Enabling Multimodal Generation on CLIP via Vision-Language Knowledge Distillation

Wenliang Dai[1], **Lu Hou**[2], **Lifeng Shang**[2], **Xin Jiang**[2], **Qun Liu**[2], **Pascale Fung**[1]
[1]Hong Kong University of Science and Technology, [2]Huawei Noah's Ark Lab
wdaiai@connect.ust.hk, pascale@ece.ust.hk,
{houlu3, shang.lifeng, jiang.xin, qun.liu}@huawei.com

Abstract

The recent large-scale vision-language pre-training (VLP) of dual-stream architectures (e.g., CLIP) with a tremendous amount of image-text pair data, has shown its superiority on various multimodal alignment tasks. Despite its success, the resulting models are not capable of multimodal generative tasks due to the weak text encoder. To tackle this problem, we propose to augment the dual-stream VLP model with a textual pre-trained language model (PLM) via vision-language knowledge distillation (VLKD), enabling the capability for multimodal generation. VLKD is pretty data- and computation-efficient compared to the pre-training from scratch. Experimental results show that the resulting model has strong zero-shot performance on multimodal generation tasks, such as open-ended visual question answering and image captioning. For example, it achieves 44.5% zero-shot accuracy on the VQAv2 dataset, surpassing the previous state-of-the-art zero-shot model with 7× fewer parameters. Furthermore, the original textual language understanding and generation ability of the PLM is maintained after VLKD, which makes our model versatile for both multimodal and unimodal tasks.

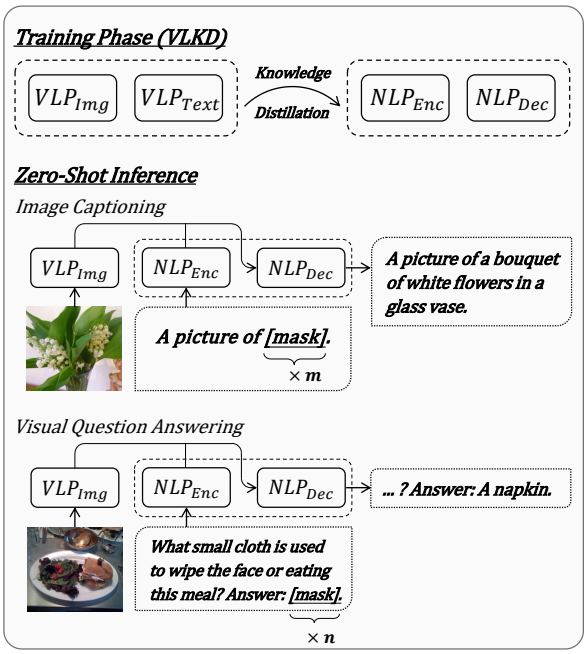


Figure 1: Intuition of our proposed approach. After VLKD, the model can fill in the masked locations with meaningful words to describe the image without further finetuning. Moreover, it can answer questions with proper reasoning over the given images and pre-trained knowledge inside PLMs, e.g., *a napkin* is for wiping the face at meals.

1 Introduction

Recent large-scale dual-stream Vision-Language Pre-training (VLP) models like CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021), have shown remarkable performance on various downstream multimodal alignment tasks, e.g., image-text retrieval and image classification. These models are pre-trained using cross-modal contrastive learning on tremendous image-text pairs and learn strong multimodal representations. Despite their success, as mentioned by Radford et al. (2021), their text encoder is relatively weak by only having a discriminative multimodal pre-training objective, which makes them incompetent on generative multimodal tasks such as image captioning and open-ended visual question answering (VQA).

Meanwhile, the Transformer-based (Vaswani et al., 2017) auto-regressive large-scale pre-trained language models (PLMs), such as GPT (Radford and Narasimhan, 2018; Brown et al., 2020), have been dominating in the natural language generation (NLG) tasks. These models are usually trained with causal self-attention, which only allows the model to attend to past outputs (unidirectional) to satisfy their generative nature. More recently, BART (Lewis et al., 2020) and T5 (Raffel et al., 2020) propose to augment the auto-regressive decoder with a bidirectional Transformer encoder to

Findings of the Association for Computational Linguistics: ACL 2022, pages 2383 - 2395
May 22-27, 2022 ©2022 Association for Computational Linguistics

further capture bidirectional information of the input. These encoder-decoder architectures excel on not only NLG but also understanding (NLU) tasks.

To tackle the aforementioned limitations of dual-stream VLP models and fully utilize PLMs, in this paper, we present ***V***ision-***L***anguage ***K***nowledge ***D***istillation (***VLKD***), a simple yet effective approach to enable CLIP to perform generative multimodal tasks through knowledge distillation. Specifically, we align the BART encoder to CLIP's joint multimodal embedding space to gain the understanding of multimodal knowledge, along with an image-conditioned language modeling loss to consort BART encoder and decoder. During training, we freeze CLIP's weights to keep its learned multimodal space. For the finetuning and inference of downstream tasks, the original CLIP text encoder is discarded, which can be interpreted as being replaced by the distilled BART. Therefore, we leverage the strengths from both sides, the expressive multimodal representation space of CLIP and the strong text generation capability of BART.

Compared to VLP from scratch, VLKD uses several magnitudes fewer image-text pairs and computational resources. As depicted in Figure 1, after VLKD pre-training, the model exhibits strong zero-shot performance on generative multimodal tasks, including open-ended VQA and image captioning. Without finetuning, it has the ability to generate answers by reasoning over the question, the visual information, and the textual knowledge embedded in the pre-trained BART. Furthermore, it can also directly generate a plausible caption given an image. Empirical results show that our model achieves 44.5% accuracy on the VQAv2 dataset and 84.6 CIDEr on COCO image caption dataset in a zero-shot manner. Moreover, the original NLU and NLG ability of BART is maintained, which makes the model versatile for both multimodal and unimodal tasks.

To summarize, our contributions are: 1) We introduce an efficient approach to distill knowledge from the dual-stream VLP model CLIP to BART. The resulting model shows strong zero-shot performance on generative multimodal tasks, as well as pure NLP tasks; 2) We exhaustively quantify these capabilities on six benchmarks under various settings; and 3) We conduct comprehensive analysis and ablation study to provide insights and grease future work on this direction.

