COMPOUND TOKENS: CHANNEL FUSION FOR VISION-LANGUAGE REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

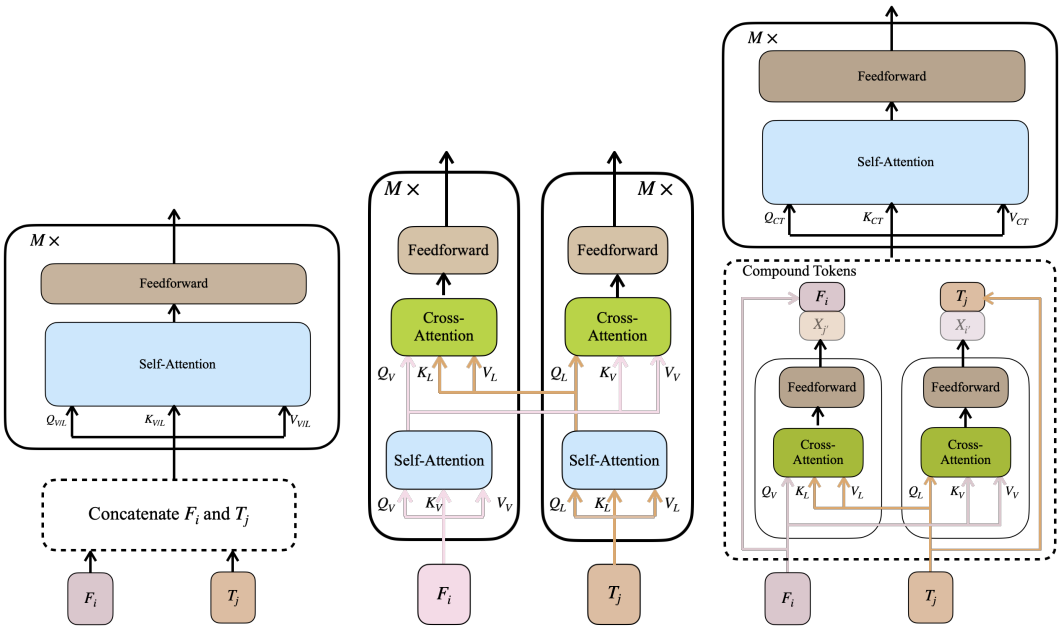
We present an effective method for fusing visual-and-language representations for several question answering tasks including visual question answering and visual entailment. In contrast to prior works that concatenate unimodal representations or use only cross-attention, we compose multimodal representations via channel fusion. By fusing on the channels, the model is able to more effectively align the tokens compared to standard methods. These multimodal representations, which we call compound tokens are generated with cross-attention transformer layers. First, vision tokens are used as queries to retrieve compatible text tokens through cross-attention. We then chain the vision tokens and the queried text tokens along the channel dimension. We call the resulting representations compound tokens. A second group of compound tokens are generated using an analogous process where the text tokens serve as queries to the cross-attention layer. We concatenate all the compound tokens for further processing with multimodal encoder. We demonstrate the effectiveness of compound tokens using an encoder-decoder vision-language model trained end-to-end in the open-vocabulary setting. Compound Tokens achieve highly competitive performance across a range of question answering tasks including GQA, VQA2.0, and SNLI-VE. We will make the code public.

1 INTRODUCTION

Multimodal learning will continue to play an increasingly fundamental role as we build increasingly more general purpose artificial agents. Tasks that seek information about visual inputs based on text queries such as visual question answering (VQA) (Goyal et al., 2017; Hudson & Manning, 2019) have emerged as effective frameworks for multimodal learning as they require a thorough understanding of both visual and textual information. For example, to correctly answer the question "what type of drink is to the right of the soda bottle?", a model must be able to distinguish a soda bottle from other bottles, left from right, understand language, and recognize the drink in question. Thus, effectively mixing or fusing the joint representations is critical for these tasks that have enjoyed so much progress in recent years (Li et al., 2021a; 2022; Wang et al., 2021; Yu et al., 2022; Wang et al., 2022b; Alayrac et al., 2022; Wang et al., 2022a).

One common strategy for fusing multimodal representations is to simply concatenate the vision and text tokens together, and feed them into a multimodal transformer encoder with self-attention layers. This approach, which we use as our main comparison reference is christened *merged attention*, and has been employed extensively in several vision-language models (Luowei Zhou, 2019; Hendricks et al., 2021; Dou et al., 2022; Piergiovanni et al., 2022a). A second multimodal fusion method called *co-attention* (shown in Figure 1), feeds the text and visual tokens separately into independent transformer encoders, and leverages cross-attention to communicate information between the two modalities (Tan & Bansal, 2019; Bugliarello et al., 2021; Hendricks et al., 2021; Li et al., 2021a; Dou et al., 2022).

While merge attention based models may struggle to align complementary tokens across different modalities effectively, co-attention based models forfeit the benefits of global self-attention across all tokens. Interestingly, Dou et al. (2022) observed a performance boost from co-attention compared to merged attention, suggesting a beneficial effect from the cross-attention mechanism. The downside of co-attention, however, is that it is parameter inefficient compared to merged attention as it requires



(a) Merged Attention (b) Co-Attention (c) Compound Tokens

Figure 1: **Different multimodal fusion methods**: Illustrations of two types of fusions methods in previous works: (a) co-attention, and (b) merged attention from the perspective of one visual token F_i, and one text token T_j. Our proposed compound tokens fusion method is illustrated in (c). Note that we use only one cross-attention layer for each modality compared to co-attention which uses both cross-attention and self-attention in all blocks. We concatenate the input query to the cross-attention module with the cross-attention output along the channel dimension. Q, K, and V denote the input query, keys and values respectively to the attention module. X represents the cross-attention layer's output. Finally, the subscripts V, L, and CT respectively identify an input as visual features, text features or compound tokens, e.g., Q_V indicates an input query that is composed of visual tokens.

separate sets of parameters for vision and language features. Our work unifies merge attention and co-attention in an efficient pipeline that produces more powerful multimodal representations than either method for several multimodal tasks.

