cls}, w_i^{cls}\}_{i=1}^N$. It is assumed that these representations have been normalized to have a unit ℓ_2 norm. The contrasive objective is designed to align the image and text representations:

$$\begin{aligned}\mathcal{L}_{ITC} = &-\frac{1}{2N}\left[\sum_i^N \log\left(\frac{\exp(\langle v_i^{cls}, w_i^{cls}\rangle/\tau)}{\sum_{j=1}^N \exp(\langle v_i^{cls}, w_j^{cls}\rangle/\tau)}\right)\right] \\ &-\frac{1}{2N}\left[\sum_i^N \log\left(\frac{\exp(\langle w_i^{cls}, v_i^{cls}\rangle/\tau)}{\sum_{j=1}^N \exp(\langle w_i^{cls}, v_j^{cls}\rangle/\tau)}\right)\right],\end{aligned} \tag{3}$$

where $\langle\cdot,\cdot\rangle$ refers to the inner product, and τ is the temperature to scale the logits. The contrastive objective that pulls the representations of paired image-text close together while pushing apart unmatched pairs, promoting alignment in a shared semantic space.

Image Captioning. Image captioning objective requires the model to generate tokenized texts T_i with precise details in an autoregressive manner, compared with ITC treating the inputs as single entities. We use the image encoder E_I to encode image representations and train the text decoder D_T to maximize the conditional likelihood of the text T_i by utilizing forward autoregressive factorization:

$$\mathcal{L}_{IC} = -\sum_{j=1}^{|T_i|} \log D_T\left(T_i^j | T_i^{<j}, E_I(I_i)\right). \tag{4}$$

The text decoder is trained using parallel computation for enhanced learning efficiency. This training objective enables the model to learn fine-grained representations with a strong alignment through cross-modal prediction, fostering the acquisition of joint-modality representations useful for various multimodal understanding and generation tasks. Nonetheless, this interaction is unidirectional and asymmetrical, primarily focusing on understanding visual elements in language, without exploiting the comprehension of textual elements in the visual context.

Text-Guided Masked Image Modeling. To address the inherent imbalance in cross-modal interaction within the image captioning task, we design a novel training objective called text-guided masked image modeling. Our SyCoCa framework utilizes the image decoder, denoted as D_I, to reconstruct the masked image by predicting the pixel values for each masked patch. The output linear projection and reshape process is similar to that of the MAE. For a given image $I = \{p^1, ..., p^P\}$ consisting of P patches, we employ patch-wise masking using a masking map $M = \{m^1, ..., m^P\}$, where $m^i \in \{0, 1\}$ indicates whether the patch is masked. The TG-MIM loss is computed as the L1 loss between the masked patches and their reconstructed counterparts in the pixel space:

$$\{\hat{p}^1, ..., \hat{p}^P\} = D_I(E_I(I \oplus M), E_T(T)), \tag{5}$$

$$\mathcal{L}_{TM} = \sum_{i=1}^{P} m^i \cdot \parallel p^i - \hat{p}^i \parallel^1, \tag{6}$$

where $\hat{p}^i$ refers to the reconstructed patches. Note that the TG-MIM objective, proposed in this paper, distinguishes itself from the image-pretraining objective MIM by dedicated efforts on the guidance offered by the cross-attended text representation. The text-guided modeling focuses on the interaction between visual and textual representations, with the goal of enhancing the fine-grained understanding of textual elements within the visual context through reconstruction process. The introduction of this objective complements the uni-directional prediction of image captioning.

3.3. Attentive Masking

Intuitively, the reconstruction of image regions that reveal the same semantics or show highly related contents to the text shall be a more effective and robust cross-modal interaction, rather than reconstructing those irrelevant image regions. Therefore, we suggest generating masking maps that eliminate image patches that are relevant to language captions, and subsequently reconstruct these patches based on the text captions.

A straightforward idea to evalaute the correlation is by measuring the similarity between each patch embedding v^i and the global text embedding w^{cls}. However, w^{cls} provides a coarse-grained summary of the entire sentence. To mask the vision elements mentioned in the caption as accurately as possible, we propose to calculate the token-wise maximum similarity between image and text embeddings. This method is effective in modeling fine-grained cross-modal similarities in a previous study (Yao et al., 2021). Formally, for the i-th embedding v^i of an image, we calculate its similarities with all text embeddings $\{w^1, ..., w^S\}$ and select the highest similarity

$$s^i = \max_{j=1}^{S} \langle v^i, w^j \rangle \tag{7}$$

to represent its correlation with the text.

Image patches to be masked are selected based on their scores $\{s^1, ..., s^P\}$. To ensure stable training in TG-MIM, a fixed ratio r_h of top scoring patches (represented by the *high attn mask* in Figure 2) in the input image are masked. Conversely, in the caption task, the patches with the lowest scores (represented by the *low attn mask* in Figure 2) in the input image are masked at a fixed ratio r_l. The purpose is to enhance cross-modal interaction by encouraging the model to prioritize visually salient regions that play a significant role in understanding and representing the image for caption generation.

3.4. Overall Objective

Finally, we pretrain SyCoCa with all these losses combined:

$$\mathcal{L} = \mathcal{L}_{ITC} + \lambda_{IC}\mathcal{L}_{IC} + \lambda_{TM}\mathcal{L}_{TM}, \tag{8}$$

where λ_{IC} and λ_{TM} are the hyper-parameters weighting between IC and TG-MIM. All the modules of SyCoCa are trained from scratch.

4. Experiments

To demonstrate the effectiveness of our proposed SyCoCa, we conduct extensive experiments on 5 downstream tasks. Initially, we present the experimental setup, including the model architecture, pretraining datasets, downstream tasks, and implementation details. Subsequently, we compare SyCoCa with the baseline model, CoCa, on image-text retrieval, image classification, image captioning and visual question answering tasks. Finally, we perform a series of ablation studies to further analyze and evaluate our model.

4.1. Experimental Setup

Model Architecture. To ensure a fair comparison, we conduct our experiment using the open-source implementation of CoCa[1]. Both SyCoCa and the baseline model utilize the same CoCa-Base configuration for the image encoder, text encoder, and text decoder. Furthermore, in SyCoCa, we introduce a new image decoder that shares the same architecture as the text decoder yet with minor modifications. Specifically, we replace the token-prediction head with a

[1] https://github.com/mlfoundations/open_clip

Table 1. Zero-shot image-text retrieval evaluation results on Flickr30K and MSCOCO dataset.