2 Related Work

2.1 Vision-language Pre-training

Based on how the two modalities interact, recent VLP models mainly fall into two categories: single-stream and dual-stream models. Single-stream models (Chen et al., 2020; Li et al., 2019; Ramesh et al., 2021; Lin et al., 2021; Kim et al., 2021a; Shen et al., 2022) concatenate the patch-wise or regional visual features and textual embeddings and feed them into a single model. Dual-stream models (Lu et al., 2019; Radford et al., 2021; Jia et al., 2021; Zhai et al., 2021; Yao et al., 2022) use separate encoders for images and texts, allowing efficient inference for downstream multimodal alignment tasks like image-text retrieval, by pre-computing image/text features offline. However, these models can not be directly used for multimodal generation tasks. In this paper, we propose an efficient method to align the dual-stream VLP model CLIP's multimodal embedding space with a powerful PLM BART to gain multimodal generation ability.

There are also VLP models that can perform multimodal generation tasks, by expensive pre-training with objective of image-conditioned auto-regressive language modeling (Lin et al., 2021; Wang et al., 2021; Hu et al., 2021; Li et al., 2022). However, the pre-training of these models requires a large number of image-text pairs and numerous computation resources. Other models like (Agrawal et al., 2019; Li et al., 2019, 2020; Cho et al., 2021; Li et al., 2021) rely on an extra pre-trained object detector such as Faster-RCNN with labeled bounding-box data to extract image regional features offline and are less scalable.

2.2 Knowledge Distillation

Knowledge distillation (KD) in deep learning is first proposed by Hinton et al. (2015), which transfers knowledge embedded in the logits learned in a cumbersome teacher model to a smaller student model without sacrificing too much performance. Besides logits, other forms of knowledge like the intermediate representations and attentions (Jiao et al., 2019; Hou et al., 2020) have also been used in transferring the knowledge embedded in Transformer-based models. Recently, contrastive representation distillation (Tian et al., 2019) distills the knowledge from the teacher network to the student network by maximizing the mutual information between the two networks, and is recently extended to transfer the knowledge from the pre-

trained multimodal model CLIP for zero-shot detection (Gu et al., 2021) and multilingual setting (Jain et al., 2021). In this paper, we apply the conventional KD as well as the contrastive KD to transfer the knowledge from the pre-trained CLIP to BART. Besides, we also propose to transfer the knowledge in CLIP image encoder to BART decoder through the cross-attention.

3 Proposed Method

We propose to distill multimodal knowledge from CLIP to BART for generative multimodal tasks, which takes the strengths from both sides (powerful multimodal representations of CLIP and text generation ability of BART). To this end, we propose three objectives (Section 3.2). The overall architecture is illustrated in Figure 2.

3.1 Model Architecture

CLIP. CLIP (Radford et al., 2021) is a dual-stream VLP model pre-trained with a contrastive loss on 400 million image-text pairs. It consists of a text encoder which is a GPT (Radford et al., 2019) style Transformer model, and an image encoder which can be either a Vision Transformer (ViT) (Dosovitskiy et al., 2020) or Residual Convolutional Neural Network (ResNet) (He et al., 2016). CLIP learns a joint multimodal embedding space with its text encoder and image encoder aligned. Given an input image-text pair, the image encoder first reshapes the image into a sequence of 2D patches and then maps them into 1D embeddings with a prepended `[CLS]` token using a trainable linear projection. These embeddings are fed into the CLIP image encoder together with positional encodings. The output embedding of the `[CLS]` token can represent the whole image. For the text sentence, it is bracketed with `[SOS]` and `[EOS]` tokens, and the output embedding of the latter is used as the sentence-level representation. In this paper, we explore four CLIP variants, including ViT-B/16, ViT-L/14, RN50×16, and RN50×64.

BART. BART is a Transformer-based (Vaswani et al., 2017) sequence-to-sequence model that has a bi-directional encoder and a uni-directional (left-to-right) decoder, which can be seen as a generalization of the BERT (Devlin et al., 2019) and GPT (Radford and Narasimhan, 2018). It is pre-trained on 160GB text data in a self-supervised way by performing the text span infilling task with the input sentences corrupted and shuffled. Similar to the CLIP text encoder, BART also tokenizes and converts the input text into a sequence of embeddings, which are then fed into the BART encoder. BART excels at both NLG (e.g., abstractive summarization) and NLU tasks.

3.2 Training Objectives

To distill multimodal knowledge from CLIP to BART, we propose three objective functions: 1) Text-Text Distance Minimization (*TTDM*); 2) Image-Text Contrastive Learning (*ITCL*); and 3) Image-Conditioned Text Infilling (*ICTI*). During training, the model parameters of CLIP are frozen constantly, i.e. no gradients will be back-propagated through them (marked as *SG* in Figure 2), to ensure its two encoders are still aligned and the multimodal knowledge is not forgotten.

For each training batch with B image-text pairs, denote the k-th image-text pair as $\mathbf{x}^k = \{\mathbf{x}_I^k, \mathbf{x}_T^k\}$, and the output of multimodal encoders of CLIP and BART encoder as

$$\begin{aligned}
\text{CLIP}_I(\mathbf{x}_I^k) &\rightarrow \mathbf{V}^k = [\mathbf{v}_{cls}^k, \mathbf{v}_1^k, \ldots, \mathbf{v}_{n_1}^k], \\
\text{CLIP}_T(\mathbf{x}_T^k) &\rightarrow \mathbf{T}^k = [\mathbf{t}_{sos}^k, \mathbf{t}_1^k, \ldots, \mathbf{t}_{n_2}^k, \mathbf{t}_{eos}^k], \\
\text{BART}_{enc}(\mathbf{x}_T^k) &\rightarrow \mathbf{E}^k = [\mathbf{e}_{bos}^k, \mathbf{e}_1^k, \ldots, \mathbf{e}_{n_3}^k, \mathbf{e}_{eos}^k].
\end{aligned}$$

