# Supplementary

# A. Model Details

## A.1. Model Architecture

Our model utilizes the vision encoder and text encoder to learn uni-modal representations, and the fusion encoder to conduct cross-modal interactions, respectively. The whole architecture is displayed in Fig. 5.

## A.2. Previous Pre-training Tasks

**Contrastive Learning (CL).** We conduct CL on the global representations from vision and text encoders. Given a batch of image-text pairs, for an image (text), the paired text (image) is tread as the positive sample, and other texts (images) are negative samples. We use the InfoNCE loss as follows:

$$\begin{aligned} \mathbf{NCE}_{V2T} &= -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(s(V_i,T_i)/\tau)}{\sum_{n=1}^{N}\exp(s(V_i,T_n)/\tau)}, \\ \mathbf{NCE}_{T2V} &= -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(s(T_i,V_i)/\tau)}{\sum_{n=1}^{N}\exp(s(T_i,V_n)/\tau)}, \end{aligned} \tag{5}$$

where $N$ is the batchsize and $\tau$ serves as a learnable temperature parameter. The similarity function is formatted as cosine similarity, $s(V,T) = \frac{\phi_v(V)^T\phi_t(T)}{||\phi_v(V)||\cdot||\phi_t(T)||}$, where $\phi$ is a linear projection head. The vision-text contrastive loss is defined as:

$$\mathcal{L}_{CL} = \mathbf{NCE}_{V2T} + \mathbf{NCE}_{T2V}\,. \tag{6}$$

For the video-text data, we use the mean pooling of $M$ frame [CLS] features to denote the global representation of a video and then also use Eq. (5) for contrastive learning.

**Vision-Text Matching (VTM).** The model is required to predict whether a pair of image-text (video-text) is matched or not. Specifically, we conduct a binary classification on the concatenation of the visual and textual global features. The loss is defined as:

$$\mathcal{L}_{VTM} = \mathbf{CE}(\phi(\text{concat}[V,T]), y), \tag{7}$$

where $V, T$ are [CLS] features and $y$ is the ground truth. **CE** is Cross-Entropy loss and $\phi$ refers to a binary classifier. The image-text (video-text) pairs serve as positive samples, and we randomly replace the image (video) in a data pair with another image (video) to build negative sample.

**Masked Language Modeling (MLM).** We adopt MLM following BERT [24], which conducts a classification on the vocabulary list to predict masked words. We randomly mask out 15% text tokens, and replace them with the [MASK] token, random words, or left unchanged, with the probability of 80%, 10% and 10%, respectively. The classification loss is as follows:

$$\mathcal{L}_{MLM} = \mathbf{CE}(\phi(T_{mask}), y), \tag{8}$$

where $T_{mask}$ is the output masked token feature, $\phi$ serves as a classifier, and $y$ is the original token ID.

The overall training objective of our model is:

$$\mathcal{L} = \mathcal{L}_{CL} + \mathcal{L}_{ITM} + \mathcal{L}_{MLM} + \mathcal{L}_{SCL}, \tag{9}$$

where $\mathcal{L}_{SCL}$ is defined in Eq. (3).

# B. Experiments Details

## B.1. Pre-training Settings

In the image-text pre-training phase, we train the model for 100k steps totally using a batch size of 4096 on 64 NVIDIA A100 GPUs. We adopt the AdamW optimizer with a weight decay of 0.01. The learning rate of uni-modal encoders is warmed up from 0 to $1e-5$ in first 10% steps and then decayed linearly. The fusion transformer has a five times higher learning rate. As for the video-text pre-training, the model is trained for 10k steps with the same batch size. The maximal learning rate of uni-modal encoders is $5e-6$, and other settings are similar to the first phase.

In terms of the model architecture, We utilize CLIP-ViT-224/16 [39] and RoBERTa [34] to initialize vision and language encoders following METER [9]. The fusion encoder consists of dual-stream cross-modal blocks of 6 layers, each with a hidden dimension of 768 and 12 heads in the multi-head attention. As for data pre-processing, the image size

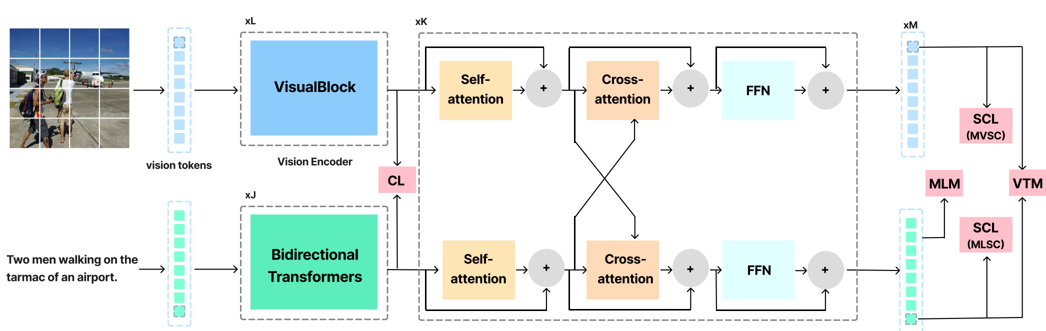


Figure 5. The whole architecture of our model.

is set to $288 \times 288$ for pre-training and $384 \times 384$ for fine-tuning, respectively. RandAugment [7] is applied for data augmentation. We resize each frame of video to $224 \times 224$ and uniformly sample 4 frames as video input. Moreover, the maximum length of input text is 50. The temperature hyper-parameter $\tau$ in Eq. (2) is set as 0.03.

