MaMMUT: A Simple Architecture for Joint Learning for MultiModal Tasks

Anonymous authors
Paper under double-blind review

Abstract

The development of language models have moved from encoder-decoder to decoder-only designs. In addition, the common knowledge has it that the two most popular multimodal tasks, the generative and contrastive tasks, tend to conflict with one another, are hard to accommodate in one architecture, and further need complex adaptations for downstream tasks. We propose a novel paradigm of training with a decoder-only model for multimodal tasks, which is surprisingly effective in jointly learning of these disparate vision-language tasks. This is done with a simple model, called MaMMUT. It consists of a single vision encoder and a text decoder, and is able to accommodate contrastive and generative learning by a novel two-pass approach on the text decoder. We demonstrate that joint learning of these diverse objectives is simple, effective, and maximizes the weight-sharing of the model across these tasks. Furthermore, the same architecture enables straightforward extensions to open-vocabulary object detection and video-language tasks. The model tackles a diverse range of tasks, while being modest in capacity. Our model achieves the state of the art on image-text and text-image retrieval, video question answering and open-vocabulary detection tasks, outperforming much larger and more extensively trained foundational models. It shows very competitive results on VQA and Video Captioning, especially considering its capacity. Ablations confirm the flexibility and advantages of our approach.

1 Introduction

Vision-language learning has become critical in improving both visual-understanding and multimodal vision-language tasks. Large foundational vision-language models, which are designed to be extended to multiple downstream tasks, follow two main training strategies, typically exemplified by disjoint architectures. Some vision-language pre-training approaches apply a contrastive loss, in a dual-encoder style architecture, e.g. CLIP, Align, Florence (Radford et al., 2021; Jia et al., 2021; Yuan et al., 2021). Contrastive training has been shown to produce strong backbones, which lead to successful image understanding and cross-modal retrieval tasks, e.g. image-to-text or text-to-image retrieval.

Alternatively, the autoregressive and masked token modeling objectives, well known from language modeling, are very popular with vision-language models which need to generate text. They are often referred to as split-captioning objectives. The split-captioning training is typically beneficial to text-generative tasks, e.g. Visual Question Answering (VQA). The most common architectures used in these scenarios are the encoder-decoder ones, which use a separate vision and text encoders, or a joint vision-text encoder, before a joint decoder, applying decoding losses from language learning (Cho et al., 2021; Chen et al., 2022; Piergiovanni et al., 2022a; Wang et al., 2021; 2022c). Architectures with cross-attention over frozen or partly frozen language models have also been popular (Alayrac et al., 2022).

Combining these two types of architectures and loss functions has proven to be challenging, with recent approaches such as Align-Before-Fuse (ALBEF) and CoCa (Li et al., 2021; Yu et al., 2022; Yan et al., 2022) requiring relatively complex models of multiple components to accommodate more tasks especially video tasks (Yan et al., 2022). Simultaneously, many pure language-only models have adopted simple decoder-only architectures with great success (Liu et al., 2018; Brown et al., 2020; Du et al., 2021) and with the added benefit of significant parameter reduction (Liu et al., 2018).

We here propose a simple approach to unify contrastive learning, localization aware and autoregressive captioning pretraining, by using a single language decoder and an image encoder. Our formulation is more general and allows maximal weight-sharing and parameter efficiency between the contrastive and generative tasks. To address the unique

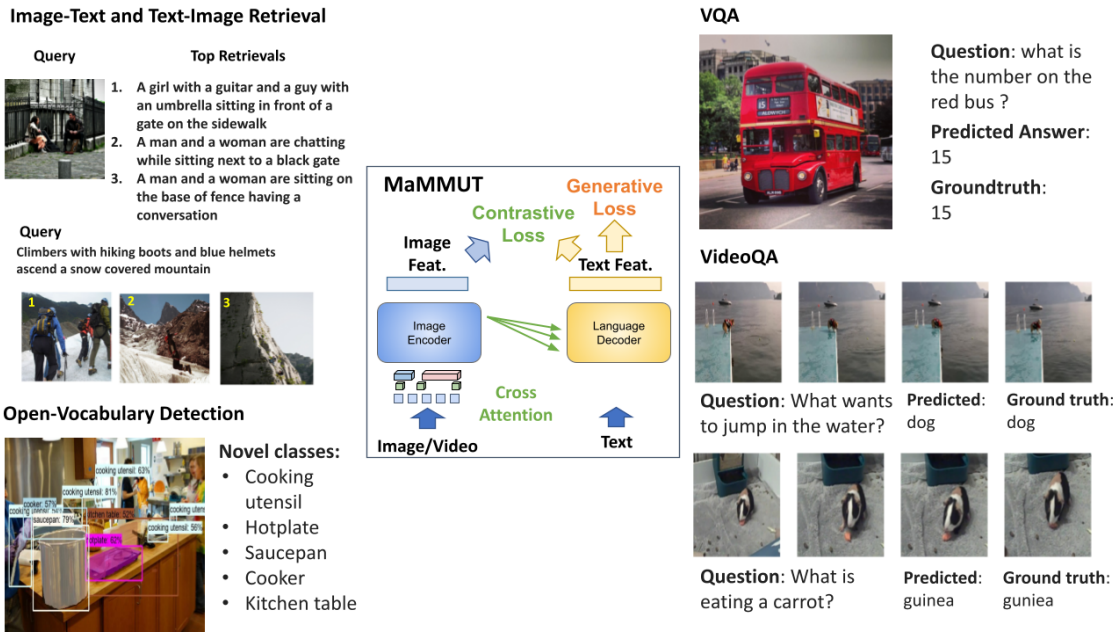

Figure 1: The MaMMUT model is a simple vision-encoder and text-decoder architecture, which serves as foundational model for both image-language and video-language tasks. Despite its relatively small size, the model outperforms SOTA on many diverse tasks. Example results on Image-Text/Text-Image retrieval, Visual Question Answering (VQA), Open-vocabulary detection, Video Question Answering (VideoQA) are shown.

challenge for multi-modal models of reconciling the unconditional sequence-level representation learning needed for contrastive learning with the token-conditioned next-token prediction, we propose a two-pass learning strategy using the text decoder. In one pass of the training, we utilize cross attention and causal masking to learn the caption generation task, where the text features can attend to the image features and predict the tokens in sequence; in the other pass we disable the cross-attention and causal masking, which learns the contrastive task without visibility into the image features. We further modify the contrastive training objective to be localization-aware, further equipping the model for object detection tasks. With this training strategy our model can address a diverse range of tasks, e.g. retrieval, text generative or detection. This provides a simpler alternative to previous approaches (Li et al., 2021; Yu et al., 2022; Singh et al., 2022; Yuan et al., 2021), where our model architecture is more compact and shared more broadly across tasks in the model. Furthermore, our model allows us to apply this architecture to video by a seamless adaptation based on the TubeViT approach (Piergiovanni et al., 2023a). This allows us to address video-text tasks, such as Video Question Answering (VideoQA) and Video Captioning successfully, outperforming prior large image-text or video foundatinoal models, such as 80B Flamingo, by pre-training on image-and-text dataset only. Furthermore, we extend the approach to leverage the pre-trained model for open-vocabulary detection, demonstrating the localization capabilities of the model. We call the model MaMMUT.

Our model is simple, but at the same time represents a union of tasks that other models find it challenging to put together, or need more specialized of complex adaptations for. In comparison, many of the prior approaches, despite reporting results on 10, 20 or more tasks or training bigger models (Chen et al., 2022; Singh et al., 2022; Alayrac et al., 2022; Yuan et al., 2021), accommodate fewer categories of tasks than our model does. One of the largest models, PaLI (Chen et al., 2022) addresses only classification, VQA and image captioning task categories, but is not designed to handle retrieval, detection, or video tasks. GIT (Wang et al., 2022b) extends the model to videos, but does not handle retrieval or detection. FLAVA (Singh et al., 2022), Florence (Yuan et al., 2021), ALBEF (Li et al., 2021), BLIP (Li et al., 2022a) which are based on constrastive learning, accommodate more tasks, but do not perform video-language or object localization tasks, or need significant downstream adaptations to do so (Yuan et al., 2021). Flamingo (Alayrac et al., 2022) is not able to do retrieval or object detection. With MaMMUT, we address all these tasks.