Compound Tokens fusion aligns multimodal tokens using cross-attention without losing the advantages of global self-attention over all vision and text tokens. We use the tokens from one modality to query the other modality, and concatenate the output with the query tokens on the channels. An analogous process is repeated where we switch the roles of the two modalities. The resulting sets of compound tokens are concatenated and fed into a multimodal transformer encoder. Different from merged attention, we concatenate the vision and text tokens along the channel dimension. Unlike co-attention that uses both functions in every block, we use only two cross-attention functions at the beginning to facilitate channel concatenation.

Combining the query features and the cross-attention outputs on the channels (illustrated in Figure 1c) does not increase the token length, thus eliminating any additional computational or memory overheads in the multimodal transformer, and decoder modules. To further ensure that our method is efficient, we first embed each modality into half of their original feature dimension before compounding them. We show in Table 1 that other ways of mixing the input queries and the cross-attention outputs, such as weighting or element-wise product, are less effective compared to channel concatenation.

We evaluate compound tokens through extensive experiments in the challenging open-vocabulary setting via exact matching. In this setting, the generated responses must match exactly with the ground truth answers to be considered correct. This is notably more difficult than making predictions from a small predefined set of responses as in encoder only models. We adopt a generation

pipeline following works such as Cho et al. (2021); Wang et al. (2022b); and Piergiovanni et al. (2022a) that demonstrated the flexibility of that approach and its relevance to practical scenarios. However, as observed in Dou et al. (2022), this setting is less suitable for small models like ours. Accordingly, we provide separate results for VQA (Goyal et al., 2017) using an encoder only model. Even in the encoder only setting, our pretraining setup still uses an encoder-decoder architecture as in Piergiovanni et al. (2022a) and illustrated in Figure 2.

Compound Tokens fusion outperforms both merged attention and co-attention on GQA (Hudson & Manning, 2019), SNLI-VE (Xie et al., 2019), and VQA (Goyal et al., 2017) with and without vision-language pretraining. Our proposed fusion method obtained 82.87% on SNLI-VE beating METER (Dou et al., 2022) by 2.26%. Additionally, Compound Tokens recorded 82.43% on GQA significantly outperforming CFR (Nguyen et al., 2022) by 8.83%. Our VQA score of 70.62% on the VQA dataset is competitive among existing models.

To summarize, our work contributes a novel multimodal fusion method for vision-language tasks that enjoys the benefits of both cross-attention and self-attention without substantial additional computational overhead. We show the superiority of the proposed method over other fusion methods across several question answering tasks.

2 Related Works

Similar to the remarkable impact models such as T5 (Raffel et al., 2020), BERT (Devlin et al., 2019), and GPT-3 (Brown et al., 2020) have had on natural language process by pretraining on large amounts of text data, multimodal models like VilBERT (Lu et al., 2019), BEiT-3 Wang et al. (2022a), SimVLM (Wang et al., 2022b), Flamingo (Alayrac et al., 2022), and PaLI (Chen et al., 2022b) have increasingly shown significant advantages from pretraining on large scale and diverse multimodal data. Unsurprisingly, vision-and-language tasks including visual dialog (Das et al., 2017; Kottur et al., 2018; Chen et al., 2022a), visual reasoning (Suhr et al., 2017; Zellers et al., 2019), entailment (Xie et al., 2019; Chen et al., 2020), visual question answering (Antol et al., 2015; Goyal et al., 2017; Jiang et al., 2020; Wang et al., 2022b), caption generation (Anderson et al., 2018; Changpinyo et al., 2021), and cross-modality retrieval (Mao et al., 2016; Kamath et al., 2021) have all made great strides in recent years.

Important architectural innovations in vision-and-language models have been instrumental in accelerating these impressive scaling capabilities. One such innovation is the switch from expensive object detectors such as Faster-RCNN (Ren et al., 2015) in earlier models (Tan & Bansal, 2019; Lu et al., 2019; Li et al., 2019; 2020; Zhang et al., 2021) to simpler modules such as ResNet (He et al., 2015) or Vision Transformer (Dosovitskiy et al., 2021) for encoding visual features. Removing the object detectors reduced the need to train on clean human-annotated datasets such as Visual Genome (Krishna et al., 2016), thus paving the way to more impactfully use copious amounts of weakly-supervised image-text datasets from the internet.

Pretraining vision-and-language models with different cross-modal objectives has been another major axis of exploration in recent works. Contrastive learning (Li et al., 2021b; Yu et al., 2022), image captioning (Anderson et al., 2018), image-text matching (Lee et al., 2018; Lu et al., 2019), prefix language modeling (Wang et al., 2022b), word-patch alignment (Kim et al., 2021), etc., are some of the variety of losses proposed recently. Other works combine several losses during pretraining (Li et al., 2022; Dou et al., 2022), while still more methods unify several question answering tasks into a multi-task framework (Nguyen & Okatani, 2019; Lu et al., 2020; Piergiovanni et al., 2022a).

While a majority of works on vision-and-language representation learning have concentrated on improving feature extraction of the distinct modalities (e.g., using an object detector versus a convolutional neural network or a Transformer for vision feature extraction) or devising novel objective functions, efforts on improving the fusion of the multimodal representations have garnered relatively little attention. Most researchers simply adopt concatenation as described in merged attention (Dou et al., 2022) for fusion (Luowei Zhou, 2019; Piergiovanni et al., 2022a; Wang et al., 2022b). These methods differ sometimes on whether merging is done early in the model or at a deeper stage after processing each modality with large independent backbones. Co-attention is another popular method for mixing the multimodal features (Tan & Bansal, 2019; Li et al., 2021a; Dou et al., 2022). In co-