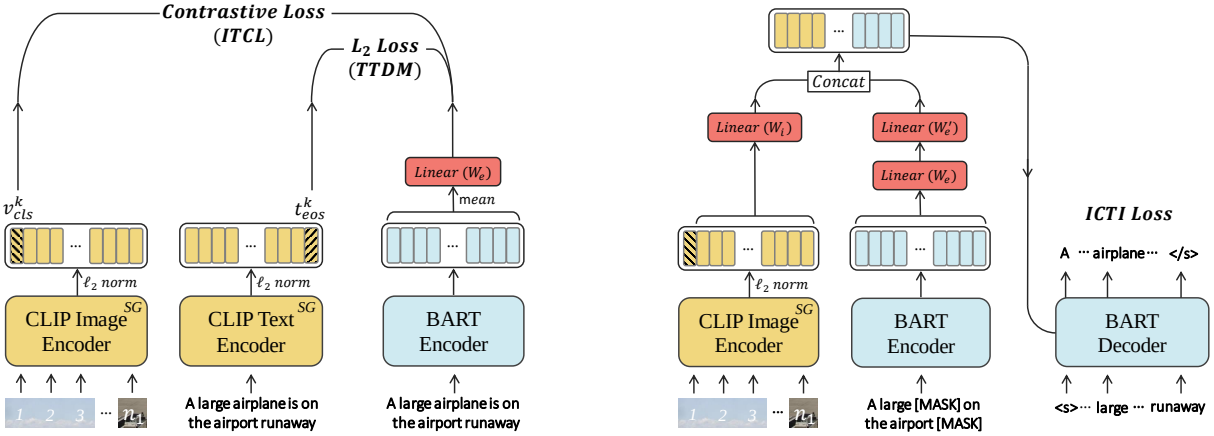
Here, n_1 is the number of image patches, n_2 and n_3 denote the sequence lengths of the text encoder of CLIP and BART, respectively. $\mathbf{v}_*^k, \mathbf{t}_*^k \in \mathbb{R}^{d_1}$ represents the ℓ_2-normalized output embedding from the CLIP image and text encoder at a certain position. $\mathbf{e}_*^k$ is the unnormalized raw output embedding from the BART encoder. In the following, we elaborate on the three distillation objectives.

3.2.1 Text-Text Distance Minimization

To align the CLIP text encoder and BART encoder, i.e. making their output representations close given the same input text, we propose to minimize the ℓ_2 distance between their sequence-level output representations. Specifically, for the k-th input text, it can be formulated as

$$\begin{aligned}
\bar{\mathbf{e}}_{\text{norm}}^k &= \mathbf{W}_e \bar{\mathbf{e}}^k / \|\mathbf{W}_e \bar{\mathbf{e}}^k\|_2, \\
\mathcal{L}_{TTDM} &= \frac{1}{B} \sum_{k=1}^{B} \|\mathbf{t}_{eos}^k - \bar{\mathbf{e}}_{\text{norm}}^k\|^2,
\end{aligned}$$

where $\bar{\mathbf{e}}^k \in \mathbb{R}^{d_2}$ is the average of all output embeddings from the BART encoder, and $\mathbf{W}_e \in \mathbb{R}^{d_1 \times d_2}$ is a weight matrix to linearly project the output of BART encoder to CLIP's multimodal space.



(a) The *TTDM* and *ITCL* losses.

(b) The *ICTI* loss.

Figure 2: Architecture of the proposed VLKD method to distill multimodal knowledge from CLIP to BART. (a) shows the *TTDM* and *ITCL* losses between the dual-stream CLIP encoders and BART encoder. (b) illustrates the *ICTI* loss for image-conditioned language modeling. *SG* denotes the *stop gradient* operation, indicating that no gradients will be back-propagated through that part of model parameters.

3.2.2 Image-Text Contrastive Learning

Contrastive training has been shown to be very effective in cross-modal representation learning (Tian et al., 2020; Sigurdsson et al., 2020; Zhang et al., 2020; Radford et al., 2021). To further adapt the BART encoder to CLIP's multimodal space, we optimize a symmetric InfoNCE loss between the output representations of the BART encoder and CLIP image encoder. The image-to-text contrastive loss $\mathcal{L}_{i2t}$ is formulated as

$$\mathcal{L}_{i2t} = -\frac{1}{B}\sum_{k=1}^{B}\log\frac{\exp\left(\mathbf{v}_{cls}^{k\top}\bar{\mathbf{e}}_{\text{norm}}^{k}/\tau\right)}{\sum_j \exp\left(\mathbf{v}_{cls}^{k\top}\bar{\mathbf{e}}_{\text{norm}}^{j}/\tau\right)},$$

where τ is a learnable temperature parameter. Different from Radford et al. (2021), we find that not clamping the τ shows a slight improvement. Similarly, the text-to-image contrastive loss $\mathcal{L}_{t2i}$ is

$$\mathcal{L}_{t2i} = -\frac{1}{B}\sum_{k=1}^{B}\log\frac{\exp\left(\mathbf{v}_{cls}^{k\top}\bar{\mathbf{e}}_{\text{norm}}^{k}/\tau\right)}{\sum_j \exp\left(\mathbf{v}_{cls}^{j\top}\bar{\mathbf{e}}_{\text{norm}}^{k}/\tau\right)}.$$

Then, the *ITCL* loss can be calculated as

$$\mathcal{L}_{ITCL} = \frac{1}{2}(\mathcal{L}_{i2t} + \mathcal{L}_{t2i}).$$

Note that when computing the *ITCL* and *TTDM* losses, we do not introduce any new linear projections to the CLIP output features to avoid destroying the pre-trained alignment between its image and text encoders. Instead, we add one linear layer (parameterized by $\mathbf{W}_e$) to project the BART encoder to CLIP's representation space and match their feature dimension.

3.2.3 Image-Conditioned Text Infilling

With only *TTDM* and *ITCL*, the BART decoder is not updated at all. To consort BART encoder and decoder, we propose to perform the text span infilling task conditioned on the corresponding image features. As depicted in Figure 2b, for the k-th image-text pair, following Lewis et al. (2020), we corrupt the input text by masking 15% of whole-word tokens with span lengths drawn from a Poisson Distribution with $\lambda = 3$.