	Flickr30K						MSCOCO					
	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
CoCa	44.2	72.4	81.7	31.3	59.7	70.8	16.3	37.1	48.9	15.3	35.8	47.0
Ours	**46.6**	**76.3**	**83.6**	**35.9**	**64.6**	**75.6**	**18.7**	**41.0**	**53.6**	**17.2**	**39.7**	**51.6**
%Gains	+5.4	+5.4	+2.3	+14.7	+8.2	+6.8	+14.7	+10.5	+9.6	+12.4	+10.9	+9.8

Table 2. Comparison with CoCa on image captioning (MSCOCO, NoCaps) and vision question answering (VQA). B: BLEU@4. M: METEOR. C: CIDEr. S: SPICE.

	Image Captioning										
	MSCOCO				NoCaps-Test				VQA		
	B@4	M	C	S	B@4	M	C	S	val	dev	test
CoCa	21.3	22.6	71.1	16.2	12.7	22.4	53.4	10.4	43.0	39.1	39.6
Ours	**22.4**	**23.6**	**75.6**	**16.8**	**12.9**	**22.8**	**54.8**	**10.5**	**46.9**	**42.6**	**42.9**
%Gains	+5.2	+4.4	+6.3	+3.7	+1.6	+1.8	+2.6	+1.0	+9.1	+9.0	+8.3

pixel-prediction head. All model parameters are initialized using a Gaussian distribution with a mean of 0 and a standard deviation of 0.01, allowing the training process to start from scratch.

Pretraining Data. We use the Conceptual Captions 12M (Changpinyo et al., 2021) (CC12M) dataset with 12 million image-caption pairs, as the multi-modal pretraining data for all models. Although this dataset is smaller in scale compared to the large custom datasets employed by the state-of-the-art pretraining models, such as 400M pairs in CLIP (Radford et al., 2021) and 3B pairs in CoCa (Yu et al., 2022a), it is a good match to the computation resources available to us. Moreover, CC12M has been widely adopted for benchmark evaluations in various studies on vision-language pretraining (Mu et al., 2022; Singh et al., 2022; Goel et al., 2022; Li et al., 2022c; Zhai et al., 2022). Additionally, we conduct a evaluation of our SyCoCa using a large-scale image-text dataset to enable a direct comparison with state-of-the-art methods. Detailed results and discussions of these supplementary experiments are provided in Appendix A.

Downstream Tasks. We evaluate SyCoCa on 3 downstream vision-language tasks and 2 classification tasks, including image-text retrieval, image captioning, visual question answering, zero-shot image classification, and fine-tuned image classification.

Implementation Details. We conduct our model training on two machines, each equipped with 8 NVIDIA A100 GPUs, for a total of 20 epochs. The batch size during training is set to 2048, and the resolution of pretraining images is set to 224×224. We use the AdamW optimizer (Loshchilov & Hutter, 2017) with an initial learning rate of $1e-4$. The learning rate schedule follows a cosine decay, including a warm-up period of 5000 steps. In terms of hyperparameters, we simply set $\lambda_{IC} = 2$ following CoCa and $\lambda_{TM} = 1$. The masking ratios r_h and r_l are both empirically set to 50%.

4.2. Results on Downstream Vision-Language Tasks

We compare the performance of SyCoCa and CoCa on downstream vision-language tasks, including image-text retrieval, image captioning, and vision question answering.

Image-Text Retrieval. In this task, the models are required to find the sample that best matches the input across modalities without finetuning. We conduct evaluation on standard image-text retrieval datasets, namely Flickr30K (Plummer et al., 2015) and MSCOCO (Lin et al., 2014), and report the results in Table 1. We can find that SyCoCa consistently outperforms CoCa showcasing the gains in the range of 5%-15% for R@1. Here, '%Gains' denotes the percentage of performance improvement over CoCa, namely:

$$\%\text{Gains} = \frac{\text{(improved perf. - original perf.)}}{\text{original perf.}} \times 100\% \quad (9)$$

Image Captioning. In this task, the models are required to generate textual descriptions for input images. We finetune both SyCoCa and CoCa using cross-entropy loss on the MSCOCO Captioning (Lin et al., 2014) dataset. Subsequently, we report the BLEU@4, METEOR, CIDEr, and SPICE scores on the Karparthy test split of MSCOCO, as well as the test split of the NoCaps (Agrawal et al., 2019) dataset. The results shown in Table 2 demonstrate that SyCoCa outperforms CoCa across all metrics. Specifically,

Table 3. Zero-shot image classification evaluation results on 8 coarse-grained datasets: ImageNet (Deng et al., 2009), ImageNet-V2 (Recht et al., 2019), ImageNet-A (Hendrycks et al., 2021b), ImageNet-R (Hendrycks et al., 2021a), CIFAR-10 (Krizhevsky & Hinton, 2009), CIFAR-100 (Krizhevsky & Hinton, 2009), STL-10 (Coates et al., 2011) and Caltech101 (Fei-Fei et al., 2004).

	IN-1K	IN-V2	IN-A	IN-R	C-10	C-100	STL-10	Caltech	Avg.
CoCa	26.1	22.4	6.5	31.7	**65.9**	29.7	88.1	64.5	41.9
Ours	**27.8**	**23.9**	**7.0**	**32.3**	62.0	**29.8**	**89.2**	**65.6**	**42.2**
%Gains	+6.5	+6.7	+7.7	+1.9	-5.9	+0.3	+1.2	+1.7	+2.5

Table 4. Fine-tuned image classification evaluation results on 6 fine-grained datasets: DTD (Cimpoi et al., 2014), Stanford Dogs (Khosla et al., 2011), CUB-200 (Wah et al., 2011), Flowers102 (Nilsback & Zisserman, 2008), MNIST (LeCun et al., 2010), Food101 (Bossard et al., 2014). Both models are transfered using linear probing.

	DTD	Dogs	CUB	Flowers	MNIST	Food	Avg.
CoCa	58.1	49.0	40.6	69.1	93.3	69.1	63.2
Ours	**60.5**	**51.2**	**41.3**	**70.6**	**94.5**	**71.1**	**64.9**
%Gains	+4.1	+4.5	+1.7	+2.2	+1.3	+2.9	+2.7

SyCoCa improves by 4%-6% on MSCOCO and 1%-3% on NoCaps compared to CoCa.