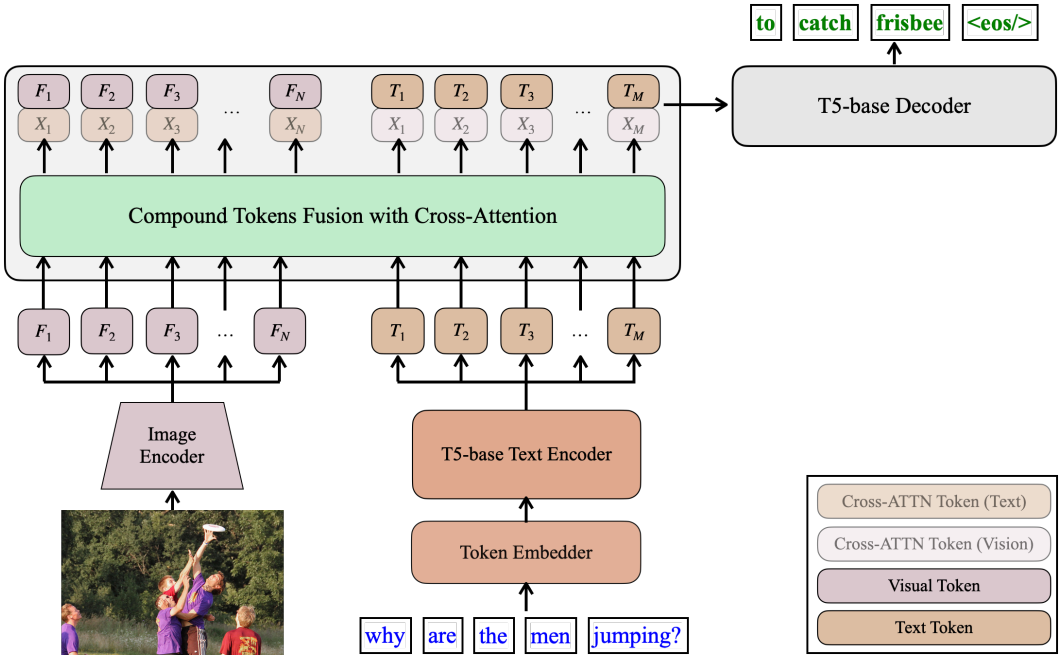


Figure 2: **Model Architecture:** Compound Tokens Fusion is illustrated in Figure 1c. ResNet-50 (He et al., 2015) and T5-base (Raffel et al., 2020) are used for the image and text encoders respectively.

attention, the vision and text features are modeled independently in separate transformer encoders, with a cross-attention mechanism serving as the bridge between the two modalities.

This work focuses on improving the multimodal fusion of the different representations for question answering tasks. As a result, we do not dwell heavily on the type of backbone encoders, the pre-training style nor the loss functions used. We propose a novel fusion of multimodal tokens through channel concatenation to encourage better alignment between the tokens, while preserving the capability to use self-attention over the joint representations. We use (1) a vision-to-text cross-attention, and (2) a text-to-vision cross-attention to retrieve more aligned representations. These aligned representations are concatenated with the query tokens on the feature dimension to form what we call compound tokens. The vision-to-text and text-to-vision compound tokens are merged and fed into a transformer encoder. See Figure 1 for illustrative comparisons between merged attention, co-attention and compound tokens fusion.

3 Compound Tokens

3.1 Background

We now provide a high level background on relevant functions necessary for the understanding of our method. We ignore layer normalization and multi-layer perceptrons in attention blocks in this overview for simplicity. For the same reason, we do not discuss residual connections between layers.

Attention: Given a set of query vectors $\mathbf{Q} \in \mathbb{R}^{N\times d}$ and a set of key vectors $\mathbf{K} \in \mathbb{R}^{M\times d}$, an attention layer gathers information from context vectors $\mathbf{V} \in \mathbb{R}^{M\times c}$ proportional to the normalized scores between the elements of $\mathbf{Q}$ and $\mathbf{K}$. Specifically, for softmax dot-product attention (Vaswani et al., 2017), the scalar output $z_{i,\ell}$, of an attention layer for query vector $q_i \in \mathbf{Q}$ and key vector $k_j \in \mathbf{K}$, is the weighted sum of the elements of $\mathbf{V}$,

$$a_{i,j} = \frac{q_i^T k_j}{\sqrt{d}} \qquad \alpha_{i,j} = \frac{\exp(a_{i,j})}{\sum_\ell \exp(a_{i,\ell})} \qquad z_{i,\ell} = \sum_j \alpha_{i,j} \mathbf{V}_{j,\ell} \,. \tag{1}$$

An attention mechanism is called self-attention when the query vectors are also members of the context vectors, i.e., $q_i \in \mathbf{K}\ \forall i$. It is known as cross-attention otherwise.

Multimodal Fusion: Token concatenation followed by self-attention is one of the most adopted approaches for cross-modal learning in recent vision-language architectures (Li et al., 2019; Piergiovanni et al., 2022a; Wang et al., 2022b; Chen et al., 2022b). Formally, given a sequence of N image tokens, $\mathcal{I} \in \mathbb{R}^{N \times d}$, and M text tokens, $\mathcal{T} \in \mathbb{R}^{M \times d}$, most methods concatenate $\mathcal{I}$ and $\mathcal{T}$ into a single representation $\mathcal{O} \in \mathbb{R}^{(N+M) \times d}$ which is then fed into a multimodal transformer for further modeling. The target outputs are produced using either a linear layer or a decoder. Besides concatenation, other methods such as (Tan & Bansal, 2019; Li et al., 2021a; 2022) use multimodal transformers composed of both self-attention and cross-attention in every block.