Considering that $\mathbf{V}^k$ and $\mathbf{W}_e\mathbf{E}^k$ are already aligned in the CLIP's multimodal space through *TTDM* and *ITCL*, and having a different feature dimension with the BART decoder, we further project them to the BART decoder dimension with $\mathbf{W}_i$ and $\mathbf{W}'_e$. Then, we concatenate them together as $\mathbf{C}^k$ before feeding into the BART decoder as shown in Eq.(1). As mentioned in Section 3.1, we explore two variants of CLIP. With a slight abuse of notation, for ResNet-based CLIP, $\mathbf{V}^k$ is composed of representations of all image patches $\{\mathbf{v}_i^k\}_{i=1}^{n_1}$, while for ViT-based CLIP, $\mathbf{V}^k$ consists of the representation of the `[CLS]` token $\mathbf{v}_{cls}^k$ only.

Note that the weight matrix $\mathbf{W}'_e$ is initialized to be the pseudo-inverse of $\mathbf{W}_e$, such that text representations after the two projections $\mathbf{W}'_e\mathbf{W}_e\mathbf{E}^k$ are the closest to the original pre-trained BART encoder space at initialization[1]. The BART decoder then interacts with $\mathbf{C}^k$ through standard Transformer cross-attention layers. We optimize a lan-

[1]The pseudo inverse matrix $\mathbf{W}'_e$ satisfies $\mathbf{W}'_e = \arg\min_{\mathbf{X}} \|\mathbf{W}_e\mathbf{X} - \mathbf{I}\|_F^2$, where $\mathbf{I}$ is the identity matrix and $\|\cdot\|_F$ denotes the Frobenius Norm.

guage modeling loss $\mathcal{L}_{ICTI}$ by minimizing the negative log-likelihood in Eq.(2), in which $\mathbf{w}_j$ denotes the token to be predicted at each decoding step.

$$\mathbf{C}^k = \text{concat}(\mathbf{W}_i\mathbf{V}^k, \mathbf{W}'_e\mathbf{W}_e\mathbf{E}^k), \quad (1)$$

$$\mathcal{L}_{ICTI} = -\frac{1}{B}\sum_{k=1}^{B}\sum_{j}\log P(\mathbf{w}_j^k|\mathbf{w}_{<j}^k, \mathbf{C}^k). \quad (2)$$

The *ICTI* loss is crutial for for our methodology to work, as it not only coordinates the BART encoder and decoder, but also enables the BART decoder to understand the multimodal information by recovering texts with visual clues.

Finally, we simultaneously optimize the summation of three losses $\mathcal{L}$ as

$$\mathcal{L} = \gamma\mathcal{L}_{TTDM} + \mathcal{L}_{ITCL} + \mathcal{L}_{ICTI},$$

where γ is set to 10^3 by default, as $\mathcal{L}_{ITCL}, \mathcal{L}_{ICTI}$ are about three magnitudes larger than $\mathcal{L}_{TTDM}$.

3.3 Datasets for VLKD

Our model is trained on the Conceptual Captions (CC3M) (Sharma et al., 2018) dataset, which contains 3 million image-text pairs crawled from the Internet. For larger model variants (ViT-L/14 and RN50x64), we further include the Visual Genome Caption data which contains ∼700K image-text pairs. No images for pre-training appear in the downstream datasets. Compared to previous VLP work (Radford et al., 2021; Jia et al., 2021; Wang et al., 2021), VLKD is much cheaper by leveraging several magnitudes less data. Furthermore, we experiment with even smaller data (1M, 100K) by uniformly sampling a subset of CC3M to test the limit of dataset size of VLKD, with results discussed in Section 5.

4 Experiments

To demonstrate the effectiveness of VLKD, we evaluate it on generative multimodal tasks for both zero-shot and finetuning. Specifically, we test the image captioning task, and also the VQA task under the open-ended scenario. Furthermore, we also run the model on NLU and NLG tasks to investigate the influence of VLKD on the text processing ability of the original pre-trained BART.

4.1 Finetuning Datasets

Image Captioning. Image captioning requires the model to generate a relevant description given an image. We use the COCO image caption dataset (Lin et al., 2014) with the Karpathy split (Karpathy and Fei-Fei, 2017). Additionally, we use the NoCaps (Agrawal et al., 2019) dataset to test the model performance when there are out-of-domain objects.

Open-Ended VQA. Unlike previous works (Anderson et al., 2018; Chen et al., 2020; Li et al., 2020; Yu et al., 2021a; Zhang et al., 2021; Kim et al., 2021b) that treat the VQA task as a discriminative problem, we let the model generate answers freely, which is more aligned with the real-world scenario of this task. We use the standard VQAv2 (Goyal et al., 2017), and also OK-VQA (Marino et al., 2019) which requires knowledge to answer questions correctly.

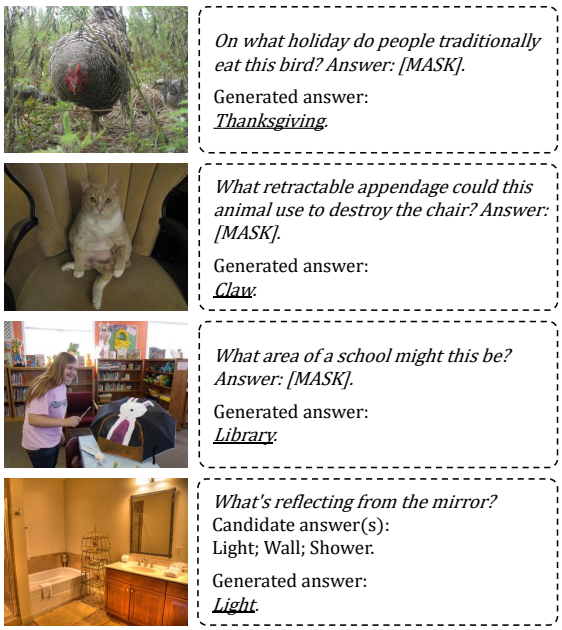
NLU and NLG. For NLU, we test our model on the GLUE benchmark (Wang et al., 2019), which consists of nine text classification tasks. We exclude the WNLI task as it is problematic[2]. For NLG, we test the abstractive summarization task on XSUM (Narayan et al., 2018) dataset, which requires the model to comprehend long texts and generate short summaries with key information.

