

PERCEIVER-VL: Efficient Vision-and-Language Modeling with Iterative Latent Attention

Zineng Tang* Jaemin Cho* Jie Lei Mohit Bansal
 UNC Chapel Hill

{terran, jmincho, jielei, mbansal}@cs.unc.edu

Abstract

We present PERCEIVER-VL, a vision-and-language framework that efficiently handles high-dimensional multi-modal inputs such as long videos and text. Powered by the iterative latent-cross-attention of Perceiver, our framework scales with linear complexity, in contrast to the quadratic complexity of self-attention used in many state-of-the-art transformer-based models. To further improve the efficiency of our framework, we also study applying LayerDrop on cross-attention layers and introduce a mixed-stream architecture for cross-modal retrieval. We evaluate PERCEIVER-VL on diverse video-text and image-text benchmarks, where PERCEIVER-VL achieves the lowest GFLOPs and latency, while maintaining competitive performance. In addition, we also provide comprehensive analyses over various aspects of our framework, including pretraining data, scalability of latent size and input size, dropping cross-attention layers at inference to reduce latency, modality aggregation strategy, positional encoding, and weight initialization strategy.¹

1. Introduction

During the past several years, there has been increased interest in vision-and-language learning. Many recent models [69, 52, 9, 68, 45, 72, 41] adopt the transformer [74] architecture to encode vision-and-language inputs. These methods have improved the performance of various tasks, such as text-based image/video retrieval [8, 57, 80, 62, 7, 2, 43, 40] and visual question answering [3, 29, 79, 42, 82]. However, the transformer is based on the self-attention module [74] with a quadratic computational cost in relation to its input length. This makes it difficult for models to process high-dimensional data, such as long videos.

To this end, we propose PERCEIVER-VL, an end-to-

*equal contribution

¹Our code and checkpoints are available at: https://github.com/zinengtang/Perceiver_VL

end vision-and-language architecture that efficiently handles high-dimensional multi-modal inputs. PERCEIVER-VL is built on the iterative latent cross-attention of the recently proposed PERCEIVER [28, 27]. Concretely, we map a multi-modal input array of size M to a latent array of size N with cross-attention. This changes the computational complexity of the attention modules from $O(M^2)$ to $O(NM)$. Since vision-and-language models often handle very long input arrays (e.g., $M > 1000$),² this greatly improves the efficiency for vision-and-language tasks. To further enhance the efficiency of our framework, we also study reducing the number of cross-attention layers based on LayerDrop [17] and using a mixed-stream architecture for cross-modal retrieval tasks. By varying the number of cross-layer attention layers that take the most computation, we allow users to flexibly control the latency at inference. The mixed-stream architecture combines the widely used single-stream and multi-stream architectures and improves the retrieval performance of the multi-stream architecture with minimum increase in latency.

We evaluate PERCEIVER-VL on various video-text (MSRVTT, DiDeMo, LSMDC, ActivityNet, TGIF-QA, MSRVTT-QA) and image-text (Flickr30k, VQAv2, NLVR²) tasks. Overall, PERCEIVER-VL achieves performance competitive to recent vision-and-language models, while maintaining significantly higher efficiency with the lowest GFLOPs and latency. We demonstrate that PERCEIVER-VL scales more efficiently than the transformer-based architecture with respect to video length and frame size. In addition, we show that our method also allows for flexible adaptions to further improve its efficiency: (1) Decreasing the size of latent array during finetuning reduces the computation significantly, with only minimal accuracy drop; (2) Mixed-stream architecture achieves a reasonable accuracy-latency trade-off: higher accuracy than multi-stream and lower latency than

²The Frozen-in-time [4] visual encoder takes video inputs of frame length 8, frame size 224×224, and patch size 16×16. The resulting input length $M=(224/16)^2 \times 8=1568$, which is much larger than a typical latent array size $N=128$ in PERCEIVER-VL.

single-stream for text-to-video retrieval. (3) We apply LayerDrop during training. This allows users to control the latency by reducing the number of cross-attention layers at inference, again with only minimal accuracy drop. Moreover, LayerDrop also acts as a regularizer; training with LayerDrop improves model performance. Lastly, we conduct a comprehensive ablation study, including weight initialization, pretraining dataset, and comparing modality aggregation methods. We find it helpful to initialize the parameters from ViT [16] and CLIP [59] and pretrain jointly on video-text and image-text pairs. We do not find a meaningful difference in whether to combine the two input modalities in joint or separate attention modules, and whether to use learned [21, 74] or Fourier [66, 54, 28] positional encoding for the latent array.