3.2 Proposed Fusion Method

Our method, illustrated in Figures 1c & 2, draws from both co-attention and merged-attention. Compound Tokens fusion first projects the visual and language tokens into half of the embedding space so that the total number of features is maintained after channel concatenation: $\widetilde{\mathcal{I}} \in \mathbb{R}^{N \times \frac{d}{2}}$; $\widetilde{\mathcal{T}} \in \mathbb{R}^{M \times \frac{d}{2}}$ for the image and text tokens respectively. Next, we employ only two cross-attention layers (unlike co-attention (Dou et al., 2022) that uses cross-attention and self-attention in every block) to create visual and language compound tokens

$$\widehat{\mathcal{I}} = \mathcal{A}\left(\widetilde{\mathcal{I}}, \widetilde{\mathcal{T}}, \widetilde{\mathcal{T}}\right) \quad \in \mathbb{R}^{N \times \frac{d}{2}} \qquad \widehat{\mathcal{T}} = \mathcal{A}\left(\widetilde{\mathcal{T}}, \widetilde{\mathcal{I}}, \widetilde{\mathcal{I}}\right) \quad \in \mathbb{R}^{M \times \frac{d}{2}} \tag{2}$$

$$\mathcal{I}_{cmpd} = \text{C-Concat}\left(\widetilde{\mathcal{I}}, \widehat{\mathcal{I}}\right) \quad \in \mathbb{R}^{N \times d} \qquad \mathcal{T}_{cmpd} = \text{C-Concat}\left(\widetilde{\mathcal{T}}, \widehat{\mathcal{T}}\right) \quad \in \mathbb{R}^{M \times d}, \tag{3}$$

where $\mathcal{A}(q, k, v)$ is the cross-attention function with q, k, and v as queries, keys, and values respectively. C-Concat(u, v) concatenates tensors u and v along the feature dimension. We combine vision-to-text compound tokens $\mathcal{I}_{cmpd}$, and text-to-vision compound tokens $\mathcal{T}_{cmpd}$, into a set of output compound tokens as in merged attention architectures

$$\mathcal{O}_{cmpd} = \text{Concat}\left(\mathcal{I}_{cmpd}, \mathcal{T}_{cmpd}\right) \in \mathbb{R}^{(N+M) \times d}. \tag{4}$$

Following previous methods, we feed $\mathcal{O}_{cmpd}$ into a self-attention multimodal encoder before generating the target outputs with an auto-regressive decoder. We also show results in Figure 3 and Table 2 where we do not use a multimodal encoder: $\mathcal{O}_{cmpd}$ is passed directly into the decoder to produce the outputs.

4 Experimental Setup

4.1 Model

We use ResNet-50 (He et al., 2015) as our image encoder and T5-base (Raffel et al., 2020) as our text encoder. The output of the image and text encoders are provided to our novel fusion method described in Section 3.2. A T5-base decoder consumes the output of the fusion module and generates free form text for all question answering tasks. The image encoder is pretrained on ImageNet-1k (Deng et al., 2009) while the text encoder and decoder use pretrained T5 weights.

4.2 Datasets

SNLI-VE (Xie et al., 2019) is a dataset of approximately 500,000 image-text pairs used for visual entailment (VE). Given an image and a proposed statement, the task for this dataset requires determining whether the statement is neutral, entails, or contradicts the image.

Visual Question Answering (VQA2.0) (Goyal et al., 2017) is a widely used benchmark for many question-answering models and contains 400,000 image-text pairs spanning 3,130 output categories. Each image-question pair is associated with 10 answers.

GQA (Hudson & Manning, 2019) is a vision question answering dataset of complex compositional questions comprising scene-object relations formed from Visual Genome (Krishna et al., 2016) with approximately 22 million question-answer pairs and 113 thousand images.

We emphasize that for all tasks, our model must generate a correct answer in an open-vocabulary setting of about 32,000 words irrespective of the number of categories in the task. A generated

response is counted as correct if and only if it matches exactly with the ground-truth answer. We use the VQA metric[1] for VQA2.0 and simple accuracy for GQA and SNLI-VE as evaluation metrics.

In addition to the downstream datasets, we also use CC3M[2] (Sharma et al., 2018) and COCO Captions (Lin et al., 2014) for pretraining. The pretraining setup uses a mixture of these datasets across four objectives: (1) **image-text matching** where the model predicts whether an image-text pair is a match or not, (2) **captioning** where the model generates the full caption given an image, (3) **caption completion** where the model completes a masked caption, and (4) **masked-language modeling** as in BERT (Devlin et al., 2019).

Unless otherwise stated, we pretrain our models for 300k iterations using a batch size of 512 and perform an additional 100k iterations of finetuning at a batch size of 128 on the downstream tasks: SNLI-VE, VQA, and GQA. When pretraining, the image resolution is set to 224×224 which is increased to 384×384 during finetuning or when training from scratch without vision-language pretraining (VLP). The input text length is set to 32. The output text length is 32 during pretraining and reduced to 8 during finetuning. Table 6 in the Appendix documents all our hyper-parameter settings including learning rates, weight-decay, etc. Generally, we use SNLI-VE and GQA for ablations as performance on those datasets in our setup is more stable than results on VQA.

5 Experimental Results

5.1 Why Channel Concatenation?

To determine the best way of composing compound tokens, we examined a number of options with a prime objective to not increase the token length. To this end, we sampled four combination methods and compared them on SNLI-VE and GQA as the performances on these datasets in our setup are more stable compared to VQA. Given input queries q and cross-attention layer's outputs X, we explored the following: (1) *channel concatenation* where we concatenate q and X along the feature dimension as described in Section 3.2. (2) *weighting* uses the operation $Y = \alpha q + \beta X$ where α and β are learnable scalars initialized randomly, and Y is the output. (3) In *Element-wise product*, $Y = q \odot X$. (4) Finally, we tested a simple summation of the tensors, $Y = q + X$. All these methods use approximately the same number of flops and parameters. The results in Table 1 show channel concatenation is better than the other methods, hence our use of channel concatenation in the rest of the paper.

Table 1: **Different Methods of Formulating Compound Tokens**: Channel concatenation obtains the highest accuracy on SNLI-VE and GQA.

Method	GFlops	SNLI-VE	GQA
Channel Concatenation	20.71	**80.85**	**80.79**
Weighting	20.71	80.63	80.61
Summation	20.71	80.75	80.35
Element-wise Product	20.71	80.81	78.31

5.2 Comparison of Compound Tokens with Merged Attention

We first compare merged attention and compound tokens fusion (our method) in Figure 3 without vision-language pretraining to establish some baseline results. We then incorporate vision-language pretraining and reassess the performance of each method in Table 2. All three downstream tasks for each fusion method uses the same pretrained model.

For these baseline comparisons, the fusion modules do not use a multimodal encoder. Merged attention simply feeds a concatenation of the multimodal tokens to the decoder while compound tokens fusion passes the tokens to the decoder immediately after channel chaining.