4.2 Implementation Details

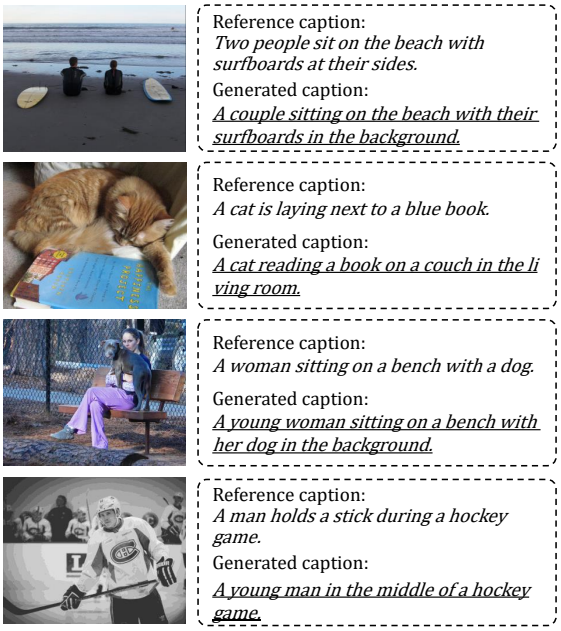
We use BART-large as the pre-trained backbone NLP model, which has 12 layers in both encoder and decoder with a hidden size of 1024 and 16 heads in each multi-head attention (MHA) layer. In total, it contains 406M parameters. For the pre-trained CLIP (Radford et al., 2021) model, we report four variants with different visual backbones, including ViT-B/16, ViT-L/14, RN50×16, and RN50×64.

We use 64 Nvidia V100 GPUs for VLKD and 8 for the finetuning of downstream tasks. In total, we pre-train the model for 10 epochs, which takes about 5 hours. We use a batch size of 4608 for ViT-B/16 and ViT-L/14, 4096 for RN50x16 and 3840 for RN50x64. All of the models are optimized by the AdamW (Loshchilov and Hutter, 2019) optimizer. The learning rate is warmed up to $2.4e^{-4}$ within the first 2% steps and then linearly decay to 0. More information of VLKD pre-training and the finetuning of each downstream task can be found in Appendix A.

[2] https://gluebenchmark.com/faq



(a) Zero-shot VQA.



(b) Zero-shot image captioning.

Figure 3: Examples of (a) zero-shot VQA and (b) image captioning. Our model shows the ability to recognize visual objects and generate appropriate sentences based on their properties and relationship. Furthermore, the model can bind visual objects to text conceptual knowledge that is learned in the PLMs when generating answers given questions.

4.3 Multimodal Zero-Shot Evaluation

Benefit from the knowledge distillation, especially the *ICTI* loss, our model can perform various downstream multimodal tasks in a zero-shot manner.

4.3.1 Zero-Shot Image Captioning

During knowledge distillation, the *ICTI* loss can be seen as a simple version of the image captioning task, which asks the model to fill in the corrupted locations of image descriptions. If the masking ratio increases to 100%, it reduces to the image captioning task. Therefore, it is intuitive to test the zero-shot performance of our model.

Following Radford et al. (2021) and Wang et al. (2021), we compose the input with a text prompt and also m mask tokens, i.e., "A picture of `[MASK]`$\times m$.", for the model to generate the caption for the image. The zero-shot results are included in Table 1. Our zero-shot model achieves comparable overall performance to the finetuned UpDown (Agrawal et al., 2019) model on NoCaps dataset. As shown in Figure 3b, the zero-shot generated captions are plausible with correct objects, relationships, and actions. However, sometimes details like colors could be omitted.

In our experiments, we use $m = 6$ for COCO and $m = 8$ for NoCaps. Although it could potentially limit the length of generation, we find that it has negligible influence to the performance, as for each `[MASK]` token, the model is learned to fill one to three tokens depending on the context. Furthermore, this could be used to control the length of generated texts for different senarios. See Section 5 for a more detailed discussion about the effects of number of the masks.

4.3.2 Zero-Shot VQA

Zero-shot VQA is much more challenging than image captioning, as it requires reasoning over both the image and question, which is very different from the *ICTI* loss during the knowledge distillation. As illustrated in Figure 1, we construct the input by appending a text prompt "Answer: `[MASK]`$\times n$." to the question Given the context (image+question+prompt), the model is required to predict the answer by *recovering* the textual token in the `[MASK]` positions. In our experiments, we use $n = 2$ for the VQAv2, which is found performing best among $n \in \{1, 2, 3\}$.

In Table 2, compared to the strong baseline Frozen (Tsimpoukelli et al., 2021), our model improves the zero-shot accuracy by 13.1% on the VQAv2 validation set and 7.4% on the OK-VQA test set with 7$\times$ fewer parameters, indicating the

Methods	#Pretrain Image-text Pairs	OD	OT	COCO Caption Karpathy Test				NoCaps Validation							
								In		Near		Out		Overall	
				B@4	C	M	S	C	S	C	S	C	S	C	S
BUTD†	1.5M	✓	✓	36.3	120.1	27.7	21.4	80.0	12.0	73.6	11.3	66.4	9.7	73.1	11.1
$\text{OSCAR}^{\dagger}_{\text{Large}}$	6.5M	✓	✓	41.7	140.0	30.6	24.5	85.4	11.9	84.0	11.7	80.3	10.0	83.4	11.4
$\text{VinVL}_{\text{Large}}$	6.5M	✓	✓	41.0	140.9	31.1	25.2	103.7	13.7	95.6	13.4	83.8	11.9	94.3	13.1
VL-T5	9.2M	✓	✗	34.6	116.1	28.8	21.9	-	-	-	-	-	-	-	-
VL-BART	9.2M	✓	✗	34.2	114.1	28.4	21.3	-	-	-	-	-	-	-	-
$\text{LEMON}_{\text{Huge}}$	203M	✓	✓	42.6	145.5	31.4	25.5	118.0	15.4	116.3	15.1	120.2	14.5	117.3	15.0
$\text{SIMVLM}_{\text{Huge}}$	1.8B	✗	✗	40.6	143.3	33.7	25.4	113.7	-	110.9	-	115.2	-	112.2	-
VLKD *(Zero-shot)*															
ViT-B/16	3M	✗	✗	16.7	58.3	19.7	13.4	-	-	-	-	-	-	-	-
RN50×16	3M	✗	✗	18.2	61.1	20.8	14.5	52.6	9.7	52.9	9.6	58.6	9.3	54.0	9.6
RN50×64	3.7M	✗	✗	25.8	85.1	23.1	16.9	64.8	13.6	62.3	13.6	66.9	9.9	63.6	12.8
VLKD *(Finetuned)*															
ViT-B/16	3M	✗	✗	37.2	128.0	28.8	22.4	-	-	-	-	-	-	-	-
RN50×16	3M	✗	✗	38.9	131.1	29.6	23.9	92.3	12.6	82.0	11.8	70.3	10.4	81.1	11.7
RN50×64	3.7M	✗	✗	40.3	135.7	30.5	24.3	105.1	14.5	99.7	13.8	90.2	12.1	97.6	13.6