Experiments on image-text, text-image retrieval, VideoQA and Open-Vocabulary Object Detection show above-SOTA performances, outperforming much larger models. MaMMUT shows very competitive performance in Video Captioning and Visual Question Answering (VQA), which is done in the challenging open-ended generation setting. We note that our pre-training is done on image-text noisy pairs only and does not include video or labeled object localization

information or supervised labels. Ablation studies confirm the challenges and benefits of developing a simple and flexible model to unify multiple tasks.

Our contributions are: (i) a simple, compact and extensible foundational vision-language model which addresses multiple diverse multimodal tasks, such as image/text retrieval, Visual Question Answering, Open-Vocabulary Detection, Video Question Answering and Video Captioning; (ii) a novel two-pass learning method which trains jointly reconciling the unconditional sequence-level representation learning needed for contrastive learning with the token-conditioned captioning-like learning, using a single shared decoder-only model; (iii) a seamless extension to video tasks with no additional pre-training needed, creating a powerful video-language model.

2 Related Work

Vision and language pretraining has gained considerable popularity, where many successful and wide-ranging approaches have been created. Following token masking or next-token prediction losses used in text modeling, image-language pre-training methods extended these ideas for image and text inputs, where a language modeling loss is applied to a model which considers image inputs (Wang et al., 2022b; Chen et al., 2022; Wang et al., 2021; 2022e; Chen et al., 2020b; Zhou et al., 2020; Zhang et al., 2021a; Tan & Bansal, 2019; Li et al., 2020; Lu et al., 2019; Weijie Su, 2020; Dou et al., 2022). This technique enables text-generative vision-and-language tasks, such as VQA, Image Captioning or classification. Extensions to these approaches, where masking is done in the image space (Bao et al., 2021), have also been considered. Some of the above-mentioned approaches also feature similar architecture as ours, with a simple decoder e.g. GIT (Wang et al., 2022b), but the tasks addressed are limited to the text generative ones.

Contrastive vision and language pretraining has been popularized by the CLIP and Align models (Radford et al., 2021; Jia et al., 2021), demonstrating that contrastive learning can peoduce powerful embedding for many downstream applications. While the contrastive loss has been widely applicable, e.g. in self-supervised learning (Chen et al., 2020a), the appeal of these methods is the ability to learn across two modalities and produce high quality embeddings from large amounts of noisy image-language pairs which are harvested automatically from the web. Many generative pre-training approaches have also subsequently leveraged such noisy and easily obtainable datasets for vision-language pre-training purposes. Contrastive vision-language models are typically two-tower models, where nontrivial modifications are needed for downstream tasks (Yuan et al., 2021).

Several prior works proposed approaches to combine contrastive and generative vision-language pre-training, on the premise of two-tower models and cross-attention or cross-modal masking to align the modalities, e.g. ALBEF, FLAVA, BLIP, CoCa (Li et al., 2021; Singh et al., 2022; Li et al., 2022a; Yu et al., 2022). ALBEF (Li et al., 2021) applies a contrastive loss to an image and text encoder models and adds a decoder for generative tasks. A similar approach is taken by CoCa (Yu et al., 2022) where where the text encoder is actually a decoder-style causal masking model. Furthermore, a second decoder uses cross-attention to connect image representations with text to generate the output text. In these approaches, the contrastive loss is applied to a part of the model and is combined with the generative objectives from the final decoder. Other approaches, OFA and UnifiedIO (Wang et al., 2022d; Lu et al., 2022) unify understanding and generation tasks in the same sequence-to-sequence framework, spanning a larger number of tasks, including tasks, such as, referring expressions (Yu et al., 2016). These works are not easy to extend to retrieval, video or object detection tasks. BEiT-3 (Bao et al., 2021), propose to start from the same core model, but re-configure it according to the type of task. However this approach requires significant changes of the model depending on the task.

A number of video foundational models have been proposed (Zellers et al., 2021; Fu et al., 2021; Alayrac et al., 2022; Wang et al., 2022f;c; Cheng et al., 2022; Luo et al., 2020). The Flamingo model (Alayrac et al., 2022) extends a large frozen language model, with image and video inputs to deliver impressive results. VIOLET (Fu et al., 2021) uses masked language and masked-video modeling for joint video-text learning. Other approaches extend a pre-trained image-language model, adapting it to video (Wang et al., 2022b; Yan et al., 2022; Piergiovanni et al., 2023b), where a common approach is to just accumulate features from individual frames of the video. Our model is mostly aligned to the latter class of models, however, we instead directly process the spatio-temporal information of the video and do not need to process the model by a single frame at a time. Our approach is more advantageous, as much of the temporal information is lost in the sinlge-frame approaches. Florence (Yuan et al., 2021) does an adaptation with 3D kernels to the SWIN transformer (Liu et al., 2021), and OmniVL (Wang et al., 2022c) uses the Timesformer model (Gedas Bertasius, 2021), which preserve the temporal information, however the adaptation is generally more complex. Combining contrastive and captioning losses in video is also a popular technique. For example, InternVideo (Wang et al., 2022f)

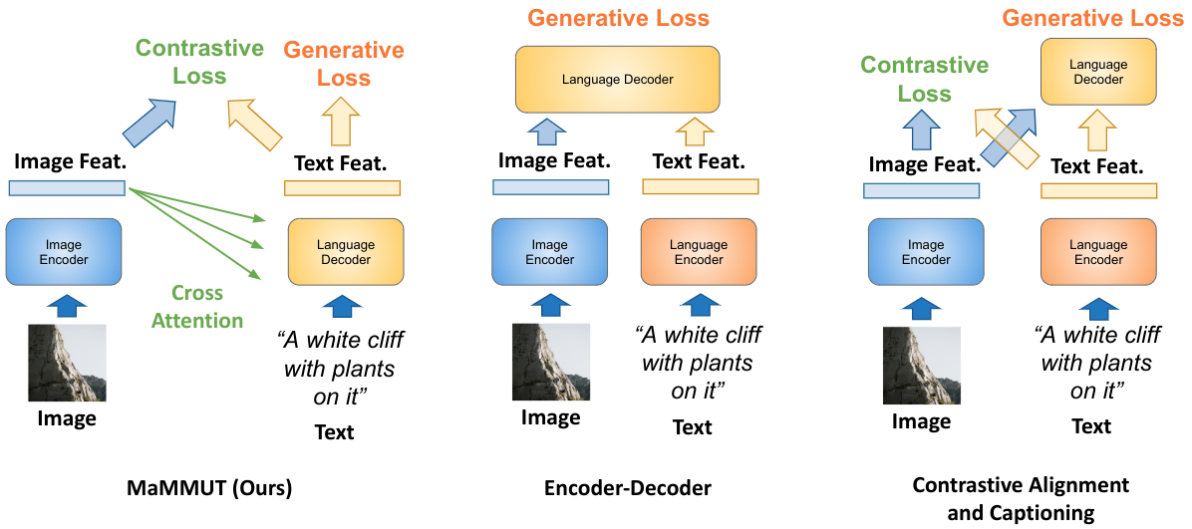

Figure 2: MaMMUT model architecture with an image encoder and text decoder (left), compared to others. Many encoder-decoder architectures (center) cannot handle the contrastive objective, for example (Cho et al., 2021; Chen et al., 2022). Approaches to combine contrastive and captioning (right), e.g. Align-Before-Fuse (ALBEF) (Li et al., 2021) or CoCa (Yu et al., 2022) develop more complex models and are hard to extend to video inputs or localization tasks. Our architecture is simpler than previous approaches and is able to accommodate more tasks.

proposed a combination of a masked video encoder-decoder and a cross-modal constrastive dual encoder in a video foundational model. In contrast, while we perform only a light fine-tuning over a image-language pre-trained model, our approach outperforms the above-mentioned more sophisticated and better-trained video models.