[1] https://visualqa.org/evaluation.html

[2] The version of the dataset we used has about 2 million samples

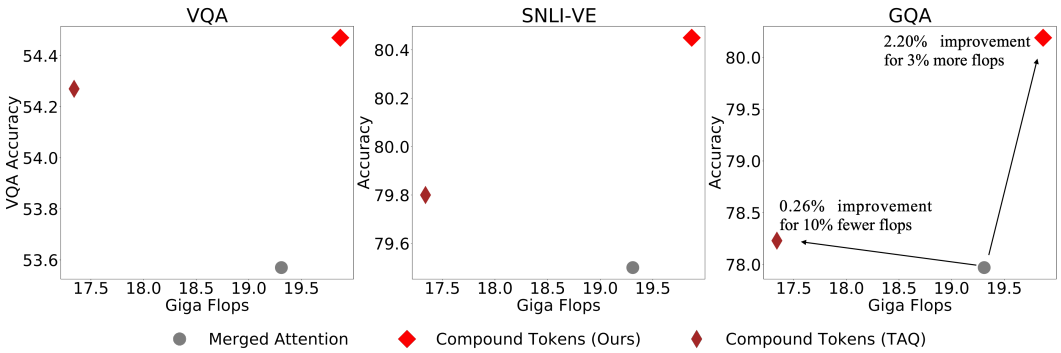


Figure 3: **Merged Attention versus Compound Tokens without Vision Language Pretraining:** With a relatively minimal amount of additional flops, Compound Tokens demonstrate a much improved performance over merged attention across all tasks. Compound Tokens (TAQ) is a more efficient version of our fusion method where we use only one cross-attention layer. For this method, we use the **T**ext tokens **A**s the input **Q**uery. We then compound the output with the text tokens along the channels as illustrated in Figure 1c. Note that even this more efficient compound tokens version outperforms merged attention, albeit with only marginal gains.

Table 2: **Merged Attention versus Compound Tokens *with* Vision Language Pretraining:** We repeat the experiments in Figure 3, but include vision-language pretraining on a mixture of CC3M and COCO captions. While pretraining generally increases performance across all methods, compound tokens continue to surpass merged attention, further underscoring the superiority of our method.

Fusion Method	GFlops	VQA	SNLI-VE	GQA
Merged Attention	19.31	53.33	81.25	78.25
Compound Tokens (Ours)	19.87	**57.51**	**81.49**	**80.45**
Compound Tokens (TAQ, Ours)	17.34	53.23	81.21	77.74

The results in Figure 3 and Table 2 show clearly that compound tokens fusion is superior to merged attention with and without vision-language pretraining at a relatively small amount of additional compute. This performance boost suggests that the use of cross-attention to align the multimodal representations is positively impactful on performance for various question answering tasks. When vision-language pretraining is employed, Compound Tokens fusion outperforms merged attention by substantial margins on VQA ($+4.18\%$) and GQA (2.20%). The improvement on SNLI-VE is a relatively modest 0.24%. Our method enjoys similar improvement margins when training from scratch without vision-language pretraining. We include a more efficient version of our method (Compound Tokens (TAQ)) where we use only the text tokens as queries to these baseline comparisons. Even this reduced capacity version of our method outperforms merged attention across all tasks when training from scratch while using fewer flops. The ablation in Table 7 in the Appendix shows that the ranking here is consistent across different image resolutions. In Table 9, we show that Compound Tokens also beat merged attention when a Vision Transformer (Dosovitskiy et al., 2021) is used as the image-encoder instead of the ResNet-50 we used in these experiments.

5.3 Multimodal Transformer Encoder

After establishing the superiority of compound tokens over merged attention when no multimodal encoder is used before the decoder, we expand the models to include a multimodal encoder with 12 self-attention blocks to match the setting in most previous vision-language models (Li et al., 2021a; Dou et al., 2022). We also compare with two other fusion methods Co-Attention (illustrated in Figure 1b), and Co-Tokenization (Piergiovanni et al., 2022b). Originally implemented for question answering tasks in videos, Co-Tokenization iteratively fuses visual features with text features using a TokenLearner (Ryoo et al., 2021). We use an adaptation of Co-Tokenization for images. The Co-Attention fusion module uses 6 blocks each for the vision and the text branches as in METER (Dou et al., 2022) where each block has a self-attention, cross-attention and feedforward layers. Co-

Table 3: **Comparisons with other Fusion models *without* Vision-Language Pretraining**: We extend the models to include a multimodal encoder with 12 self-attention layers in merged attention to match the typical setting in previous works. Compound Tokens outperform merged attention and Co-Attention with fewer parameters than both methods and fewer flops than merged attention. Co-Attention and merged attention are from Dou et al. (2022) while Co-Tokenization is from Piergiovanni et al. (2022b). The results here are our implementations of the these methods. Params shows the number of parameters in the entire model (not just the fusion module); RES is the image resolution and L is the total number of transformer blocks in the multimodal encoder: Compound Tokens uses two cross-attention blocks before the multimodal encoder.

Fusion Method	L	Params ($\times 10^6$)	RES	GFlops	SNLI-VE	GQA
Merged Attention	12	332.94	384×384	34.89	79.81	78.07
Co-Attention	12	361.26	384×384	29.61	80.20	77.75
Compound Tokens (Ours)	10	325.82	384×384	32.90	**80.52**	**78.21**
Co-Tokenization	12	392.14	384×384	57.78	**80.79**	**81.07**
Compound Tokens (Ours)	10	325.82	384×384	32.90	80.52	78.21

Tokenization uses 64 image tokens and 4 transformer blocks per each tokenization round. There are 3 tokenization rounds, constituting 12 self-attention blocks overall. A self-attention block in our implementation is made up of a self-attention function and a feedforward layer. The multimodal encoder for Compound Tokens fusion has 10 blocks to compensate for the two cross-attention blocks that it uses.

The results of these experiments are shown in Table 3. The models are trained for 300k iterations at a batch size of 128 on each downstream task without any vision-language pretraining (See Table 8 in the Appendix for results with pretraining across different resolutions). Compound Tokens fusion continues to outperform merged attention and co-attention in this setting as well, indicating the fusion mechanism remains competitive even when a multimodal encoder is used. However, it slightly underperforms the more expensive Co-Tokenization module when training from scratch.