Table 1: Results on the COCO caption (Karpathy test set) and NoCaps (validation set). B@4, C, M, and S denote BLEU-4, CIDEr, METEOR, and SPICE, respectively. OD and OT indicate whether object detectors and object tags are used or not. Numbers of previous models are taken from (Anderson et al., 2018; Li et al., 2020; Zhang et al., 2021; Cho et al., 2021; Hu et al., 2021; Wang et al., 2021). Models marked by † additionally use the constrained beam search (CBS) (Anderson et al., 2017) for the NoCaps dataset. Note that LEMON and SIMVLM use significantly more pre-training data and have more trainable model parameters than the others.

Methods	#Params	VQAv2 val / test-dev	OK-VQA test
Generative (Open-ended)			
Frozen *(Zero-shot)*	7B	29.5 / -	5.9
Frozen *(Finetuned)*		48.4 / -	19.6
VLKD *(Zero-shot)*	< 1B		
RN50×16		37.4 / 38.2	9.9
ViT-B/16		38.6 / 39.7	10.5
ViT-L/14		42.6 / 44.5	13.3
VLKD *(Finetuned)*			
RN50×16		67.4 / 68.8	36.2
ViT-B/16		69.3 / 69.8	36.3
ViT-L/14		73.9 / 74.5	39.0
Discriminative			
$\text{UNITER}_{\text{Large}}$	-	- / 73.8	-
$\text{OSCAR}_{\text{Large}}$	-	- / 73.6	-
$\text{VinVL}_{\text{Large}}$	-	- / 76.5	-
$\text{SIMVLM}_{\text{Base}}$	-	- / 77.9	-

Table 2: Accuracies(%) on the VQAv2 and OK-VQA datasets. We categorize models into two parts: answer questions in a generative or discriminative way.

efficiency and effectiveness of VLKD. Our model achieves 44.5% zero-shot accuracy on the VQAv2 test-dev set, which to the best of our knowledge is the new state-of-the-art. Furthermore, as shown in Figure 3a, our model can bind visual objects to conceptual knowledge stored in the PLM to answer

Model	In-domain	Out-of-domain
UNITER	**74.4**	10.0
VL-T5	71.4	13.1
VL-BART	72.1	13.2
VLKD (ViT-L/14)	74.9	**23.4**

Table 3: Accuracies(%) on VQAv2 Karpathy test-split.

questions. For example, it connects the visual object *Turkey* with the traditional food people usually eat at the *Thanksgiving* festival.

4.4 Multimodal Finetuning Evaluation

When finetuning VLKD on downstream multimodal tasks, we keep the same input format as zero-shot to obtain outputs in a generative way. The CLIP model parameters are still frozen during finetuning.

4.4.1 Finetuning Image Captioning

In Table 1, we demonstrate that our model can achieve decent performance when finetuned on the COCO dataset. The SCST CIDEr optimization method (Rennie et al., 2017) is used to further improve the performance. Our model outperforms VL-T5/BART (Cho et al., 2021) without using an extra object detector, which is fairly time-consuming as explained by Kim et al. (2021b). Compared to state-of-the-art models, however,

Model	CoLA	SST-2	RTE	MRPC	QQP	MNLI	QNLI	Avg.
$\text{BERT}^{\diamond}_{\text{LARGE}}$ (Devlin et al., 2019)	60.6	93.2	70.4	82.9/88.0	91.3/87.9	86.4	92.3	82.6
$\text{BART}^{\diamond}_{\text{LARGE}}$ (Lewis et al., 2020)	**62.8**	**96.6**	**87.0**	86.7/90.4	**92.5/89.3**	**90.0**	**94.9**	**87.2**
VisualBERT† (Li et al., 2019)	38.6	89.4	56.6	71.9/82.1	89.4/86.0	81.6	87.0	74.0
UNITER† (Chen et al., 2020)	37.4	89.7	55.6	69.3/80.3	89.2/85.7	80.9	86.0	73.1
VL-BERT† (Su et al., 2020)	38.7	89.8	55.7	70.6/81.8	89.0/85.4	81.2	86.3	73.6
VilBERT† (Lu et al., 2019)	36.1	90.4	53.7	69.0/79.4	88.6/85.0	79.9	83.8	72.1
LXMERT† (Tan and Bansal, 2019)	39.0	90.2	57.2	69.8/80.4	75.3/75.3	80.4	84.2	71.6
SIMVLM‡ (Wang et al., 2021)	46.7	90.9	63.9	75.2/84.4	90.4/87.2	83.4	88.6	77.4
VLKD (RN50×16)	59.1	95.5	81.2	**87.5/91.1**	92.1/89.2	89.6	94.3	85.7

Table 4: Results on the GLUE development set (single task single models). We report the Matthews correlation for CoLA, accuracy/F1 for MRPC and QQP, and accuracy for the rest of the tasks. The performance of models that are marked by $\diamond$ are taken from (Lewis et al., 2020), $\dagger$ are from (Iki and Aizawa, 2021), and $\ddagger$ are from (Wang et al., 2021). Compared to other VLP models, our VLKD model has a great advantage in text-only NLP tasks.

there is still a small performance gap, which we conjecture is mainly due to their usage of object detector/tags and much more pre-training image-text pairs. We also evaluate our VLKD models with ResNet visual backbones on the NoCaps dataset (Table 1). For zero-shot image caption, the CIDEr score on the out-of-domain set is even higher than the in- and near-domain sets, which shows the generalization of our knowledge distillation method to common visual objects. After finetuned on the COCO training set, the performance on NoCaps of our model with the RN50×64 backbone is comparable to the state-of-the-art models.