Object detection is available in some foundation models, for example VinVL (Zhang et al., 2021a) pre-train the vision-language model in order to detect objects and attributes. Florence (Yuan et al., 2021), adapt vision-language pre-trained models to object detection tasks. At the same time, Open-Vocabulary Detection, which aims at detecting novel categories of objects, is often overlooked by pre-trained vision-language models, where previous approaches focused on detecting objects in a set of pre-defined classes. VILD (Gu et al., 2022), demonstrate that large vision-language models are advantageous for detecting novel object categories. Inspired by this work, while taking a different approach, we show that open-vocabulary detection is an easy extension to our model with strong performance.

3 Method

We introduce our model, called MaMMUT, which offers a simple architecture consisting of a single image-encoder and a text-decoder (Figure 2, left). Our model combines the strengths of contrastive learning and autoregressive next-token prediction in a more flexible architecture than existing works. For example, the model is able to train and perform inference for those tasks with the same shared set of weights. Furthermore, one of the advantages of this simple model is that it allows simple extensions to video where the exact same architecture is able to consume directly video features. This is in contrast to prior work which adapt to video by processing individual frames or by more complex mechanisms. Other important downstream tasks, such as open-vocabulary object detection, can also be easily added to the model.

3.1 MaMMUT Architecture

MaMMUT is an intuitive and simple architecture for multimodal tasks, which consists of a single vision-encoder and a single text-decoder (see left of Figure 2). We encode the images into latent representation using a neural network encoder (such as ViT (Dosovitskiy et al., 2020)). The texts are both encoded and decoded with a Transformer-based decoder-only model. The MaMMUT architecture utilizes cross attention to fuse visual representation with text features anywhere in the decoder layers. This allows the whole decoder to produce both unimodal text representations (needed for a forward pass for contrastive learning) and multimodal text representation (needed to fuse the visual features with the text ones for vision+text tasks). This is done by a two pass joint training and we notably use a single combined training objective (shown later in Equation 6). Below we delve deeper into different aspects of the MaMMUT model.

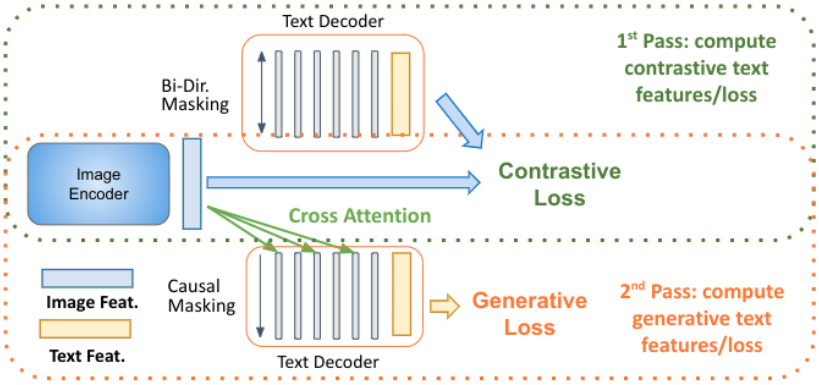

Figure 3: MaMMUT two-pass learning. We compute the image features via the image encoder. Then we compute the contrastive text features and loss by applying a non-causal masking to the language decoder, which makes it effectively a text encoder (top). Finally we compute the generative text features and loss by applying causal masking and cross attention with the image features (bottom). The decoder is visualized twice for clarity, its weights are fully shared.

Decoder-only Two-Pass learning. The main challenge to unify contrastive learning and next-token prediction is to unify the text representation, because the contrastive learning uses unconditional sequence-level representation, whereas captioning optimizes the likelihood of each token conditioned on the previous tokens.

We propose a two-pass approach to jointly learn the two conflicting types of text representations by the same model. During the first pass, to learn the contrastive task, we enable bi-directional masking within the decoder. The text features should not see the image features (which characterizes dual-encoder contrastive learner), but can attend to all tokens at once to produce the sequence-level representation. On the second pass, we utilize cross attention and causal masking to learn the caption generation task. The text features can attend to the image features and predict the tokens in a sequence (see Figure 3). As mentioned on the first pass we disable the cross-attention and causal masking in order to separate the text and image features for contrastive learning. All text-decoder weights are shared and gradients are aggregated from the two passes during training. The two passes are done interchangeably during training so their order is not important. As the two passes can share identical image representation, MaMMUT achieves considerable computation savings compared to training two separate models.

We use a vision transformer as image encoder, and project the output dimension of image encoder to that of text decoder by a linear layer. The global image (v_i) and text representation (l_i) for contrastive learning are computed by average pooling over spatial dimensions and the sequence length, respectively. We insert M cross-attention layers into N text decoder layers, where $M \approx \frac{N}{2}$. The ratio $\frac{M}{N}$ represents a trade-off between model capacity and text-generative capability, where higher M tends to benefit the text-generation tasks.

Somewhat surprisingly, MaMMUT requires no task-specific predictor heads to unify the two conflicting tasks other than a vocabulary embedding layer to map the decoder output features to text tokens. The same decoder features are average-pooled to represent the whole sequence (l_i). This is verified in our ablations as well. Compared to contrastive captioner (Yu et al., 2022), MaMMUT allows more flexible fusion of image and text features anywhere in the text decoder, and fully share the text-decoder parameters between contrastive and caption generation tasks.

Model simplicity. In contrast to other architectures, Figure 2, middle and right, we use a single vision encoder and a single text decoder. Unlike encoder-decoder models (Chen et al., 2022; Wang et al., 2022b;e) and contrastive alignment and captioning (Li et al., 2021; Yu et al., 2022), this is simpler and more flexible architecture. While using an encoder-decoder is popular, it comes with some disadvantages. For example, a pretrained visual encoder is required, and contrastive tasks are usually not feasible. Contrastive alignment and captioning methods extend the encoder-decoder models to handle retrieval, but that mechanism is not easy to adapt to video and creates a hard loss-balancing task. MaMMUT on the other hand works well for video-text tasks (Sec. 4) and can be seamlessly adapted without extending the architecture, adding more encoders or having to use an encoder multiple times.

3.2 Pretraining Losses

Our model combines the strengths of popular vision-language pre-training losses, as are specifically described below, including with our modifications.

Contrastive Loss. Compared to the fully supervised classification pretraining, the contrastive learning approach uses separate image and text encoders to compute image and text embeddings of typically web-scale image-text pairs (Radford et al., 2021; Jia et al., 2021; Yu et al., 2022; Zhai et al., 2022). This allows to model to learn from the richer supervisory signal of free-form text than a fixed label set. The two encoders are trained jointly to minimize the contrastive objective below (Oord et al., 2018):

$$L_{\text{I2T}} = -\frac{1}{B}\sum_{i=1}^{B}\log(\frac{\exp(v_i l_i/\tau)}{\sum_{i=1}^{B}\exp(v_i l_j/\tau)}). \quad L_{\text{T2I}} = -\frac{1}{B}\sum_{i=1}^{B}\log(\frac{\exp(v_i l_i/\tau)}{\sum_{j=1}^{B}\exp(v_i l_j/\tau)}). \tag{1}$$

$$L_{contrastive} = L_{\text{I2T}} + L_{\text{T2I}}. \tag{2}$$

, where v_i and l_i are the normalized image and text embeddings, and τ is the learnable temperature to scale the logits. The loss is summed over the number of elements in the batch B. As a result of learning from free-form text, the representation is very effective for zero-shot image classification, image-text retrieval, and robust to corrupted or out-of-distribution images (Radford et al., 2021).