5.4 An Encoder only Model for VQA

The performance of our models on VQA in the encoder-decoder setup is significantly lower than reported results in previous works even for small models like ours. We note again that this suboptimal performance is not unique to compound tokens fusion; we observe similar low values for all the fusion methods we tested in our architectural setup (See Table 10 in the Appendix for VQA results in the encoder-decoder architecture for all fusion methods.). We believe the low performance is an effect of the decoder not being able to learn the VQA vocabulary sufficiently. To address any problems introduced by the decoder, we use an encoder only model for the VQA task during finetuning by replacing the decoder in a pretrained model with a linear layer of size 3130. The results in Table 4 show that the encoder only model significantly outperforms the encoder-decoder model. We, thus, adopt that setup in our comparison with previous work discussed in the next section. The VQA metric is still used for evaluation in the encoder only model.

Table 4: **Encoder only versus Encoder-Decoder**: The Encoder only model outperforms the encoder-decoder model by a large margin. The models here use the same pretrained encoder-decoder model: the decoder is replaced with a linear classifier when transitioning to an encoder-only version. We finetune both models for 100k steps after pretraining for 300k steps.

Fusion Method	Architecture	GFlops	VQA Accuracy
Compound Tokens	Encoder-Decoder	35.50	58.14
Compound Tokens	Encoder Only	31.77	70.39

5.5 Comparison with Existing Approaches

Finally, we compare our results with various competitive recent models such as METER (Dou et al., 2022), ALBEF (Li et al., 2021a), and CFR (Nguyen et al., 2022). The models in Table 5 generally have approximately the same number of parameters, but differ significantly on the pretraining datasets, pretraining objectives, and backbone encoders. For example, while we use Conceptual Captions (Sharma et al., 2018) and COCO (Lin et al., 2014) as our pretraining datasets, METER used Conceptual Captions, COCO, Visual Genome (Krishna et al., 2016) and SBU Captions (Ordonez et al., 2011). ALBEF used all the datasets in METER in addition to Conceptual Captions 12M (Changpinyo et al., 2021).

The model we use for this comparison has 340 million parameters in total. We pretrain it for 500k iterations with a batch-size of 512 using an image resolution of 224×224 and further finetune for 200k iterations on each of the downstream tasks at resolution 384×384 with batch size 128. This model uses a multimodal encoder with 12 blocks.

Except for SimVLM (Wang et al., 2022b) which has about 1.5 billion parameters and uses a significantly large pretraining data (a 1.8 billion private dataset), our model outperforms all other methods on SNLI-VE and GQA by large margins. We are confident that further pretraining and increasing image resolution will improve our already competitive result on the VQA dataset. Scaling up the model may also yield additional performance improvements.

Table 5: **Comparison with SOTA:** Compound Tokens outperforms all other models on SNLI-VE and GQA in an open-vocabulary evaluation except SimVLM (Wang et al., 2022b) which used a privat dataset of 1.5B samples. For VQA, we present the results in the closed-vocabulary setting for fair comparisons with the other methods: our open-set evaluation is significantly worse than the closed-set evaluation model on this task. The best values among the models besides SimVLM are in **bold**. The second best values are underlined. *The flops are based on our calculations. Our model is extremely more efficient than the rest partly because we use a short text sequence length of 32 and a ResNet-50 backbone that produces 49 visual tokens.

Approach	Params	GFlops*	VQA	SNLI-VE	GQA
SimVLM$_{Huge}$ (Wang et al., 2022b)	1.5B	890	*80.34*	*86.32*	-
VisualBERT (Li et al., 2019)			66.70	75.69	-
UNITER (Chen et al., 2020)			73.82	79.39	-
LXMERT (Tan & Bansal, 2019)			69.90	-	60.00
ALBEF (Li et al., 2021a)	418M	122	75.84	80.91	-
METER (Dou et al., 2022)	336M	130	**77.68**	80.61	-
BLIP (Li et al., 2022)	475M	122	77.54	-	-
12-in-1 (Lu et al., 2020)			71.30	-	60.50
VinVL (Zhang et al., 2021)			75.95	-	65.05
VL-T5 (Cho et al., 2021)			70.30	-	60.80
CFR (Nguyen et al., 2022)			69.80	-	73.60
Compound Tokens (Ours)	340M	36	70.62	**82.87**	**82.43**

6 Conclusion

We introduce Compound Tokens, a new multimodal fusion method for vision-and-language representation learning. Our method beat super competitive models such as ALBEF and METER on SNLI-VE by close to 2%. Furthermore, Compound Tokens performance on GQA beats the next best model we are aware of by more than 8 percentage points. Finally, we demonstrated through numerous comparative experiments that our method is better than merged attention and co-attention across three popular question answering tasks. We consistently outperformed these standard methods with and without pretraining on image-text pairs, across different image resolutions and image encoding backbones.

With this strong demonstration as an effective fusion method, we hope that Compound Tokens will inspire other methods of modeling multimodal representations beyond token concatenation.

7 CODE OF ETHICS AND REPRODUCIBILITY STATEMENTS

We describe the datasets we used in Section 4.2, our main model in 4.1, and hyper-parameter in Sections 4 & A. We also list all hyper-parameters for the ablation experiments in Table 6 to enhance reproducibility of our method. Finally, we will make the code public to the research community.

All datasets used in this work are publicly available. The work did not involve any human subjects and has no immediate harmful insights or implications for society. We are confident that our fusion method will help advance multimodal learning and machine learning at large.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning, 2022. URL https://arxiv.org/abs/2204.14198.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Emanuele Bugliarello, Ryan Cotterell, Naoaki Okazaki, and Desmond Elliott. Multimodal Pretraining Unmasked: A Meta-Analysis and a Unified Framework of Vision-and-Language BERTs. *Transactions of the Association for Computational Linguistics*, 9:978–994, 09 2021. ISSN 2307-387X. doi: 10.1162/tacl_a_00408. URL https://doi.org/10.1162/tacl_a_00408.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, 2021.