Our contributions can be summarized as: (1) We propose PERCEIVER-VL, an efficient vision-and-language framework with linear scalability, on-demand depth reduction, and mixed-stream retrieval architecture. (2) We demonstrate that our framework achieves significantly higher efficiency than recent transformer-based models on various vision-and-language benchmarks, with overall competitive performance. (3) We provide a comprehensive analysis of the efficiency, architectural components, and training strategies of our framework. We hope that our research allows the community to use the highly efficient framework for diverse vision-and-language tasks and inspires future research.

2. Related Work

2.1. Efficient Transformers

Many research works have proposed to reduce the quadratic computation complexity of self-attention in transformers [74] based on different methods, including hashing [35], sparse attention [11, 5, 73], kernel trick [32], low-rank key/value projection [76], blockwise attention [58], past memory compression [60], and inducing point methods [39]. Unlike these methods, PERCEIVER [28] proposes using iterative cross-attention to map an input array to a smaller latent array and apply self-attention to the latent array, which makes the computation scale linearly. PERCEIVER-IO [27] adds a decoder to PERCEIVER to allow the model to tackle various downstream tasks with structured prediction.

To our knowledge, the iterative cross-attention of PERCEIVER for multi-modal inputs has only been studied on audio-video autoencoding task [27]. In this work, we present PERCEIVER-VL, which extends the PERCEIVER framework in vision-and-language domain. We also evaluate PERCEIVER-VL on diverse video-text and image-text benchmarks and conduct extensive experiments to analyze its efficiency. In addition, we introduce new techniques including cross-attention drop and mixed-stream architecture

for cross-modal retrieval.

2.2. Vision-and-Language Pretraining

Large-scale pretraining of transformers [74, 14] has achieved huge success in natural language processing [51, 81, 38, 15, 65, 61, 13]. Following this success, image-text [69, 52, 9, 47, 86, 44, 41, 12, 59, 30] and video-text [68, 88, 45, 72, 84, 83, 70, 71] multi-modal transformers have achieved improvements on various vision-and-language tasks [3, 8, 80, 87, 42]. Such models take both visual and textual inputs and are pretrained on large image-text/video-text pairs, with multi-modal masked language modeling and vision-text matching objectives under a standard transformer architecture [74]. One prominent issue with such models is that they are hard to scale because of the quadratic computation cost of a standard transformer. In this work, we propose a new vision-and-language pretraining framework that scales more efficiently than the transformer-based frameworks mentioned above.

3. Perceiver-VL

PERCEIVER-VL architecture consists of an input array, a latent array, and an encoder-decoder network. In the following, we explain the details of each component and how PERCEIVER-VL processes high-dimensional vision-and-language data efficiently. In Fig. 1, we illustrate the PERCEIVER-VL architecture.

3.1. Vision and Language Embedding

We extend the single-modal input array of PERCEIVER to the vision-and-language domain, creating the input array c as a concatenation of visual and text embeddings. The embeddings are created as the sum of (1) modality embedding; (2) temporal embedding; (3) positional embedding; and (4) patch/token embedding. Modality embedding is a learned embedding of a binary modality indicator $\in \{V, T\}$. Temporal embedding is a learned embedding of input video frames $\in \{1 \dots L^V\}$, where L^V is the frame length. Note that temporal embedding is only used for videos; we do not use temporal embedding for images or text. Positional embedding is a learned embedding of 2D patches $\in \{1 \dots L^P\}$ for image/video or token indices for text $\in \{1 \dots L^T\}$, where L^P is the number of patches for images, and L^T is the number of text tokens. Patch embedding is learned with a linear projection of non-overlapping image/video input patches (e.g., 32×32 pixels) [16]. We treat an image as a single-frame video so that our model can flexibly process image and video input with a single architecture [4]. Token embedding is a learned embedding of text tokens.