Image Captioning loss. Dual-encoder contrastive models treat the text as a single entity to be encoded, while the encoder-decoder image captioner seeks a more detailed token-level understanding by predicting each word in sequence (see middle of Figure 2). This is achieved through an encoder-decoder architecture, where the image encoder creates a latent representation of the image using ViT (Dosovitskiy et al., 2020) or ConvNet backbone. The text decoder then generates the tokens autoregressively by maximizing the likelihood of each predicted token given the previously generated tokens in the sequence, resulting in the following forward autoregressive formulation:

$$L_{captioning} = -\sum_{t=1}^{T}\log P_\theta(y_t|y_{1,2,\ldots,t-1}, x). \tag{3}$$

To achieve maximum learning efficiency and parallelize computation, the encoder-decoder architecture is trained using a technique called teacher-forcing, which trains the model to predict the tokens at all time steps in parallel.

Focal Contrastive Loss. Contrastive learning typically relies on large batch size to extract supervisory signal from noisy image-text data. Our goal is to learn from the more informative and challenging examples than what is possible with the standard cross entropy loss. The focal loss (Lin et al., 2017) presents a compelling alternative as it allows us to finely tune the weights assigned to challenging examples, demonstrating improved performances for object classification or detection scenarios. It has been recently shown, that applying focal loss achieves very competitive performance with significantly smaller batch size for contrastive learning (Kim et al., 2023). More specifically the focal loss is applied as follows: Let v_i and l_i be the normalized image and text embeddings, and the image-to-text (I2T) focal contrastive loss be L_{focal}. Using the same notation as Equation 1, we can write L_{focal} mathematically as:

$$L_{\text{focal_contrastive}} = -\frac{1}{B}\sum_{i=1}^{B}\sum_{j=1}^{B}(1-p_i)^\gamma \log(p_i), \tag{4}$$

where p_i denotes the true class probability as below:

$$p_i = \begin{cases} \sigma(v_i l_j/\tau) & \text{if } i = j \\ 1-\sigma(v_i l_j/\tau) & \text{if } i \neq j \end{cases} \tag{5}$$

where σ denotes the sigmoid function. For simplicity, we use the non-alpha-balanced focal loss (Lin et al., 2017). The total loss is the sum of both I2T and T2I losses as in Equation 2. The focal contrastive loss modification provides additional sensitivity to objects, which equips the model for downstream detection tasks so we prefer it to the classical contrastive loss in our model.

Final loss Combining the losses above, our final loss is as follows. We note that while balancing of the losses is needed as is typical, the losses are simply combined and are shared with the full model.

$$L_{total} = \lambda_{cap} L_{captioning} + \lambda_{focal} L_{\text{focal_contrastive}}, \tag{6}$$

where λ_{cap}, λ_{focal}, are loss balancing hyper-parameters.

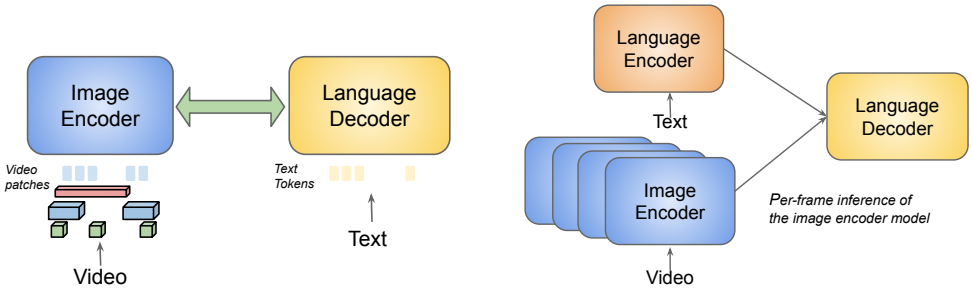

Figure 4: MaMMUT video model (left) efficiently and seamlessly extends the image-language model, by adding learnable spatio-temporal features. The model is applied only once, as opposed to other models, processing each individual frame independently (right). Our model uses image-text pre-training only.

Learning from Scratch with Noisy Image-Text Supervision. Unlike many existing methods that train model components in multiple stages using different data sources or modalities (Chen et al., 2022; Alayrac et al., 2022; Wang et al., 2022e;b), MaMMUT is pretrained end-to-end from scratch, without relying on any prior training or external sources. We use only a web alt-text dataset (Jia et al., 2021) for training. As the image encoder is typically the computation bottleneck in contrastive learning (Radford et al., 2021), our pretraining approach incurs only a relatively light overhead in training efficiency over a pure contrastive learner (Radford et al., 2021) ($\approx 16\%$). This is highly desirable as scaling up model and data size have shown consistent benefits in contrastive and captioning pretraining (Radford et al., 2021; Chen et al., 2022; Alayrac et al., 2022).

Learned Positional Embeddings for Localization Awarenesss. Existing vision and language pretraining approaches and detection finetuning have a mismatch in how they use positional embeddings. Pretraining approaches typically use full-image positional embeddings during training and apply the same embeddings for downstream tasks. But for detection finetuning, recognition occurs at the region level, requiring the full-image positional embeddings to generalize to regions not seen during pretraining. To address this gap, we adopted the Cropped Positional Embedding (Kim et al., 2023). The idea is to up-sample the positional embeddings from the pretraining image size (e.g., 224) to the detection task image size (e.g., 1024). Then, a randomly cropped and resized region from the up-sampled positional embeddings is used as the image-level positional embedding during pretraining. This method trains the model to view each image not as a full image, but as a region crop from a larger, unknown image, which better matches the downstream use case of detection where recognition occurs at the region level instead of the image level.

Implementation Details. Our model consists of a standard ViT-Huge (Dosovitskiy et al., 2020; Xiaohua Zhai & Beyer, 2022) image encoder (of 650M parameters) and a transformer text decoder of 1B parameters. The cross attention layers are applied every two decoder layers, where the model and hidden dimension follow the text decoder (420M params). Ablation results are conducted with ViT-Base image encoder and a smaller text decoder using 128M parameters, and a smaller 4K batch size. The optimal hyper-parameters are then used to run the large model. The large model is trained for 500K steps using a batch size of 16K. We use AdamW optimizer with weight decay value 0.01. Our initial learning rate is 0.001, and both generative and contrastive loss weights are set to 1.0. We first resize every image to 272x272 and randomly crop a 224x224 patch out for pretraining. We apply 10K warmup steps before applying linear LR decay to the end of training. The temperature in contrastive learning is learnable and initialized to 1.0. We use the standard Sentencepiece tokenizer and set the text length to 64 following existing works (Jia et al., 2021; Yu et al., 2022). We use an alt-text dataset of 1.8B image-text pairs (Jia et al., 2021), as is common in prior methods. The dataset is used for both contrastive and generative pretraining. To better match the downstream detection task, we finetune the model for 100K iterations with the Cropped Positional Embedding for further localization awareness.