Cheng Chen, Yudong Zhu, Zhenshan Tan, Qingrong Cheng, Xin Jiang, Qun Liu, and Xiaodong Gu. Utc: A unified transformer with inter-task contrastive learning for visual dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022a. URL https://arxiv.org/abs/2205.00423.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme, Andreas Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu

Soricut. Pali: A jointly-scaled multilingual language-image model, 2022b. URL https://arxiv.org/abs/2209.06794.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *ECCV*, 2020.

Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 1931–1942. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/cho21a.html.

Ekin Dogus Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V. Le. Autoaugment: Learning augmentation policies from data. In *CVPR*, 2019. URL https://arxiv.org/pdf/1805.09501.pdf.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M.F. Moura, Devi Parikh, and Dhruv Batra. Visual Dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, 2019. URL https://aclanthology.org/N19-1423.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, Zicheng Liu, and Michael Zeng. An empirical study of training end-to-end vision-and-language transformers. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. URL https://arxiv.org/abs/2111.02387.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.

Lisa Anne Hendricks, John Mellor, Rosalia Schneider, Jean-Baptiste Alayrac, and Aida Nematzadeh. Decoupling the role of data, attention, and losses in multimodal transformers. *Transactions of the Association for Computational Linguistics*, 9:570–585, 2021. URL https://aclanthology.org/2021.tacl-1.35.

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

Huaizu Jiang, Ishan Misra, Marcus Rohrbach, Erik Learned-Miller, and Xinlei Chen. In defense of grid features for visual question answering. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10264–10273, 2020.

Aishwarya Kamath, Mannat Singh, Yann LeCun, Ishan Misra, Gabriel Synnaeve, and Nicolas Carion. Mdetr–modulated detection for end-to-end multi-modal understanding. *arXiv preprint arXiv:2104.12763*, 2021.

Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pp. 5583–5594, 2021. URL http://proceedings.mlr.press/v139/kim21k.html.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. URL http://arxiv.org/abs/1412.6980.

Satwik Kottur, José M. F. Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. Visual coreference resolution in visual dialog using neural module networks. In *The European Conference on Computer Vision (ECCV)*, 2018.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. In , 2016. URL https://arxiv.org/abs/1602.07332.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *ECCV*, 2018.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 9694–9705. Curran Associates, Inc., 2021a. URL https://proceedings.neurips.cc/paper/2021/file/505259756244493872b7709a8a01b536-Paper.pdf.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. URL https://arxiv.org/pdf/2201.12086.pdf.

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

Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. UNIMO: Towards unified-modal understanding and generation via cross-modal contrastive learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 2592–2607. Association for Computational Linguistics, 2021b. URL https://aclanthology.org/2021.acl-long.202.

Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. *ECCV 2020*, 2020.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and Larry Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Skq89Scxx.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/c74d97b01eae257e44aa9d5bade97baf-Paper.pdf.

Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Lei Zhang Houdong Hu Jason J. Corso Jianfeng Gao Luowei Zhou, Hamid Palangi. Unified vision-language pre-training for image captioning and vqa. *arXiv preprint arXiv:1909.11059*, 2019.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, 2016.

Binh X. Nguyen, Tuong Do, Huy Tran, Erman Tjiputra, Quang D, and Anh Nguyen Tran. Coarse-to-fine reasoning for visual question answering. In *Multimodal Learning and Applications (MULA) Workshop, CVPR*, 2022.

Duy-Kien Nguyen and Takayuki Okatani. Multi-task learning of hierarchical vision-language representation. In *CVPR*, 2019. URL https://arxiv.org/abs/1812.00500.

Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011. URL https://proceedings.neurips.cc/paper/2011/file/5dd9db5e033da9c6fb5ba83c7a7ebea9-Paper.pdf.

AJ Piergiovanni, Wei Li, Weicheng Kuo, Mohammad Saffar, Fred Bertsch, and Anelia Angelova. Answer-me: Multi-task open-vocabulary visual question answering. In *European Conference on Computer Vision (ECCV)*, 2022a. URL https://arxiv.org/abs/2205.00949.

AJ Piergiovanni, Kairo Morton, Weicheng Kuo, Michael S. Ryoo, and Anelia Angelova. Video question answering with iterative video-text co-tokenization. In *ECCV*, 2022b. URL https://arxiv.org/abs/2208.00934.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper/2015/file/14bfa6bb14875e45bba028a21ed38046-Paper.pdf.

Michael Ryoo, AJ Piergiovanni, Anurag Arnab, Mostafa Dehghani, and Anelia Angelova. Tokenlearner: Adaptive space-time tokenization for videos. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 12786–12797. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/6a30e32e56fce5cf381895dfe6ca7b6f-Paper.pdf.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*, 2018.

Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. A corpus of natural language for visual reasoning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 217–223. Association for Computational Linguistics, 2017. URL https://aclanthology.org/P17-2034.

Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Wenhui Wang, Hangbo Bao, Li Dong, and Furu Wei. VLMo: Unified vision-language pre-training with mixture-of-modality-experts, 2021.

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. Image as a foreign language: Beit pretraining for all vision and vision-language tasks, 2022a. URL https://arxiv.org/abs/2208.10442.

Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. In *International Conference on Learning Representations (ICLR)*, 2022b. URL https://arxiv.org/abs/2108.10904.

Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*, 2019.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models, 2022. URL https://arxiv.org/abs/2205.01917.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5579–5588, June 2021.

APPENDIX

A HYPER-PARAMETER SETTINGS

We provide full details of our hyper-parameter settings in this section. We use Adam (Kingma & Ba, 2015) to optimize all our models. The learning rate starts from zero and warms up linearly to the base rate after 8k iterations. Cosine annealing (Loshchilov & Hutter, 2017) with a cycle rate of 100k steps is then used to decay the rate to zero by the end of training. We use gradient clipping with a maximum norm of 1.0 in all our experiments.

We do not use any data augmentation beyond resizing and normalization in all the ablation experiments and finetuning experiments. We apply random cropping and AutoAugment (Cubuk et al., 2019) during pretraining of our main model.