Vision Question Answering. In this task, the models are required to predict an answer based on both an image and a question. To apply SyCoCa and CoCa to this task, we modify the text encoder to a multi-modal decoder, which allows for the fusion of image representation and question input. We adapt the text decoder to generate answers in an autoregressive manner, derived from the output of the multi-modal decoder. To finetune the models, we utilize the VQA (Goyal et al., 2017) dataset. During inference, we constrain the decoder to generate answers only from a set of 3,128 candidate answers to make a fair comparison. The results in Table 2 clearly indicate that SyCoCa surpasses CoCa in all cases, where SyCoCa achieves remarkable improvements of 8%-9% on the validation, test-dev, and test splits of VQA.

Our evaluation on downstream vision-language tasks has confirmed the advancements achieved by SyCoCa in terms of multi-modal alignment and cross-modal understanding. On one hand, our zero-shot retrieval experiments have demonstrated the importance of bidirectional mutual interaction in enhancing the performance of CoCa in terms of modality alignment. On the other hand, the experimental results presented in Table 2 highlight the critical role played by the bidirectional interaction mechanism in fostering mutual understanding and capturing fine-grained elements across different modalities.

4.3. Results on Image Classification Tasks

We conduct a comparison between CoCa and SyCoCa regarding their classification performance on both zero-shot and fine-tuned image classification tasks. For zero-shot image classification, we evaluate the models on 8 coarse-grained image classification datasets that include common categories like airplanes and dogs. The results are presented in Table 3. SyCoCa outperforms CoCa in 7 out of 8 cases, resulting in an average accuracy improvement of 2.5%. These findings highlight the effectiveness of bidirectional understanding in bridging the gap between visual and textual representations.

In case of the fine-tuned image classification tasks, we use 6 fine-grained image classification datasets that encompass subcategories within a specific categorie, such as different breeds of dogs (Khosla et al., 2011) like Border Collie and Golden Retriever. During the fine-tuning process, we employ linear probing to gauge the image encoders' capability to discern intricate details within images. We report our results in Table 4. The results indicate that SyCoCa outperforms CoCa in all datasets, resulting in an average performance improvement of 2.7%. This demonstrates the effectiveness of incorporating bidirectional prediction to enhance the understanding and differentiation of fine-grained elements within the visual representation domain.

4.4. Qualitive Analysis of SyCoCa

To obtain an intuitive comprehension of the advantages of SyCoCa, we use Grad-CAM (Selvaraju et al., 2017), a commonly used ”visual explanation” toolkit, to generate attention location maps for the patch embedding layer in the image encoder. As shown in Figure 3, compared with CoCa, the improved version SyCoCa can capture fine-grained visual elements related to informative words in text. For instance, in the case of Image 1, SyCoCa exhibits the attention towards regions that are pertinent to the words ”*holding*”,

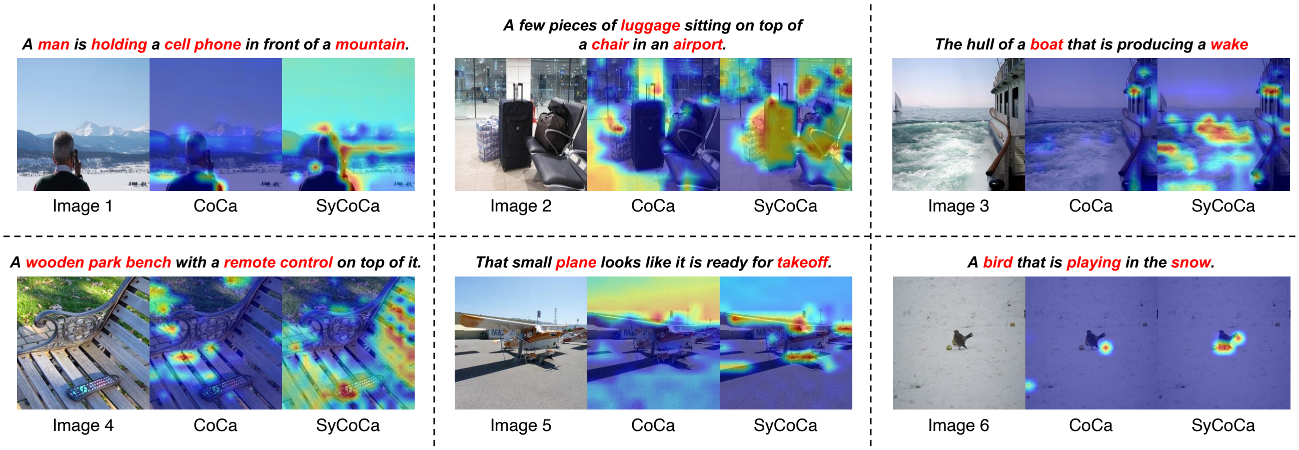

A man is holding a cell phone in front of a mountain.
Image 1
CoCa
SyCoCa
A few pieces of luggage sitting on top of a chair in an airport.
Image 2
CoCa
SyCoCa
The hull of a boat that is producing a wake
Image 3
CoCa
SyCoCa
A wooden park bench with a remote control on top of it.
Image 4
CoCa
SyCoCa
That small plane looks like it is ready for takeoff.
Image 5
CoCa
SyCoCa
A bird that is playing in the snow.
Image 6
CoCa
SyCoCa

Figure 3. Qualitative analysis of the proposed SyCoCa. We visualize the attention localization map of the first convolution layer in image encoder by the toolkit Grad-CAM.