4 MaMMUT for Video Tasks

Our video model is an efficient and seamless extension to the main image-language model, based on the TubeViT idea (Piergiovanni et al., 2023a). It extracts video tubes which are then projected to patches similar to 2D image projections (Figure 4, left). The model is applied only once. Some other image-language models adapted to video by processing each individual frame by the image encoder, e.g. (Wang et al., 2022b; Piergiovanni et al., 2023b; Yan et al., 2022) (Figure 4, right). This is a limitation as only a relatively small number of frames can be processed due

Method	image model size	MS COCO (5K test set) image-to-text R@1	R@5	R@10	text-to-image R@1	R@5	R@10	Flickr30K (1K test set) image-to-text R@1	R@5	R10	text-to-image R@1	R@5	R@10
CLIP (Radford et al., 2021)	302M	58.4	81.5	88.1	37.8	62.4	72.2	88.0	98.7	99.4	68.7	90.6	95.2
ALIGN (Jia et al., 2021)	408M	58.6	83.0	89.7	45.6	69.8	78.6	88.6	98.7	99.7	75.7	93.8	96.8
FLAVA (Singh et al., 2022)	86M	42.7	76.8	-	38.4	67.5	-	67.7	94.0	-	65.2	89.4	-
FILIP (Yao et al., 2021)	302M	61.3	84.3	90.4	45.9	70.6	79.3	89.8	99.2	99.8	75.0	93.4	96.3
Florence (Yuan et al., 2021)	637M	64.7	85.9	-	47.2	71.4	-	90.9	99.1	-	76.7	93.6	-
CoCa-L (Yu et al., 2022)	303M	65.4	85.6	91.4	50.1	73.8	81.8	91.4	99.2	**99.9**	79.0	95.1	97.4
CoCa (Yu et al., 2022)	1B	66.3	86.2	91.8	51.2	74.2	82.0	92.5	99.5	**99.9**	80.4	95.7	97.7
MaMMUT (ours)	630M	**70.7**	**89.1**	**93.7**	**54.1**	**76.8**	**84.2**	**94.9**	**99.5**	**99.9**	**82.5**	**96.0**	**98.0**

Table 1: **Zero-shot image-text retrieval results on COCO and Flickr30K benchmarks.** We evaluate our pretrained model compared to other methods. We achieve state-of-the-art results by a margin the on image-to-text / text-to-image retrieval benchmarks with comparable model capacity. (**bold**: best).

to memory and runtime constraints. There is also some evidence that for a large number of frames the number of effective tokens becomes too large which leads to deteriorated performance (Piergiovanni et al., 2023a).

We follow the TubeViT approach (Piergiovanni et al., 2023a). However, we apply several changes. One main challenge is that TubeViT requires fixed position embedding, whereas this does not match well with the learned positional embeddings of the main encoder. Since the video tubes are sparse and can overlap, TubeViT found that the fixed position embeddings were important. To enable those here, we propose using both position embeddings, and adding a weighted connection to the newly added fixed embeddings. Next, we use the same 2D patches, but at a sparse temporal stride, and finally add the tube projections following the settings used in TubeViT. All these tokens are concatenated together, and passed through the shared ViT backbone. Our TubeViT adaptation to video is very lightweight and no additional components or losses are needed. In comparison, previous approaches tend to exhibit higher complexity. InternVideo (Wang et al., 2022f) for example, supports a video masked encoder for MAE style losses in addition to a module similar to ALBEF. Flamingo (Alayrac et al., 2022) is able to process either image and video inputs, but it trains a separate image-encoder, the text part of which is discarded when integrating in the main model. We note that we outperform these in our experiments on video tasks.

We note that this change does not need any additional video data pre-training, and all experiments conducted in the paper are done by directly fine-tuning of the MaMMUT image-text model on a video dataset.

5 Experiments

5.1 Zero-Shot Image Retrieval Results

As Zero-shot image retrieval is a good indicator of the capability of the model without fine-tuning, we first evaluate the performance of MaMMUT on Zero-shot image-text retrieval tasks. Table 1 shows the image-to-text and text-to-image results, compared to the best SOTA methods on two popular retrieval benhcmarks MS COCO (Chen et al., 2015) and Flickr (Plummer et al., 2015). As established by previous approaches, we evaluate both image-to-text and text-to-image retrieval, following the same evaluation protocol. Our approach significantly outperforms the state-of-the-art models on both image-to-text and text-to-image retrieval, for example, by 2-4 points on the recall@1 metrics. We note that these are challenging benchmarks, which many approaches have been evaluated on.

5.2 Visual Question Answering Results

We report the performance on the VQAv2 benchmark (Agrawal et al., 2015) in Table 2. Inspired by recent VQA approaches (Chen et al., 2022; Alayrac et al., 2022; Piergiovanni et al., 2022a), we conduct the experiments in the open-ended text generative setting using an English vocabulary size of 256K. Most prior approaches (Wang et al., 2021; Yu et al., 2022; Wang et al., 2022e;d) address the VQA task in the classification setting where the best answer is selected from a predefined set of answers (typically of size 3K). Some recent works (Li et al., 2021; 2022a) train the model in an open-ended settings but restrict the decoder to generate only the 3K candidate answers during inference. In contrast, we allow the decoder to use the whole vocabulary during inference. The VQA-as-open-generation setting poses two key challenges: firstly, the produced text must precisely match the expected answer in order to be considered correct, and secondly, our vocabulary size is much larger than the ones utilized in the classification settings.

Method	Test-Dev	Test-Std
FLAVA (Singh et al., 2022)	72.8	-
METER (Dou et al., 2022)	77.7	77.6
Unified-IO (Lu et al., 2022)	77.9	-
OmniVL (Wang et al., 2022c)	78.3	78.4
Florence (Yuan et al., 2021)	80.2	80.4
SimVLM (Yu et al., 2022)	80.0	80.3
OFA (Wang et al., 2022d)	82.0	82.0
CoCa (Yu et al., 2022)	82.3	82.3
BEiT-3 (Yu et al., 2022)	84.2	84.0
ALBEF (Li et al., 2021)	75.8	76.0
AnswerMe (Piergiovanni et al., 2022a)	73.6	-
BLIP (Li et al., 2022a)	78.2	78.3
GIT (Wang et al., 2022b)	78.6	78.8
Flamingo-80B (Alayrac et al., 2022)	82.0	82.1
PaLI-3B (Chen et al., 2022)	79.3	-
PaLI-15B (Chen et al., 2022)	80.8	-
PaLI-17B (Chen et al., 2022)	84.3	84.3
MaMMUT (2B)	80.7	80.8

Table 2: **Visual Question-Answering (VQAv2).** We benchmark the performance in an open-ended generation setting. Approaches that perform VQA in closed-vocabulary settings are marked in gray. MaMMUT is very competitive among existing open-ended generation methods given its modest model capacity.

Overall	Yes/No	Number	Other
80.84	93.41	63.89	73.78

Table 3: **Visual Question-Answering (VQAv2) analysis.** Performance by question types on the test-std split. "Yes/No" questions perform best, whereas questions about how many objects are present, are the most challenging.

MaMMUT achieves 80.8 accuracy on this benchmark, which is very competitive among the open-ended generation approaches. For example, MaMMUT outperforms the PaLI-3B by 1.4 points, while using 1.5x fewer parameters (2B total params). Compared to the PaLI-15B, ours achieves the same performance while being 7.5x smaller. In addition, Flamingo and PaLI use a combination of interleaved image-text, alt-text, human-annotated data sources for training (Alayrac et al., 2022; Chen et al., 2022), whereas we use only an alt-text dataset (Jia et al., 2021). Larger models, such as PaLI-17B, outperform ours. Table 3 shows MaMMUT performance on VQA test-std set by question types. MaMMUT performs the best on yes/no question type, and less well on questions that require counting.

5.3 Video Question Answering Results

In this section we present the results of our model on the Video Question Answering task, which is a challenging task answering questions about activities, events, objects, or repetition counting within a video. Our VideoQA results are obtained using the image-text pre-trained model, namely we directly fine-tune on a VideoQA dataset without any video-text pre-training. This is in contrast with other works which use video data pre-training, and indicates that our model is already very strong without additional video-text pre-training.

The results are presented in Table 4a, comparing to the state of the art on MSRVTT-QA (Jun Xu & Rui, 2016) and MSVD-QA (Xu et al., 2017) datasets. MaMMUT outperforms the best SOTA approaches, among which are both video foundational models, e.g VIOLET, MERLOT, InternVideo, image-text models adapted to video e.g. GIT and GIT2 (Wang et al., 2022b), and large vision-language models, such as Flamingo (Alayrac et al., 2022). We note that, similar to the VQA results, our VideoQA results are conducted in the more challenging open-ended generation setting. Our model is 2.5 times smaller than GIT2 (5B parameters), and 40 times smaller than Flamingo (80B parameters).