### B.2. Downstream Tasks

**Visual Question Answering (VQA).** Given an image and its corresponding question, the model needs to understand visual and textual information simultaneously to predict the answer. We concatenate output [CLS] features of the image and question, and then conduct a classification on the candidates set of 3,129 answers.

**Visual Reasoning (NLVR2).** Given a pair of images and a description, the model is expected to reason whether their relationship is consistent. Specifically, this task is transformed to a binary classification problem.

**Image-Text Retrieval.** There are two sub-tasks: (1) using images as queries to retrieve texts (TR); (2) using texts as queries to retrieve images (IR). The recall ratio is employed as the evaluation metrics. We evaluate our model on Flickr30K [38] and COCO [33]. Flickr30K contains 1K images and 5K texts for evaluation, and COCO includes 5K images and 25K texts. Generally, there are five correct captions for an image.

**Video-Text Retrieval.** Similar to exiting methods [10, 14, 18], we focus on text-to-video recall metrics. Our pre-trained model is evaluated on MSRVTT [51] and LSDMC [40], which both contain 1K video-text pairs for testing.

## C. Visualization Cases

The proposed SCL encourages the global representations to learn global-to-local alignment, which implies that they have a more accurate attention distribution on local information of the other modality. To illustrate this, we show more visualization cases of [CLS] tokens' attention maps in Fig. 6.

## D. Broader Impacts

Since our model predicts content based on learned statistics of pre-training datasets and we do not filter out possible inappropriate image- or video-text pairs (e.g., of violence and blood), our model may be used to retrieve unhealthy videos for spreading.

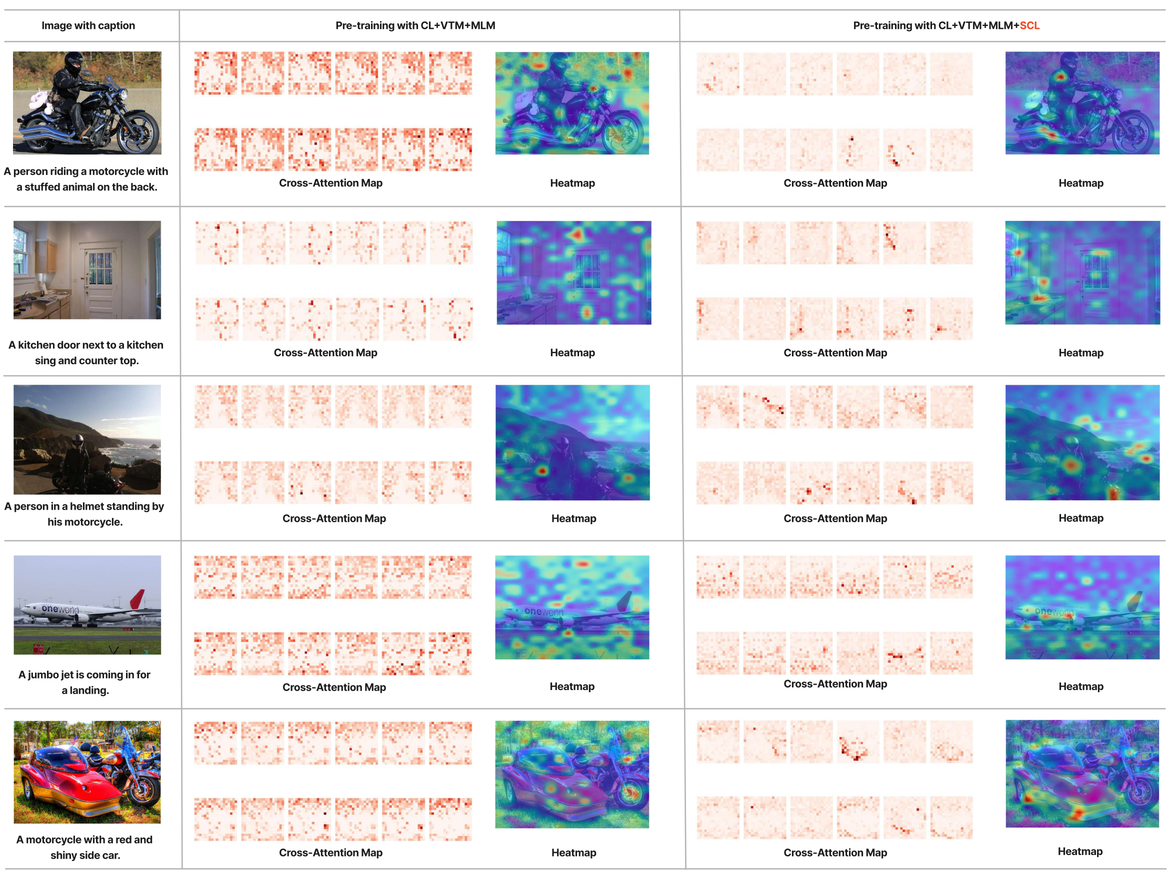


Figure 6. The cross-attention visualization of text [CLS] on the whole image for the model pre-trained with or without SCL.