Table 5. Ablation study on training objectives. ITC: image-text contrasting IC: image captioning. MIM: masked image modeling. TG-MIM: text-guided masked image modeling. RM: applying random masking in IC and MIM/TG-MIM. AM: applying attentive masking in IC and MIM/TG-MIM. mIR/mTR refers to the corresponding mean value of R@1, R@5 and R@10.

							zero-shot retrieval				zero-shot classification				
							Flickr30K		MSCOCO		IN-1K	C-10	STL-10	Caltech	
	ITC	IC	MIM	TG-MIM	RM	AM	mTR	mIR	mTR	mIR					Avg.
	✓						32.0	24.4	13.5	14.8	10.7	41.0	68.4	36.7	30.2
CoCa	✓	✓					37.5	28.7	16.1	16.0	10.4	39.6	69.2	37.8	31.9
	✓	✓				✓	41.1	31.4	17.5	18.3	11.4	40.0	71.7	39.1	33.8
	✓	✓	✓		✓		36.0	27.1	15.8	16.3	11.8	46.6	71.5	38.5	33.0
	✓	✓	✓			✓	37.7	27.8	15.2	16.2	11.6	43.5	72.1	35.9	32.5
	✓	✓		✓	✓		35.8	26.4	14.4	15.2	**11.9**	44.8	71.4	39.2	32.4
Ours	✓	✓		✓		✓	**42.6**	**32.4**	**18.3**	**18.4**	11.6	**51.1**	**73.6**	**39.5**	**35.9 (+12.5%)**

”*cell phone*” and ”*mountain*” whereas CoCa neglects these essential details. In contrast, in the case of Image 2, SyCoCa successfully identifies the objects ”*luggage*” and ”*chair*”, as well as the scenario ”*airport*”, whereas CoCa falls short in recognizing these elements. This enhanced performance is attributed to the fine-grained cross-modal understanding ability enabled by bidirectional local interaction and attentive masking, which proves advantageous for downstream tasks such as image captioning and vision language answering.

4.5. Ablation Study

We conduct a series of experiments to evalute the impact of training objectives and hyper-parameter settings of SyCoCa. Due to limited computing resources, we train the models on Conceptual Captions 3M (CC3M) (Sharma et al., 2018), which is a small dataset consisting of filtered image-text pairs. This dataset has been widely used for evaluations of vision-language pretraining (Yang et al., 2022b; Zhong et al., 2022; Wang et al., 2023b; Dong et al., 2023).

Objective Analysis. To analyze the impact of individual training objectives in SyCoCa, we conduct a series of experiments. We compare the performance of different variants on image-text retrieval and classification tasks, as presented in Table 5. We can observe that:

- By incorporating bidirectional cross-modal interaction tasks namely IC and TG-MIM, our proposed SyCoCa achieves an average improvement of 12.5% compared to CoCa on CC3M dataset.
- When applying a random masking mechanism to both IC and TG-MIM, the performance of models using TG-MIM slightly decreases compared to those using MIM. That is because the randomly masked patches may not be relevant to the text description, thereby introducing additional noise in the modal interaction and consequently affecting the performance of TG-MIM.

Table 6. Comparison of different masking ratios. mIR/mTR refers to the corresponding mean value of R@1, R@5 and R@10.

			zero-shot retrieval				zero-shot classification				
			Flickr30K		MSCOCO		IN-1K	C-10	STL-10	Caltech	
	r_l	r_h	mTR	mIR	mTR	mIR					Avg.
	25%	75%	40.1	31.0	17.5	17.8	11.1	49.7	73.3	39.4	35.0
(a)	50%	50%	42.6	32.4	18.3	18.4	11.6	51.1	73.6	39.5	35.9
	75%	25%	40.7	31.1	17.5	17.3	11.1	47.6	72.2	39.8	34.7
	25%	50%	42.5	32.4	18.3	18.3	11.1	48.0	73.9	39.4	35.5
(b)	50%	50%	42.6	32.4	18.3	18.4	11.6	51.1	73.6	39.5	35.9
	75%	50%	42.8	32.6	18.4	18.4	11.9	48.3	72.7	40.1	35.6
	50%	25%	42.4	32.3	18.3	18.3	11.5	49.9	74.1	36.5	35.4
(c)	50%	50%	42.6	32.4	18.3	18.4	11.6	51.1	73.6	39.5	35.9
	50%	75%	43.2	33.3	18.6	18.8	11.7	51.2	73.1	39.0	36.1

Table 7. Comparison of different weights for the TG-MIM objective. mIR/mTR refers to the corresponding mean value of R@1, R@5 and R@10.

	zero-shot retrieval				zero-shot classification				
	Flickr30K		MSCOCO		IN-1K	C-10	STL-10	Caltech	
λ_{TM}	mTR	mIR	mTR	mIR					Avg.
0.1	42.1	**32.4**	**18.3**	18.2	11.3	46.5	73.2	39.8	35.2
0.5	**42.8**	32.2	18.2	18.2	11.5	46.1	**73.6**	38.6	35.1
1.0	42.6	**32.4**	**18.3**	**18.4**	**11.6**	**51.1**	**73.6**	39.5	**35.9**
2.0	40.5	31.3	17.7	17.8	11.3	47.3	72.2	**40.0**	34.8

- Introducing TG-MIM alone, without the presence of AM, has little impact on performance. This is because attention masking is essential for TG-MIM to rely on text information in order to reconstruct highly correlated visual areas, thereby enhancing the cross-modal interaction between the vision and text modalities.

Impact of Masking Ratio. We train SyCoCa with different high/low masking ratios to evaluate their impact on model performance. Our evaluation is designed from three aspects: (a) dividing all image patches into high or low attentive masking, (b) maintaining a fixed low masking ratio while varying the high masking ratio, and (c) vice versa. The results are presented in Table 6. From Table 6 (a), we can observe that dividing the image patches evenly yields better results compared to biased partitioning, as it balances the visual information used in objectives in terms of image captioning and Tg-MIM. Moreover, in Table 6 (b)-(c) SyCoCa is not sensitive to changes in the mask ratio.

Training Objective Weights. We also investigate the impact of different weights for the TG-MIM, by varying the weights λ_{TM} assigned to the TG-MIM loss while keeping other objectives' weights fixed. The results in Table 7 indicate that within the specific range of $[0.1, 1.0]$, the variation of λ_{TM} has minimal impact on the performance. We select $\lambda_{TM} = 1.0$ for training our model.