5.4 Video Captioning Results

Video captioning results on the MSRVTT (Jun Xu & Rui, 2016) and MSVD (Chen & Dolan, 2011; Xu et al., 2017) datasets are presented in Table 4b. The results here too are obtained with image-text pre-training only. Our approach

performs well related to SOTA. It outperfoms prior approaches on the MSVD benchmark by large margins, and outperforms most others with the exception of GIT/GIT2 on the MSRVTT video captioning dataset.

Method	MSRVTT-QA	MSVD-QA
Just Ask (Yang et al., 2021)	41.5	46.3
MERLOT (Zellers et al., 2021)	43.1	-
OmniVL (Wang et al., 2022c)	44.1	51.0
VindLU (Cheng et al., 2022)	44.6	-
Iterative Co-Tok (Piergiovanni et al., 2022b)	45.7	48.8
All-in-one (Wang et al., 2022a)	46.8	48.3
Video-Coca (Yan et al., 2022)	46.3	56.9
VIOLET (Fu et al., 2021)	43.9	47.9
VIOLETv2 (Fu* et al., 2023)	44.5	54.7
Dynamic Pretr. (Piergiovanni et al., 2023b)	45.1	47.1
GIT (Wang et al., 2022b)	43.2	56.8
GIT2 (Wang et al., 2022b)	45.6	58.2
InternVideo (Wang et al., 2022f)	47.1	55.5
Flamingo (Alayrac et al., 2022)	47.4	-
MaMMUT (ours)	**49.5**	**60.2**

(a) **Video QA Results.** MaMMUT outperforms the SOTA on both MSRVTT-QA and MSVD-QA datasets. We note that image-language pre-training is the only one used here. MaMMUT outperforms video-first models e.g. VIOLET, InternVideo, image+video models, e.g. Flamingo, and large pre-trained image-text models adapted to video, e.g. GIT2.

Method	MSRVTT	MSVD
ORG-TRL (Zhang et al., 2020)	50.9	95.2
OpenBook (Zhang et al., 2021b)	52.9	-
SWINBert (Lin et al., 2022)	53.8	120.6
VIOLETv2 (Fu* et al., 2023)	58.0	130.2
MV-GPT (Seo et al., 2022)	60.0	-
Vid2Seq (Yang et al., 2023)	64.6	146.2
Video-Coca (Yan et al., 2022)	73.2	-
GIT (Wang et al., 2022b)	73.9	180.2
GIT2 (Wang et al., 2022b)	**75.9**	185.2
MaMMUT (ours)	73.6	**195.6**

(b) **Video Captioning Results.** MaMMUT performs well on both MSRVTT and MSVD Video Captioning Benchmarks, outperforming SOTA on MSVD by large margins. CIDEr scores are shown. As for VideoQA experiments, only image-language pre-training is used, directly fine-tuning on each dataset.

Table 4: Video Question Answering (VideoQA) and Video Captioning results.

5.5 Open-vocabulary Detection Results

The Open-Vocabulary detection task refers to the ability to detect and name objects (providing bounding boxes) of categories that are unknown to the model. We evaluate our work on the challenging LVIS dataset (Gupta et al., 2019) which features 1200+ different object categories. We report performances using the Average Precision (AP) metrics for rare classes, as previously established in the literature (Gu et al., 2022; Minderer et al., 2022). The open-vocabulary detector is initialized with the pretrained ViT backbone during finetuning. It adopts the simple feature pyramid and windowed attention to handle higher resolution images (e.g., 1024) as proposed in ViTDet (Li et al., 2022b), and Mask R-CNN heads and class-agnostic box/mask heads as in (Du et al., 2022; Gu et al., 2022; Zareian et al., 2021; Zhong et al., 2022). The model is trained with base categories, and tested to detect novel category objects at inference.

Table 5 presents our results on Open-Vocabulary detection which evaluates how well it does on detecting novel, previously unseen object categories. MaMMUT achieves 31.0 AP_r, which outperforms the existing ViT-based method OWL-ViT by 5.4 points. We also report AP performance for all classes for context, but note that our model is not optimized for detection AP, as is for example OWL-ViT which is a detection-only model.

6 Ablation studies

We here present ablation studies to understand the model characteristics and its design choices. We use a ViT-Base image encoder and a 128M language decoder, training on 4K batch size for 500K iterations unless noted otherwise.

Cross-task benefits. We first explore cross-task benefits between contrastive and text-generative pretraining. We find that joint training is generally favorable to tasks, but it affects tasks differently (Table 6). Compared to the contrastive-only pretraining, joint training achieves signficantly better VQA performance by +8 points, which is expected. Joint training improves text-to-image retrieval. This is likely because the generative modeling enhances text representation, and while remaining very competitive on the image-to-text retrieval, it is not as beneficial, which can indicate potential competition of tasks (please see our later experiments which explore this). Compared to the generative-only pretraining, joint training outperforms on VQA by +1.8 points, likely because the contrastive learning helps to improve the joint image-text representation. We attribute this to better representation learning rendered from contrastive training,

Method	APr	AP
DetPro-Cascade (Du et al., 2022)	20.0	27.0
Detic-CN2 (Zhou et al., 2022)	24.6	32.4
RegionCLIP (Zhong et al., 2022)	22.0	32.3
ViLD-Ensemble (Gu et al., 2022)	21.7	29.6
ViLD-Ensemble (Gu et al., 2022)	26.3	29.3
VL-PLM (Zhao et al., 2022)	17.2	27.0
Rasheed et al. (Rasheed et al., 2022)	21.1	25.9
OWL-ViT (Minderer et al., 2022)	23.3	**35.3**
OWL-ViT (Minderer et al., 2022)	25.6	34.7
MaMMUT (ours)	**31.0**	32.8

Table 5: **Open-Vocabulary Detection Results.** MaMMUT performs well on detecting novel objects, scoring much higher on the Average Precision (AP) of rare objects AP_r.

		MS COCO		Flickr30K		VQA
Contrastive	Generative	I2T	T2I	I2T	T2I	Acc.
✓		**54.8**	38.2	**82.6**	67.1	63.5
	✓	-	-	-	-	69.9
✓	✓	54.3	**38.7**	80.6	**67.5**	71.7

Table 6: **Cross-task benefits.** Combining the contrastive and generative objectives yield benefits for generative task in our setup, while maintaining the performance of discriminative tasks.

also evidenced by other works (Radford et al., 2021; Jia et al., 2021). More importantly, joint training model is able to tackle retrieval tasks, not supported by the generative pretraining, and achieving satisfying performance overall.

Cross-attention Design. Another key exploration is the role of cross-attention mechanisms in the proposed joint training and how they affect various tasks. Cross-attention provide an efficient means for communication between the two modalities. We find that tasks indeed perform better under different circumstances. Specifically, cross-attention is preferred for text generative tasks, but not as much for contrastive ones. Table 7 shows that while denser cross-attention connections (e.g. 4) benefit the VQA task, one or few layers (e.g. 2) are sufficient for retrieval task.

Balancing contrastive-vs-generative losses. We here explore how to train the joint objectives, so that the tasks benefit fully from joint training. We do observe a potential competition of these tasks, thus a balance between losses needs to be obtained. Table 8 shows the performance trade-off when training jointly. The experiment is done on a smaller model and fewer steps, as mentioned above. We observe that the two objectives indeed have competitive behaviors. In the rest of the experiments, we pick equal weights over these two losses, as we observe that it gives more advantage to the VQA performance, whereas the retrieval does not suffer as much. This experiment provides an insight as to how to tune these parameters, depending on the requirements of the application.

Model size effects. Effects of model scale on vision-language tasks are undeniable (Alayrac et al., 2022; Chen et al., 2022), so naturally larger scales help performance. However, we show here that models of relatively small sizes, achieve SOTA or above performance when trained jointly. The results are shown in Table 9. Model scale yields additional improvement, but we observe that even small models are close or in some cases outperform SOTA which is a much larger model. This is due to the benefits our our approach (we note here that, for direct comparison, we have used the same dataset as the current SOTA method). We use 16K batch size in this ablation.