4.4.2 Finetuning VQA

From Table 2, the best performance of VQAv2 is achieved by VLP models that tackle this task in a discriminative way with a set of pre-defined answers. However, this approach does not generalize to real-world scenarios and cannot be directly applied to more diverse datasets (e.g., OK-VQA). Differently, Frozen (Tsimpoukelli et al., 2021) and our proposed VLKD formulate VQA as a generative problem to generate answers conditioned on the questions and images in an open-ended manner, which also enables zero-shot VQA. Specifically, for each question-answer pair in the VQAv2 dataset, we optimize the model to generate the answer with the cross-entropy loss and a label-smoothing of 0.1. The loss is weighted by the weight of each answer candidate. In addition, we augment the training data with VG-QA (Krishna et al., 2016).

Furthermore, following (Cho et al., 2021), we test the performance on out-of-domain questions with rare answers using the Karpathy test-split. As shown in Table 3, our method shows a salient advantage on out-of-domain questions due to the benefit from VLKD and its generative nature without defining the answer list.

Model	ROUGE-1	ROUGE-2	ROUGE-L
$\text{BART}_{\text{Large}}$	45.14	22.27	37.25
VLKD	44.86	22.06	36.95

Table 5: Results of abstractive summarization on XSUM. We use the best performing checkpoint of the RN50×16 variant.

4.5 Evaluation of NLU and NLG

Table 4 shows results on the GLUE benchmark. Although prior VLP models are either initialized from the pre-trained BERT model, or trained by a text-only language modeling loss together with the vision-language (VL) losses, they generally suffer from the weakened performance of NLU. For example, SIMVLM performs significantly worse than BART, though trained with five times more textual data. We speculate that the weakened NLU ability of these models is caused by the catastrophic forgetting of the pre-trained BERT weights during the multimodal pre-training. Moreover, simultaneous optimization of multimodal and text-only objectives potentially shifts the latter to be an auxiliary loss, making the NLP ability not as effective.

On the other hand, the resulting model of VLKD performs only slightly worse than the original BART and significantly outperforms BERT, as the original knowledge embedded in BART is well maintained.

Additionally, as presented in Table 5, we also run VLKD on the abstractive summarization task to evaluate its NLG performance, since BART-based methods excel on the summarization (Lewis et al., 2020; Dou et al., 2021; Yu et al., 2021b). The gap between VLKD and its backbone BART is negligible. Overall, we empirically demonstrate that VLKD enables the backbone PLM to perform multimodal tasks without hurting its original NLP ability.

5 Ablation Study

Knowledge Distillation Objectives. Table 6 shows the ablation on the knowledge distillation objectives, except the *ICTI* loss which is necessary for our method to work. Without *TTDM* or *ITCL*, we observe a clear degradation of zero-shot performance on both VQAv2 and COCO image caption datasets. It is worth noting that *ITCL* contributes more to the image captioning task, which requires a deeper perception of visual features to generate captions. Oppositely, *TTDM* helps more for the VQA task, which involves reasoning over the question and image features. Removing both of them incurs a large performance drop, which demonstrates the importance of aligning the embedding space between CLIP and BART.

Model	VQAv2 (val)	COCO Caption (test)
$\text{VLKD}^{\text{ViT-B/16}}_{\text{ZERO-SHOT}}$	38.6	58.3
w/o *TTDM*	35.5	55.7
w/o *ITCL*	36.3	54.1
w/o *Both*	30.1	48.6

Table 6: Ablation study on three distillation objectives.

Number of Masks. Furthermore, we also test the influence of the number of masks for zero-shot image captioning in Table 7. As discussed in Section 4.3.1, it has a trivial influence as the model learns to fill a variable length of tokens for each masked position. We achieve the best performance on the COCO caption dataset when $m = 6$ and NoCaps when $m = 8$.

#masks	5	6	7	8
CIDEr	59.7	61.1	60.6	59.6

Table 7: Zero-shot image captioning on COCO test set using VLKD (RN50×16), with varying number of masks.

Dataset Size of Distillation. In Table 8, we vary the size of dataset used for knowledge distillation. VLKD only has a slight performance drop when the size is reduced from 3M to 1M, and a sharp drop when further reduced to 100K.

	VQAv2 (val)	COCO Caption (test)
$\text{VLKD}_{3\text{M}}$	38.6	58.3
$\text{VLKD}_{1\text{M}}$	38.3	56.2
$\text{VLKD}_{100\text{K}}$	33.8	45.1

Table 8: Zero-shot performance of VLKD (ViT-B/16) on two datasets, with varying dataset size for distillation.

Unfreeze CLIP Weights. To quantitatively measure the importance of freezing the model weights of CLIP during the VLKD pre-training, we tried unfreezing CLIP's weights and conduct the VLKD pre-training using the ViT-B/16 variant on CC3M without modifying other settings. It achieves 31.7 zero-shot accuracy on the VQAv2 validation set and 44.8 CIDEr on the COCO Caption test set. We speculate that unfreezing CLIP harms its pre-trained multimodal space, which further downgrades the performance of VLKD.

6 Conclusion

Recent dual-stream VLP models (e.g., CLIP) are powerful in various multimodal classification and retrieval tasks. However, their ability of multimodal generation or pure NLP tasks is highly restricted. In this paper, we propose a novel knowledge distillation method to efficiently align CLIP's multimodal encoders and BART's textual encoder to the same mutlimodal space, as well as a cross-modal LM loss to consort BART encoder and decoder. This enables multimodal generation under zero-shot and also fully-finetuned settings without losing the original BART's NLP ability. Empirical results show that our model achieves new state-of-the-art zero-shot performance on VQA and excellent performance on both NLP and multimodal tasks when finetuned, demonstrating the effectiveness of our proposed method.