5. Conclusion

In this paper, we introduce a novel vision-language pretraining method called SyCoCa, which aims to further enhance multi-modal alignment. Our approach focuses on improving the fine-grained understanding between vision and language modalities by introducing the text-guided masked image modeling (TG-MIM) training objective. By incorporating the TG-MIM training objective into the CoCa framework, we establish bidirectional local interaction, which leads to a precise and fine-grained alignment between the vision and language modalities. Additionally, we propose a new attentive masking approach for TG-MIM, which selectively masks image patches that have strong correlations with the text caption. By focusing on these highly relevant patches, we enhance the cross-modal interaction and significantly improve overall performance. Through extensive experiments on five vision-language tasks, we demonstrate the effectiveness and generalization ability of our SyCoCa.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Agrawal, H., Desai, K., Wang, Y., Chen, X., Jain, R., Johnson, M., Batra, D., Parikh, D., Lee, S., and Anderson, P. Nocaps: Novel object captioning at scale. In *ICCV*, pp. 8948–8957, 2019.

Bao, H., Dong, L., Piao, S., and Wei, F. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

Bao, H., Wang, W., Dong, L., Liu, Q., Mohammed, O. K., Aggarwal, K., Som, S., Piao, S., and Wei, F. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. In *NeurIPS*, pp. 32897–32912, 2022.

Bossard, L., Guillaumin, M., and Van Gool, L. Food-101–mining discriminative components with random forests. In *ECCV*, pp. 446–461, 2014.

Byeon, M., Park, B., Kim, H., Lee, S., Baek, W., and Kim, S. Coyo-700m: Image-text pair dataset. `https://github.com/kakaobrain/coyo-dataset`, 2022.

Changpinyo, S., Sharma, P., Ding, N., and Soricut, R. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, pp. 3558–3568, 2021.

Chen, J., Guo, L., Sun, J., Shao, S., Yuan, Z., Lin, L., and Zhang, D. Eve: Efficient vision-language pre-training with masked prediction and modality-aware moe. *arXiv preprint arXiv:2308.11971*, 2023.

Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., and Sutskever, I. Generative pretraining from pixels. In *ICML*, pp. 1691–1703. PMLR, 2020a.

Chen, Y.-C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., and Liu, J. Uniter: Universal image-text representation learning. In *ECCV*, pp. 104–120, 2020b.

Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., and Jitsev, J. Reproducible scaling laws for contrastive language-image learning. *arXiv preprint arXiv:2212.07143*, 2022.

Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., and Vedaldi, A. Describing textures in the wild. In *CVPR*, pp. 3606–3613, 2014.

Coates, A., Ng, A., and Lee, H. An analysis of single-layer networks in unsupervised feature learning. In *AISTATS*, pp. 215–223, 2011.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *CVPR*, pp. 248–255, 2009.

Doersch, C., Gupta, A., and Efros, A. A. Unsupervised visual representation learning by context prediction. In *ICCV*, pp. 1422–1430, 2015.

Dong, X., Bao, J., Zheng, Y., Zhang, T., Chen, D., Yang, H., Zeng, M., Zhang, W., Yuan, L., Chen, D., et al. Maskclip: Masked self-distillation advances contrastive language-image pretraining. In *CVPR*, pp. 10995–11005, 2023.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Fei-Fei, L., Fergus, R., and Perona, P. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *CVPRW*, pp. 178–178, 2004.

Goel, S., Bansal, H., Bhatia, S., Rossi, R., Vinay, V., and Grover, A. Cyclip: Cyclic contrastive language-image pretraining. In *NeurIPS*, pp. 6704–6719, 2022.

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., and Parikh, D. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, pp. 6904–6913, 2017.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. Masked autoencoders are scalable vision learners. In *CVPR*, pp. 16000–16009, 2022.

Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *ICCV*, pp. 8340–8349, 2021a.

Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., and Song, D. Natural adversarial examples. In *CVPR*, pp. 15262–15271, 2021b.

Hu, X., Gan, Z., Wang, J., Yang, Z., Liu, Z., Lu, Y., and Wang, L. Scaling up vision-language pre-training for image captioning. In *CVPR*, pp. 17980–17989, 2022.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., and Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, pp. 4904–4916, 2021.

Khosla, A., Jayadevaprakash, N., Yao, B., and Fei-Fei, L. Novel dataset for fine-grained image categorization. In *CVPRW*, June 2011.

Kim, T., Song, G., Lee, S., Kim, S., Seo, Y., Lee, S., Kim, S. H., Lee, H., and Bae, K. L-verse: Bidirectional generation between image and text. In *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16526–16536, 2022.

Kim, W., Son, B., and Kim, I. Vilt: Vision-and-language transformer without convolution or region supervision. In *ICML*, pp. 5583–5594, 2021.

Krizhevsky, A. and Hinton, G. Learning multiple layers of features from tiny images. 2009.

Kwon, G., Cai, Z., Ravichandran, A., Bas, E., Bhotika, R., and Soatto, S. Masked vision and language modeling for multi-modal representation learning. In *The Eleventh International Conference on Learning Representations*, 2023.

LeCun, Y., Cortes, C., and Burges, C. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

Li, C., Xu, H., Tian, J., Wang, W., Yan, M., Bi, B., Ye, J., Chen, H., Xu, G., Cao, Z., et al. mplug: Effective and efficient vision-language learning by cross-modal skip-connections. *arXiv preprint arXiv:2205.12005*, 2022a.

Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., and Hoi, S. C. H. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*, pp. 9694–9705, 2021.

Li, J., Li, D., Xiong, C., and Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, pp. 12888–12900, 2022b.

Li, J., Li, D., Savarese, S., and Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023a.

Li, L. H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.-N., et al. Grounded language-image pre-training. In *CVPR*, pp. 10965–10975, 2022c.

Li, X., Wang, Z., and Xie, C. Clipa-v2: Scaling clip training with 81.1 *arXiv preprint arXiv:2306.15658*, 2023b.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *ECCV*, pp. 740–755, 2014.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Mu, N., Kirillov, A., Wagner, D., and Xie, S. Slip: Self-supervision meets language-image pre-training. In *ECCV*, pp. 529–544, 2022.

Nilsback, M.-E. and Zisserman, A. Automated flower classification over a large number of classes. In *ICVGIP*, pp. 722–729, 2008.

Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., and Efros, A. A. Context encoders: Feature learning by inpainting. In *CVPR*, pp. 2536–2544, 2016.

Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., and Lazebnik, S. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, pp. 2641–2649, 2015.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *ICML*, pp. 8748–8763, 2021.

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do imagenet classifiers generalize to imagenet? In *ICML*, pp. 5389–5400, 2019.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *NIPS*, 35:25278–25294, 2022.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pp. 618–626, 2017.