	MS COCO		Flickr30K		VQA
# Cross-Att.	I2T	T2I	I2T	T2I	Acc.
1	55.3	39.6	81.7	67.2	68.7
2	**56.6**	**40.1**	81.9	**67.6**	70.8
4	55.7	39.9	**82.2**	67.3	**71.5**

Table 7: **Cross-attention Design.** Denser cross-attention layers are beneficial for the VQA task, whereas few cross-attention layers are sufficient for retrieval tasks.

weights	MS COCO I2T	MS COCO T2I	Flickr30K I2T	Flickr30K T2I	VQA Acc.
(2.0, 0.5)	**56.71**	**40.77**	**82.52**	**67.77**	70.08
(1.0, 1.0)	56.25	39.73	81.74	67.50	**71.48**

Table 8: **Balancing the losses.** Zero-shot retrieval results on COCO and Flickr30K benchmarks (R@1 shown) and VQA accuracy. Weight coefficients for the contrastive and generative loss, denoted as (constrastive, generative) affect final performance. We use the smaller model, but with fewer training steps, compared to the settings in Table 9.

method	image backbone	MS COCO I2T	MS COCO T2I	Flickr30K I2T	Flickr30K T2I
CLIP (Radford et al., 2021)	302M	58.4	37.8	88.0	68.7
ALIGN (Jia et al., 2021)	408M	58.6	45.6	88.6	75.7
FILIP (Yao et al., 2021)	302M	61.3	45.9	89.8	75.0
CoCa (Yu et al., 2022)	1B	66.3	51.2	92.5	80.4
MaMMUT (ours)	300M	**66.4**	49.4	91.2	78.8
MaMMUT (ours)	630M	**70.7**	**54.1**	**94.9**	**82.5**

Table 9: **Model scaling results.** Zero-shot image-text retrieval results on COCO and Flickr30K benchmarks, R@1 shown. We see definitive improvements due to scale. However even our small models perform close or outperform much larger SOTA models (bolded). This experiment is run in the same settings as is Table 1.

Video Model Ablations. In Table 10, we compare the different design choices for adaptation of the video model. We explore the effects of removing gated connections, the fixed embeddings and even the video feature inputs. As seen, their importance increases in that order, where the addition of video tubes to process the video are of highest importance. This is not surprising as they learn the spatio-temporal video information. These ablations are done with the smaller ViT-L model.

7 Broader impacts

Developing vision-language models is very beneficial as they have many potential uses and applications, being able to understand both visual and language inputs better and to more accurately respond to questions. Some aspects of the model, specifically the text generative components might exemplify certain risks of generating off-topic, stereotypical, unwanted or other types of outputs, for which further investigation is needed. Our model has been trained on the same image-text dataset as many other previous works (Jia et al., 2021; Xiaohua Zhai & Beyer, 2022; Zhai et al., 2022; Yu et al., 2022; Yan et al., 2022) and is collected as noisy image-language pairs in a similar fashion to an even broader set of methods in the literature (Chen et al., 2022; Yuan et al., 2021; Alayrac et al., 2022). We developed the model for exploring novel research capabilities and have only used it for evaluation and visualization purposes in this paper.

8 Conclusions

We present the MaMMUT model which is a vision-encoder text-decoder model capable of multiple vision-language tasks. We propose two-pass learning which allows us to train jointly for retrieval and text-generative tasks with fully shared weights. The model is easy to adapt to video-language and object detection tasks. Our model accommodates a set of diverse tasks, such as image-text and text-image retrieval, novel (or open-vocabulary) object detection, VQA, VideoQA and Video Captioning, obtaining very strong performance on them with respect to the state-of-the art.

	MSRVTT-QA	MSVD-QA
MaMMUT- Full Model	**42.1**	**45.8**
No Gated Connection	41.8	45.5
No Fixed Embeddings	41.5	45.1
No Tubes	40.3	42.6

Table 10: **Video Tubes Adaption Experiments.** Comparing different effects of adapting the model to video.

References

Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, and Devi Parikh. Vqa: Visual question answering. In *ICCV*, 2015. 8

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. 1, 2, 3, 7, 8, 9, 10, 11, 12

Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 3

Tom B. Brown, Benjamin Mann, Nick Ryder, Jared Kaplan Melanie Subbiah, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Christopher Hesse Clemens Winter, 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. 2020. 1

David Chen and William Dolan. Collecting highly parallel data for paraphrase evaluation. In *ACL*, 2011. 9

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020a. 3

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. In *arXiv preprint arXiv:2209.06794*, 2022. 1, 2, 3, 4, 5, 7, 8, 9, 11, 12

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C. Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. In *https://arxiv.org/abs/1504.00325*, 2015. 8

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*, 2020b. 3

Feng Cheng, Xizi Wang, Jie Lei1 David Crandall, Mohit Bansal, and Gedas Bertasius. Vindlu : A recipe for effective video-and-language pretraining. *arXiv preprint arXiv:2212.05051*, 2022. 3, 10

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. 1, 4

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4, 6, 7

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 *CVPR*, 2022. 3, 9

Nan Du, Yanping Huang, Andrew M. Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, Barret Zoph, Liam Fedus, Maarten Bosma, Zongwei Zhou, Tao Wang, Yu Emma Wang, Kellie Webster, Marie Pellat, Kevin Robinson, Kathleen Meier-Hellstern, Toju Duke, Lucas Dixon, Kun Zhang, Quoc V Le, Yonghui Wu, Zhifeng Chen, and Claire Cui. Glam: Efficient scaling of language models with mixture-of-experts. 2021. 1

Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. Learning to prompt for open-vocabulary object detection with vision-language model. In *CVPR*, 2022. 10, 11

Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. Violet: End-to-end video-language transformers with masked visual-token modeling. In *arXiv:2111.1268*, 2021. 3, 10

Tsu-Jui Fu*, Linjie Li*, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. An empirical study of end-to-end video-language transformers with masked visual modeling. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 10

Lorenzo Torresani Gedas Bertasius, Heng Wang. Is space-time attention all you need for video understanding? 2021. 3

Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *ICLR*, 2022. 4, 10, 11

Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *CVPR*, 2019. 10

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021. 1, 3, 6, 7, 8, 9, 11, 12

Ting Yao Jun Xu, Tao Mei and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, 2016. 9

Dahun Kim, Anelia Angelova, and Weicheng Kuo. Region-aware pretraining for open-vocabulary object detection with vision transformers. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 6, 7

Junnan Li, Ramprasaath R. Selvaraju, Akhilesh D. Gotmare, Shafiq Joty, Caiming Xiong, and Steven C.H. Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*, 2021. 1, 2, 3, 4, 5, 8, 9

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*, 2022a. 2, 3, 8, 9

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, 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. In *ECCV*, 2020. 3

Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *ECCV*, 2022b. 10

Kevin Lin, Linjie Li, Chung-Ching Lin, Faisal Ahmed, Zhe Gan, Zicheng Liu, Yumao Lu, and Lijuan Wang. Swinbert: End-to-end transformers with sparse attention for video captioning. *CVPR*, 2022. 10

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, pp. 2980–2988, 2017. 6

Peter J. Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. Generating wikipedia by summarizing long sequences. *ICLR*, 2018. 1

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *CVPR*, 2021. 3

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *CVPR*, 2019. 3

Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. *arXiv preprint arXiv:2206.08916*, 2022. 3, 9

Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*, 2020. 3

Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, Xiao Wang, Xiaohua Zhai, Thomas Kipf, and Neil Houlsby. Simple open-vocabulary object detection with vision transformers. In *ECCV*, 2022. 10, 11

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 6

AJ Piergiovanni, Wei Li, Weicheng Kuo, Mohammad Saffar, Fred Bertsch, and Anelia Angelova. Answer-me: Multi-task open-vocabulary visual question answering. *arXiv preprint arXiv:2205.00949*, 2022a. 1, 8, 9