All our pretraining experiments use a batch size of 512 and image resolution 224×224. The batch size is divided equally among the four pretraining objectives: image captioning, caption completion, image-text matching, and masked language modeling. We also sample the same number of examples from CC3M and COCO in every iteration. The batch size and resolutions are set to 128, and 384×384 respectively whenever training from scratch or finetuning.

The datasets we used and our model are described in Section 4. The rest of the hyper-parameters are listed in Table 6.

Table 6: **Hyper-parameter Settings**: We enumerate the hyper-parameters for our ablation experiments and main model. L is the number of blocks in a multimodal encoder. Main Model is the model we used in Table 5 for comparison with existing works.

Experiment	Phase	L	Iterations	LR	Dropout	Weight Decay
Ablations	Pretraining	0 / 12	300k	$1.1e^{-4}$	$1e^{-3}$	0.1
	Finetuning	0 12	100k	$5e-5$ $3.1e^{-3}$	0 $1e^{-3}$	$1e^{-4}$
	Scratch	0 12	300k	$7.5e^{-5}$ $3e^{-5}$	$1e^{-2}$	$1e^{-3}$
Main Model	Pretraining Finetuning	12	500k 200k	$1.1e^{-4}$ $3e^{-5}$	$1e^{-3}$	0.1 $1e^{-4}$

B Further Ablations

B.1 Image Resolution

Increasing image resolution generally leads to better performance for various question answering tasks. As a consequence, most prior works use a larger resolution during finetuning compared to the pretraining resolution. For example, Wang et al. (2022b) pretrained at resolution 224×224 and finetuned at 480×480. In this work, we followed the setting in METER (Dou et al., 2022) by pretraining and finetuning at resolutions 224×224 and 384×384 respectively. We now investigate whether Compound Tokens also enjoy improved performance relative to merged attention at different resolutions in this section.

The results of this ablation is shown in Table 7 for models without a multimodal encoder and in Table 8 for models with a multi-modal encoder. The models in Table 7 do not use any pretraining on paired image-text data while the models in Table 8 are pretrained on CC3M and COCO for 300k iterations. As in prior works, increasing image resolution improves performance across all fusion methods and datasets. In all cases, Compound Tokens continue to outperform merge attention, further underlining the fact that our proposed method is more effective than traditional merge attention.

Table 7: **Impact of Image Resolution without Vision-Language Pretraining**: Increasing the resolution increases performance for both merged attention and compound tokens, with compound tokens continuing to outperform merged attention at both resolutions. **Bold** numbers shows the best results within each comparison setting.

Fusion Method	RES	GFlops	SNLI-VE	GQA
Merged Attention	224×224	9.94	78.70	75.62
Compound Tokens	224×224	10.22	**79.59**	**76.62**
Merged Attention	384×384	19.31	79.15	76.66
Compound Tokens	384×384	19.87	**80.44**	**79.02**

B.2 Type of Image Encoder

The image encoder is an important component in vision-language models. While earlier models used object detectors such as Faster-RCNN, more recent models use either a CNN (Lecun et al., 1998) or a Vision Transformer (ViT) (Vaswani et al., 2017) for image feature extraction. We used ResNet-50 for our main experiments and investigate the impact of using a transformer as the image encoder in this ablation. The results of using using a ViT as the image encoder is shown in Table 9. All models in that experiment use 224×224 as the image resolution. A patch size of 16×16 was used for the ViT based models. The ViT based models perform slightly less than the comparable ResNet based models. As is the case with using ResNet-50, Compound Tokens fusion remains superior to merged attention here as well.

Table 8: **Impact of Image Resolution with Vision-Language Pretraining**: The results here our implementations of the various methods. Params shows the number of parameters in the entire model (not just the fusion module). L is the number of self-attention blocks overall in the multi-modal encoder. RES is the image resolution during finetuning: we pretrain all models at resolution 224×224. Increasing resolution generally leads to better performance on both datasets. Compound Tokens outperform all other fusion methods across the two resolutions.

Fusion Method	L	Params ($\times 10^6$)	RES	GFlops	SNLI-VE	GQA
Merged Attention	12	332.94	224×224	16.95	81.01	77.06
Co-Attention	12	361.26	224×224	14.63	79.89	75.06
Co-Tokenization	12	391.27	224×224	28.84	81.52	77.60
Compound Tokens	10	337.67	224×224	16.55	80.44	77.55
Compound Tokens	12	339.97	224×224	17.23	**81.75**	**79.92**
Merged Attention	12	332.94	384×384	34.89	81.78	78.13
Co-Attention	12	361.26	384×384	29.61	80.50	75.92
Compound Tokens	10	337.67.	384×384	33.95	79.93	78.73
Compound Tokens	12	339.97	384×384	35.50	**82.47**	**79.55**

Table 9: **Impact of Image Encoder**: Both the ViT-base and ResNet-50 are pretrained on ImageNet but we do not use any additional image-language pretraining. All models are trained for 300k iterations. Compound Tokens obtains a higher accuracy than merged attention across all image encoders.

Image Encoder	Fusion Method	SNLI-VE	GQA
ViT-base	Merged Attention	77.44	74.02
	Compound Tokens	**78.59**	**74.74**
ResNet-50	Merged Attention	78.70	75.62
	Compound Tokens	**79.59**	**76.62**

C Encoder-Decoder VQA Model

We observed a generally low performance on VQA in our encoder-decoder model across all fusion mechanisms. We believe this is the case because our decoder is unable to generalize well to the VQA vocabulary due to our limited pretraining dataset. Besides the low performance, we also noticed that this dataset is very sensitive to hyper-parameter changes such as learning rate and dropout in our models. Faced with these challenges, we removed VQA from our ablations as indicated in the main text and show the results in this section for completeness.

Table 10: **Encoder-Decoder VQA Accuracy**: The VQA results in our encoder-decoder setup are generally low for all fusion methods and very sensitive to learning rate and dropout changes.

Setup	Merged Attention	Co-Attention	Co-Tokenization	Compound Tokens
Scratch	55.20	52.43	51.94	54.43
Pretrained	47.92	45.04	53.29	55.83