Sharma, P., Ding, N., Goodman, S., and Soricut, R. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2556–2565, 2018.

Shen, S., Li, L. H., Tan, H., Bansal, M., Rohrbach, A., Chang, K.-W., Yao, Z., and Keutzer, K. How much can clip benefit visionand-language tasks? *arXiv preprint arXiv:2107.06383*, 3, 2021.

Singh, A., Hu, R., Goswami, V., Couairon, G., Galuba, W., Rohrbach, M., and Kiela, D. Flava: A foundational language and vision alignment model. In *CVPR*, pp. 15638–15650, 2022.

Sun, Q., Fang, Y., Wu, L., Wang, X., and Cao, Y. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023.

Tan, H. and Bansal, M. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

Trinh, T. H., Luong, M.-T., and Le, Q. V. Selfie: Self-supervised pretraining for image embedding. *arXiv preprint arXiv:1906.02940*, 2019.

Wah, C., Branson, S., Welinder, P., Perona, P., and Belongie, S. The caltech-ucsd birds-200-2011 dataset. 2011.

Wang, W., Bao, H., Dong, L., Bjorck, J., Peng, Z., Liu, Q., Aggarwal, K., Mohammed, O. K., Singhal, S., Som, S., et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022.

Wang, W., Bao, H., Dong, L., Bjorck, J., Peng, Z., Liu, Q., Aggarwal, K., Mohammed, O. K., Singhal, S., Som, S., et al. Image as a foreign language: Beit pretraining for vision and vision-language tasks. In *CVPR*, pp. 19175–19186, 2023a.

Wang, W., Yang, Z., Xu, B., Li, J., and Sun, Y. Vilta: Enhancing vision-language pre-training through textual augmentation. In *ICCV*, pp. 3158–3169, 2023b.

Wang, Z., Yu, J., Yu, A. W., Dai, Z., Tsvetkov, Y., and Cao, Y. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021.

Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., and Hu, H. Simmim: A simple framework for masked image modeling. In *CVPR*, pp. 9653–9663, 2022.

Xu, H., Ye, Q., Yan, M., Shi, Y., Ye, J., Xu, Y., Li, C., Bi, B., Qian, Q., Wang, W., et al. mplug-2: A modularized multi-modal foundation model across text, image and video. *arXiv preprint arXiv:2302.00402*, 2023.

Yang, A., Pan, J., Lin, J., Men, R., Zhang, Y., Zhou, J., and Zhou, C. Chinese clip: Contrastive vision-language pretraining in chinese. *arXiv preprint arXiv:2211.01335*, 2022a.

Yang, J., Li, C., Zhang, P., Xiao, B., Liu, C., Yuan, L., and Gao, J. Unified contrastive learning in image-text-label space. In *CVPR*, pp. 19163–19173, 2022b.

Yao, L., Huang, R., Hou, L., Lu, G., Niu, M., Xu, H., Liang, X., Li, Z., Jiang, X., and Xu, C. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*, 2021.

You, H., Guo, M., Wang, Z., Chang, K.-W., Baldridge, J., and Yu, J. Cobit: A contrastive bi-directional image-text generation model. *arXiv preprint arXiv:2303.13455*, 2023.

Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., and Wu, Y. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022a.

Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., and Wu, Y. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022b.

Yuan, L., Chen, D., Chen, Y.-L., Codella, N., Dai, X., Gao, J., Hu, H., Huang, X., Li, B., Li, C., et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021.

Zhai, X., Wang, X., Mustafa, B., Steiner, A., Keysers, D., Kolesnikov, A., and Beyer, L. Lit: Zero-shot transfer with locked-image text tuning. In *CVPR*, pp. 18123–18133, 2022.

Zhai, X., Mustafa, B., Kolesnikov, A., and Beyer, L. Sigmoid loss for language image pre-training. *arXiv preprint arXiv:2303.15343*, 2023.

Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., Choi, Y., and Gao, J. Vinvl: Revisiting visual representations in vision-language models. In *CVPR*, pp. 5579–5588, 2021.

Zhao, Z., Guo, L., He, X., Shao, S., Yuan, Z., and Liu, J. Mamo: Fine-grained vision-language representations learning with masked multimodal modeling. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1528–1538, 2023.

Zhong, Y., Yang, J., Zhang, P., Li, C., Codella, N., Li, L. H., Zhou, L., Dai, X., Yuan, L., Li, Y., et al. Regionclip: Region-based language-image pretraining. In *CVPR*, pp. 16793–16803, 2022.

A. Comparison with Other Methods

To verify the effectiveness of our proposed Text-Guided Masked Image Modeling (TG-MIM) and attentive masking (AM) strategy, the experiments are conducted under the same settings of CoCa and SyCoCa in the main paper. Especially for the training dataset, we used CC12M, a small dataset with 12 million image-text pairs. As we all know, the amount of training dataset has a crucial impact on model performance. To control data factors and reflect the superiority of our algorithm, we use a large-scale image-text pairs dataset for pre-training. Specifically, we collect Laion-2B (Schuhmann et al., 2022) and COYO-700M (Byeon et al., 2022). In addition, the number of parameters will also affect the model performance. To train the collected datasets, we use a total parameter size of 930M. Tab 8 summarizes the pre-training model configurations of the related methods. The comparison results between our SyCoCa and other methods are shown in Tab 9 and Tab 10.

Image-Text Retrieval. For the image-text retrieval task, the comparison results are shown in Tab 9. Specifically, we conduct this comparison using the *dual-encoder* configuration to appropriately reveal the performance of encoders. Our SyCoCa achieves the best results in terms of text-to-image retrieval on the Flickr30K dataset, as well as text-to-image/image-to-text retrieval on the MSCOCO dataset. It is worth noting that, SyCoCa outperforms ALIGN (Jia et al., 2021) in 11 out of 12 cases, despite both models having similar model parameter sizes and pre-training dataset sizes. In addition, even though some models (such as CoCa, BEiT-3, and SigLIP) employ larger parameter sizes and datasets, our SyCoCa still achieves comparable results. For instance, the mean text-to-image/image-to-text recall results for SyCoCa are 96.0/90.5/**82.3**/**69.8**, while for CoCa, they are **97.3**/**91.3**/81.4/69.1.