AJ Piergiovanni, Kairo Morton, Weicheng Kuo, Michael Ryoo, and Anelia Angelova. Video question answering with iterative video-text co-tokenization. *ECCV*, 2022b. 10

AJ Piergiovanni, Weicheng Kuo, and Anelia Angelova. Rethinking video vits: Sparse video tubes for joint image and video learning. *CVPR*, 2023a. 2, 7, 8

AJ Piergiovanni, Weicheng Kuo, Wei Li, and Anelia Angelova. Dynamic pretraining of vision-language models. In *First workshop on Multimodal Representation Learning, International Conference on Learning Representations (ICLR)*, 2023b. 3, 7, 10

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, pp. 2641–2649, 2015. 8

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1, 3, 6, 7, 8, 11, 12

Hanoona Rasheed, Muhammad Maaz, Muhammad Uzair Khattak, Salman Khan, and Fahad Shahbaz Khan. Bridging the gap between object and image-level representations for open-vocabulary detection. *arXiv preprint arXiv:2207.03482*, 2022. 11

Paul Hongsuck Seo, Arsha Nagrani, Anurag Arnab, and Cordelia Schmid. End-to-end generative pretraining for multimodal video captioning. *CVPR*, 2022. 10

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *CVPR*, pp. 15638–15650, 2022. 2, 3, 8, 9

Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *EMNLP*, 2019. 3

Alex Jinpeng Wang, Yixiao Ge, Rui Yan, Ge Yuying, Xudong Lin, Guanyu Cai, Jianping Wu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. All in one: Exploring unified video-language pre-training. *arXiv preprint arXiv:2203.07303*, 2022a. 10

Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022b. 2, 3, 5, 7, 9, 10

Junke Wang, Dongdong Chen, Zuxuan Wu, Chong Luo, Luowei Zhou, Yucheng Zhao, Yujia Xie, Ce Liu, Yu-Gang Jiang, and Lu Yuan. Omnivl:one foundation model for image-language and video-language tasks. 2022c. 1, 3, 9, 10

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *arXiv preprint arXiv:2202.03052*, 2022d. 3, 8, 9

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022e. 3, 5, 7, 8

Yi Wang, Kunchang Li, Yizhuo Li, Yinan He1, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, Sen Xing, Guo Chen, Junting Pan, Jiashuo Yu, Yali Wang, Limin Wang, and Yu Qiao. Internvideo: General video foundation models via generative and discriminative learning. In *arXiv preprint arXiv:2212.03191*, 2022f. 3, 8, 10

Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021. 1, 3, 8

Yue Cao Bin Li Lewei Lu Furu Wei Jifeng Dai Weijie Su, Xizhou Zhu. Vl-bert: Pre-training of generic visual-linguistic representations. In *ICLR*, 2020. 3

Neil Houlsby Xiaohua Zhai, Alexander Kolesnikov and Lucas Beyer. Scaling vision transformer. *CVPR*, 2022. 7, 12

Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *ACM Multimedia*, 2017. 9

Shen Yan, Tao Zhu, ZiRui Wang, Yuan Cao, Mi Zhang, Soham Ghosh, Yonghui Wu, and Jiahui Yu. Videococa: Video-text modeling with zero-shot transfer from contrastive captioners. In *ArXiV:2212.04979*, 2022. 1, 3, 7, 10, 12

Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just ask: Learning to answer questions from millions of narrated videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1686–1697, 2021. 10

Antoine Yang, Arsha Nagrani, Paul Hongsuck Seo, Antoine Miech, Jordi Pont-Tuset, Ivan Laptev, Josef Sivic, and Cordelia Schmid. Vid2seq: Large-scale pretraining of a visual language model for dense video captioning. *CVPR*, 2023. 10

Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. In *ICLR*, 2021. 8, 12

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *TMLR*, 2022. URL https://openreview.net/forum?id=Ee277P3AYC. 1, 2, 3, 4, 5, 6, 7, 8, 9, 12

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. *ECCV*, 2016. 3

Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, Ce Liu, Mengchen Liu, Zicheng Liu, Yumao Lu, Yu Shi, Lijuan Wang, Jianfeng Wang, Bin Xiao, Zhen Xiao, Jianwei Yang, Michael Zeng, Luowei Zhou, and Pengchuan Zhang. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, November 2021. URL https://www.microsoft.com/en-us/research/publication/florence-a-new-foundation-model-for-computer-vision/. 1, 2, 3, 4, 8, 9, 12

Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *CVPR*, 2021. 10

Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. 2021. 3, 10

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *CVPR*, 2022. 6, 12

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. *arXiv preprint arXiv:2101.00529*, 2021a. 3, 4

Ziqi Zhang, Yaya Shi, Chunfeng Yuan, Bing Li, Peijin Wang, Weiming Hu, and Zheng-Jun Zha. Object relational graph with teacher-recommended learning for video captioning. *CVPR*, 2020. 10

Ziqi Zhang, Zhongang Qi, Chunfeng Yuan, Ying Shan, Bing Li, Ying Deng, and Weiming Hu. Open-book video captioning with retrieve-copy-generate network. In *CVPR*, 2021b. 10

Shiyu Zhao, Zhixing Zhang, Samuel Schulter, Long Zhao, Anastasis Stathopoulos, Manmohan Chandraker, Dimitris Metaxas, et al. Exploiting unlabeled data with vision and language models for object detection. In *ECCV*, 2022. 11

Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, and Jianfeng Gao. Regionclip: Region-based language-image pretraining. In *CVPR*, 2022. 10, 11

Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J. Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. *AAAI*, 2020. 3

Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *ECCV*, 2022. 11

A Appendix

A.1 Additional ablations

We here include additional ablation experiments which can provide further introspection into the model.

Language pre-training effects. As shown in this paper, a purely language model can be easily adapted for multiple image-language tasks. We here explore the effects of language pre-training on the vision-language tasks. Our results show that language-only pre-training is beneficial for image-text retrieval benchmarks, but the effects are minor (Table 11).

	MS COCO		Flickr30K	
pre-training?	I2T	T2I	I2T	T2I
No	54.2	**38.9**	80.5	66.5
Yes	**54.9**	38.9	**82.4**	**67.3**

Table 11: **Effects of language-side pre-training.** Zero-shot image-text retrieval results on COCO and Flickr30K benchmarks, R@1 shown. Language pre-training provides small but consistent improvements.

Projections and attention pooling. Table 12 experiments with additional projections and attention pooling. We find that they are not needed for joint learning, indicating that the text decoder can be fully shared between tasks, which is a surprising finding. We only train for 100K iterations here.

Bi-directional masking is important for contrastive learning. Since here we are using a single language decoder to perform tasks of various characteristics e.g. retrieval and VQA, it is important to understand how to construct the language decoder for such tasks. We evaluated the effect of bi-directional or causal masking on a contrastive learning task on the Flickr dataset. We note that with these tasks, our model has to accommodate retrieval tasks which tend to process data with short text, whereas generative tasks tend to generate longer-text queries. We find that bi-directional masking benefits the contrastive learning much more than causal (Table 13), which is not surprising as the text associated with retrieval tasks is fully available before final feature representation is generated.

	image-to-text			text-to-image		
Att. Pool/ Proj	R@1	R@5	R@10	R@1	R@5	R@10
N/N	53.32	80.57	86.91	39.86	67.7	77.46
Y/N	52.44	78.71	86.33	39.3	66.62	76.91
N/Y	51.37	78.81	86.13	39.49	66.97	77.05
Y/Y	50.78	76.95	85.74	38.59	66.54	76.07

Table 12: **Projection and attention pooling.** Our conclusions are that they are not needed to realize joint training. Retrieval performance on Flickr30K is shown.

	MS COCO		Flickr30K	
Masking	I2T	T2I	I2T	T2I
Causal	53.0	37.5	78.1	62.5
Bi-directional	**54.9**	**38.8**	**82.4**	**67.3**

Table 13: **Bi-directional masking.** Contrastive learning clearly benefits from Bi-directional masking in the decoder language model.