References

Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. 2019. nocaps: novel object captioning at scale. In *International Conference on Computer Vision*, pages 8947–8956.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2017. Guided open vocabulary image captioning with constrained beam search. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6077–6086.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*.

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

Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying vision-and-language tasks via text generation. Preprint arXiv:2102.02779.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Zi-Yi Dou, Pengfei Liu, Hiroaki Hayashi, Zhengbao Jiang, and Graham Neubig. 2021. GSum: A general framework for guided neural abstractive summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4830–4842, Online. Association for Computational Linguistics.

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

Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. 2021. Zero-shot detection via vision and language knowledge distillation. Preprint arXiv:2104.13921.

Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. Preprint arXiv:1503.02531.

Lu Hou, Zhiqi Huang, Lifeng Shang, Xin Jiang, Xiao Chen, and Qun Liu. 2020. Dynabert: Dynamic bert with adaptive width and depth. In *Advances in Neural Information Processing Systems*, volume 33.

Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. 2021. Scaling up vision-language pre-training for image captioning. *ArXiv*, abs/2111.12233.

Taichi Iki and Akiko Aizawa. 2021. Effect of vision-and-language extensions on natural language understanding in vision-and-language models. Preprint arXiv:2104.08066.

Aashi Jain, Mandy Guo, Krishna Srinivasan, Ting Chen, Sneha Kudugunta, Chao Jia, Yinfei Yang, and Jason Baldridge. 2021. Mural: Multimodal, multitask retrieval across languages. Preprint arXiv:2109.05125.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2019. Tinybert: Distilling bert for natural language understanding. Preprint arXiv:1909.10351.

Andrej Karpathy and Li Fei-Fei. 2017. Deep visual-semantic alignments for generating image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:664–676.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021a. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021b. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123:32–73.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Annual Meeting of the Association for Computational Linguistics*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation.

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

Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. 2021. Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. In *Annual Meeting of the Association for Computational Linguistics*.

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. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*.

Junyang Lin, Rui Men, An Yang, Chang Zhou, Ming Ding, Yichang Zhang, Peng Wang, Ang Wang, Le Jiang, Xianyan Jia, et al. 2021. M6: A chinese multimodal pretrainer. Preprint arXiv:2103.00823.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! Topic-aware convolutional neural networks for extreme summarization. In *Conference on Empirical Methods in Natural Language Processing*.

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. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*.

Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. Preprint arXiv:2102.12092.

Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1179–1195.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Annual Meeting of the Association for Computational Linguistics*, pages 2556–2565.

Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. 2022. How much can CLIP benefit vision-and-language tasks? In *International Conference on Learning Representations*.

Gunnar A. Sigurdsson, Jean-Baptiste Alayrac, Aida Nematzadeh, Lucas Smaira, Mateusz Malinowski, João Carreira, Phil Blunsom, and Andrew Zisserman. 2020. Visual grounding in video for unsupervised word translation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*.

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

Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2019. Contrastive representation distillation. In *International Conference on Learning Representations*.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2020. Contrastive multiview coding. In *European Conference on Computer Vision*, pages 776–794.

Maria Tsimpoukelli, Jacob Menick, Serkan Cabi, S. M. Ali Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. Preprint arXiv:abs/2106.13884.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*.

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

Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. 2022. FILIP: Fine-grained interactive language-image pre-training. In *International Conference on Learning Representations*.

Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. 2020. Large batch optimization for deep learning: Training bert in 76 minutes. In *International Conference on Learning Representations*.

Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021a. Ernie-vil: Knowledge enhanced vision-language representations through scene graph. In *AAAI Conference on Artificial Intelligence*.

Tiezheng Yu, Wenliang Dai, Zihan Liu, and Pascale Fung. 2021b. Vision guided generative pre-trained language models for multimodal abstractive summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3995–4007, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. 2021. Lit: Zero-shot transfer with locked-image text tuning. *CoRR*, abs/2111.07991.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. Vinvl: Revisiting visual representations in vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5575–5584.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. 2020. Contrastive learning of medical visual representations from paired images and text. Preprint arXiv:2010.00747.

Hyper-paramters	Values
Batch size	4608 (ViT-B/16 and ViT-L/14), 4096 (RN50x16), 3840 (RN50x64)
Optimizer	AdamW, $\beta = (0.99, 0.999)$
Learning rate	2.4e-4
Weight decay	0.01
Eps	1e-6
Temperature τ	Initialized to 0.07
Warmup steps	2%
#Epochs	10
Gradient clipping	3.0

Table 9: Hyper-parameters of VLKD pre-training.

Hyper-paramters	VQA	Image captioning
Batch size	72	64
Total epochs	10	10
#Masks	2	6 (COCO), 8 (NoCaps)
Beam search size	1 (greedy)	6
Optimizer	AdamW, $\beta = (0.99, 0.999)$	
Learning rate	1e-4	
Weight decay	0.01	
Eps	1e-8	
LR warmup	First epoch	
Gradient clipping	5.0	

Table 10: Hyper-parameters for two multimodal tasks.

A Hyper-parameters

In this section, we show the hyper-parameters of vision-language knowledge distillation (VLKD), as well as downstream task finetuning.

For VLKD, the hyper-parameters are shown in Table 9, for both two CLIP variants we explored. For finetuning multimodal downstream tasks, we use the hyper-parameters shown in Table 10. Within each task, we use the same setting for multiple datasets.

For the GLUE benchmark, we use the LAMB optimizer (You et al., 2020) to train for 10 epochs. We conduct a hyper-parameter grid search with batch size={16, 32, 64}, lr={1e-4, 5e-4, 1e-3}, weight decay={1e-4, 1e-3}. We warm up the learning rate in the first epoch, then linearly decay it to zero.

For XSUM, we directly follow the hyper-parameters used in Lewis et al. (2020).

B More Examples of Zero-shot Inference

In Figure 4, we show more examples of zero-shot image captioning. In Figure 5, we depict more cases of the results of zero-shot open-ended VQA.

Figure 4: More examples of zero-shot image captioning.

Figure 5: More examples of zero-shot VQA.