Image Caption. For image caption task, the comparison results are shown in Tab 10. On MSCOCO dataset, our SyCoCa is slightly lower than the best method on BLEU@4 and SPICE, but outperforms CoCa on all metrics. On NoCaps dataset, our method achieves the best performance, which shows that our attentive mask strategy has certain advantages in the caption task.

B. Discussion

B.1. Classification Performance

We note that, in Table 3, the improvements SyCoCa brings to classification tasks are subtler compared to its impact on vision-language tasks. We surmise that this could be attributed to the resolution of the test images. To investigate this, we randomly selected 100 images from the test subsets of each evaluation dataset and computed their average pixel count. The results are reported in Table 11.

Moreover, we have noted from Table 3 that SyCoCa exhibits a negative gain on CIFAR-10. To delve into this phenomenon, we constructed confusion matrices of classification results when applying both SyCoCa and CoCa to CIFAR-10, alongside their application to STL-10, which shares similar image categories. As depicted in Figure 4, SyCoCa tends to conflate visually akin categories (like trucks and automobiles) in the lower resolution (32×32) images of CIFAR. However, in the higher resolution (96×96) images of STL-10, SyCoCa demonstrates an enhanced ability to discriminate between visually similar categories.

The aforementioned outcomes suggest that SyCoCa's performance gains are constrained on low-resolution images. We conjecture that the image modeling task heightens the model's sensitivity to the loss of detail within image regions. Future work will be dedicated to further investigating this effect and enhancing the model's robustness to variations in input image resolution.

B.2. Training Cost

The training and inference costs for SyCoCa and CoCa, utilizing the CoCa-Base configuration, are depicted in Table 12. In the training stage, SyCoCa's adoption of an image decoder increases its parameter count by 30% relative to that of CoCa. Concurrently, the introduction of the TG-MIM task, along with the computations for attentive masked bidirectional local interactions, leads to a 67% uptick in the duration of training. While in the inference stage, it is noteworthy that SyCoCa maintains parity with CoCa in terms of parameter count and time consumption. This is because that the principal objective of SyCoCa is to enhance the encoder's efficacy in downstream tasks via enriched cross-modal interaction. In our future work, we aim to investigate light-weight decoder designs and optimize the training procedure to alleviate the training complexity.

Table 8. Pre-training model configurations. † WebLI is a private multilingual dataset conducted by Google, which consists of 10 billion images and 12 billion alt-texts.

Model	total #param.	precision	dataset	image size
CLIP (Radford et al., 2021)	430M	`fp16`	WIT-400M	336^2
OpenCLIP (Cherti et al., 2022)	430M	`bf16`	LAION-2B	224^2
ALIGN (Jia et al., 2021)	820M	-	ALIGN-1.8B	289^2
FILIP (Yao et al., 2021)	430M	`fp16`	FILIP-340M	224^2
Florence (Yuan et al., 2021)	890M	-	FLD-900M	224^2
CoCa (Yu et al., 2022b)	2.1B	-	JFT-3B+ALIGN-1.8B	288^2
CoCa-Large (Yu et al., 2022b)	790M	-	JFT-3B+ALIGN-1.8B	288^2
BEiT-v3 (Wang et al., 2022)	1.9B	-	15M images+160GB documents+21M pairs	224^2
L-Verse (Kim et al., 2022)	600M	`fp16`	ImageNet+MSCOCO+CC3M	256^2
CoBIT (You et al., 2023)	1.0B	-	JFT-4B+ALIGN-1.1B+WebLI-162M	256^2
EVA-CLIP-02 (Sun et al., 2023)	430M	-	LAION-2B+COYO-700M	224^2
SigLIP (Zhai et al., 2023)	430M	-	WebLI†	224^2
CLIPA-v2 (Li et al., 2023b)	1.0B	-	DataComp-1B	336^2
SyCoCa (ours)	930M	`bf16`	LAION-2B+COYO-700M	224^2

Table 9. Zero-shot image-text retrieval comparisons on Flickr30K (Plummer et al., 2015) and MSCOCO (Lin et al., 2014). Results of models that use significantly larger parameter sizes or dataset sizes are indicated in gray.

	Flickr30K						MSCOCO					
	Image → Text			Text → Image			Image → Text			Text → Image		
Model	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 (Radford et al., 2021)	88.0	98.7	99.4	68.7	90.6	95.2	58.4	81.5	88.1	37.8	62.4	72.2
OpenCLIP (Cherti et al., 2022)	88.7	98.4	99.2	75.0	92.5	95.6	62.1	83.4	90.3	46.1	70.7	79.4
ALIGN (Jia et al., 2021)	88.6	98.7	99.7	75.7	93.8	96.8	58.6	83.0	89.7	45.6	69.8	78.6
FILIP (Yao et al., 2021)	89.8	**99.2**	**99.8**	75.0	93.4	96.3	61.3	84.3	90.4	45.9	70.6	79.3
Florence (Yuan et al., 2021)	**90.9**	99.1	-	76.7	93.6	-	64.7	85.9	-	47.2	71.4	-
CoCa (Yu et al., 2022b)	92.5	99.5	99.9	80.4	95.7	97.7	66.3	86.2	91.8	51.2	74.2	82.0
BEiT-v3 (Wang et al., 2022)	94.9	99.9	100.0	81.5	95.6	97.8	-	-	-	-	-	-
CoBIT (You et al., 2023)	91.5	99.1	-	79.9	95.3	-	65.1	85.5	-	50.3	74.2	-
EVA-CLIP-02 (Sun et al., 2023)	89.7	98.6	99.2	77.3	93.6	96.8	63.7	84.3	90.4	47.5	71.2	79.7
SigLIP (Zhai et al., 2023)	-	-	-	-	-	-	70.2	-	-	52.0	-	-
CLIPA-v2 (Li et al., 2023b)	89.1	-	-	73.0	-	-	64.1	-	-	46.3	-	-
SyCoCa (ours)	89.2	99.1	99.6	**78.7**	**95.4**	**97.4**	**67.2**	**87.5**	**92.1**	**50.7**	**75.7**	**82.9**

Table 10. Image captioning comparisons on MSCOCO (Lin et al., 2014) and NoCaps (Agrawal et al., 2019). B@4: BLEU@4, M: METEOR, C: CIDEr, S: SPICE.

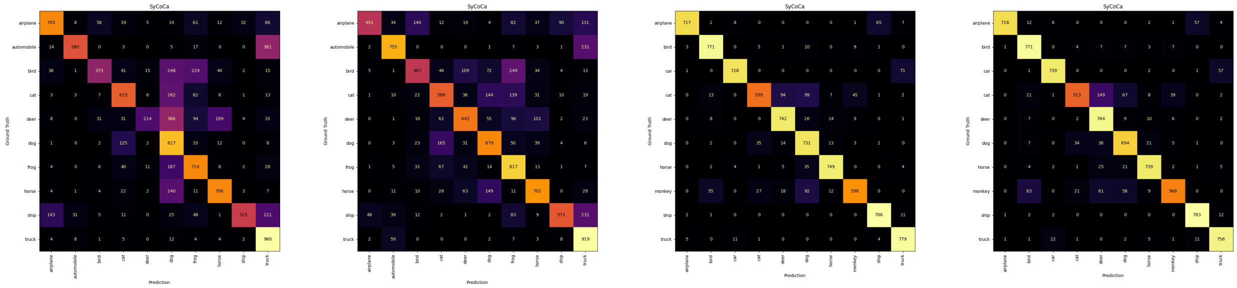
	MSCOCO				NoCaps			
					val		test	
Model	B@4	M	C	S	C	S	C	S
CLIP-ViL (Shen et al., 2021)	40.2	29.7	134.2	23.8	-	-	-	-
BLIP (Li et al., 2022b)	40.4	-	136.7	-	113.2	14.8	-	-
VinVL (Zhang et al., 2021)	41.0	31.1	140.9	**25.4**	105.1	14.4	103.7	14.4
SimVLM (Wang et al., 2021)	40.6	33.7	143.3	**25.4**	112.2	-	110.3	14.5
LEMON (Hu et al., 2022)	**41.5**	30.8	139.1	24.1	117.3	15.0	114.3	14.9
L-Verse (Kim et al., 2022)	39.9	31.4	102.2	23.3	-	-	-	-
CoBIT (You et al., 2023)	-	-	139.5	-	-	-	-	-
CoCa (Yu et al., 2022b)	40.9	33.9	143.6	24.7	122.4	15.5	120.6	15.5
SyCoCa (ours)	41.4	**34.1**	**143.7**	25.3	**122.6**	**15.8**	**121.1**	**15.6**

Table 11. Average image pixel count across datasets and the corresponding zero-shot classification performance gain of SyCoCa trained on CC12M.

Dataset	avg. #pixel	%Gains
ImageNet-1K	210843	+6.5
ImageNet-V2	207852	+6.7
ImageNet-A	188165	+7.7
ImageNet-R	212758	+1.9
CIFAR-10	1024	-5.9
CIFAR-100	1024	+0.3
STL-10	9126	+1.2
Caltech101	69410	+1.7

Table 12. Comparison of training and inference costs with CoCa-Base configuration. For inference, we report the parameter count and time consuming of image-text retrieval task.

	Training		Inference	
Method	#params.	sec./batch	#params.	sec./pair
CoCa	252M	0.50	151M	0.12
SyCoCa	328M (+30%)	0.83 (+67%)	151M	0.12



	airplane	automobile	bird	cat	deer	dog	frog	horse	ship	truck
airplane	705	8	58	19	5	14	61	12	32	86
automobile	14	580	0	3	0	5	17	0	0	381
bird	36	1	373	41	15	248	229	40	2	15
cat	3	3	7	615	6	262	82	8	1	13
deer	8	0	31	31	214	386	94	199	4	33
dog	1	0	2	125	2	817	33	12	0	8
frog	4	0	6	40	11	187	714	8	2	28
horse	4	1	4	22	2	240	11	706	3	7
ship	143	31	5	11	0	25	48	1	515	221
truck	4	8	1	5	0	12	4	4	2	960

(a) SyCoCa, CIFAR-10

	airplane	automobile	bird	cat	deer	dog	frog	horse	ship	truck
airplane	451	34	140	12	19	4	82	37	90	131
automobile	2	755	0	0	0	1	7	3	1	231
bird	5	1	467	46	109	72	249	34	4	13
cat	1	10	22	588	36	144	139	31	10	19
deer	0	1	18	62	642	55	96	101	2	23
dog	0	3	23	165	31	679	50	39	4	6
frog	1	5	33	67	42	14	817	13	1	7
horse	0	11	10	26	63	149	11	702	0	28
ship	48	39	12	2	1	2	83	9	573	231
truck	2	59	0	0	0	2	7	3	8	919

(b) CoCa, CIFAR-10

	airplane	bird	car	cat	deer	dog	horse	monkey	ship	truck
airplane	717	2	8	0	0	0	0	1	65	7
bird	3	771	0	5	1	10	0	9	1	0
car	1	0	728	0	0	0	0	0	0	71
cat	0	13	0	539	94	99	7	45	1	2
deer	0	8	0	0	742	26	14	9	0	1
dog	0	2	0	35	14	731	13	3	2	0
horse	0	2	4	1	5	35	749	0	0	4
monkey	0	55	0	27	18	92	12	596	0	0
ship	2	1	0	0	0	0	0	0	786	11
truck	5	0	11	1	0	0	0	0	4	779

(c) SyCoCa, STL-10

	airplane	bird	car	cat	deer	dog	horse	monkey	ship	truck
airplane	718	12	6	0	0	0	2	1	57	4
bird	1	771	0	4	7	7	3	7	0	0
car	0	1	739	0	0	0	2	0	1	57
cat	0	21	1	513	149	67	8	39	0	2
deer	0	7	0	2	764	9	10	6	0	2
dog	0	7	0	34	38	694	21	5	1	0
horse	0	4	2	1	25	21	739	2	1	5
monkey	0	83	0	21	61	58	9	568	0	0
ship	1	2	2	0	0	0	0	0	783	12
truck	1	1	22	1	0	2	5	1	11	756

(d) CoCa, STL-10

Figure 4. Confusion matrices for classification results: (a) SyCoCa applied to CIFAR-10, (b) CoCa applied to CIFAR-10 (c) SyCoCa applied to STL-10, and (d) CoCa applied to STL-10.