3.2. Iterative Mapping to Low-Dim Latent Space

Following [28], PERCEIVER-VL tames the quadratic computation complexity of self-attentions over high-

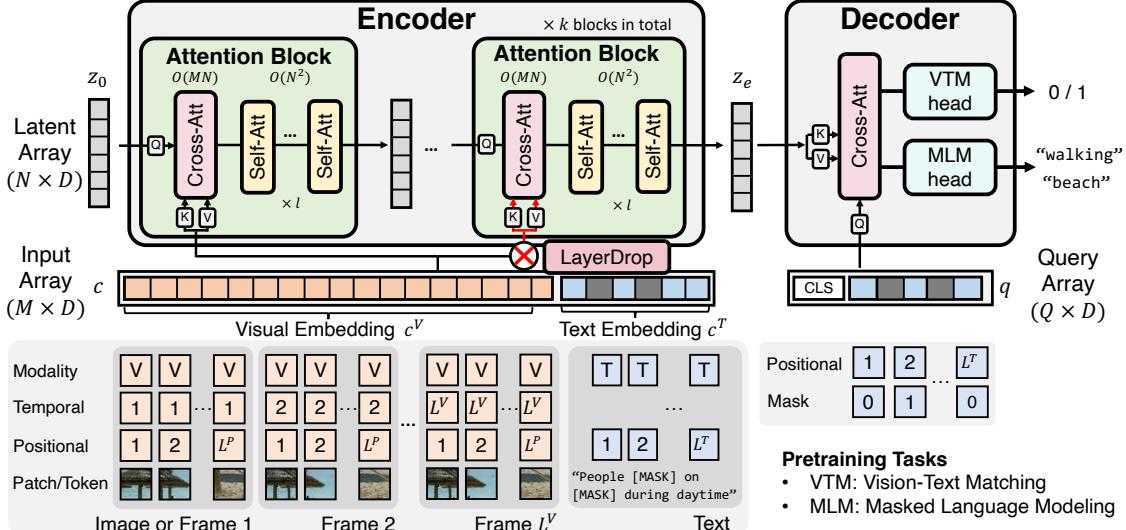


Figure 1. PERCEIVER-VL architecture for efficient vision-and-language pretraining. The encoder maps the input array c of length M (Sec. 3.1) to the latent array z of length N via iterative cross-attentions (Sec. 3.2). Since the latent arrays are smaller than input arrays for typical vision-and-language data ($N \ll M$), cross-attention based encoding has higher efficiency than standard self-attention based encoding for vision-and-language tasks. In addition, we also study dropping cross-attentions to improve latency on demand via reducing model depth (Sec. 3.3). The decoder performs structured prediction based on a cross-attention with the latent encoding z_e and task-specific query array q (Sec. 3.4).

dimensional inputs, by introducing a latent array z of size N (see ‘Latent Array’ in Fig. 1) that aggregates information from an input array c of size M via iterative cross-attentions (see ‘Cross-Att’ in Fig. 1). PERCEIVER-VL encoder consists of k attention blocks, each of which is a stack of a cross-attention and l self-attentions over a latent array z , which results in the computational complexity of $\mathcal{O}(kMN + kln^2)$. In comparison, a standard transformer encoder with the same number of self-attention modules has a computational complexity of $\mathcal{O}(kLM^2)$. Since in vision-and-language tasks where the input size M is larger than the latent array size N , the change from quadratic to linear computational complexity w.r.t. M can greatly increase efficiency (see Fig. 4). To disambiguate the latent dimensions, we add a position embedding to the latent array z . We add the learned positional embedding [21, 74] for each latent dimension. The choice of learned position encoding is based on simplicity; different from the findings from the single-modality experiments of [28], we did not find the gain from using Fourier feature position encodings [66, 54, 28], as shown in the appendix.

3.3. LayerDrop on Cross-Attention for Reducing Depth on Demand

It is the cross-attention layers that take the highest computation in the attention blocks. Therefore, to further improve the efficiency of PERCEIVER-VL, we apply LayerDrop [17] to cross-attention layers, which allows users to control the latency by changing the number of cross-

attention layers during inference. Concretely, we apply dropout [17] to each cross-attention layer with probability p^{LD} during pretraining (see ‘LayerDrop’ in Fig. 1). Note that we do not apply LayerDrop to the first cross-attention layer, to ensure that the model always receives the signal from input. We study the effect of different p^{LD} and the effect of varying the number of cross-attention layers during inference (see Sec. 5.2.4 for details).

3.4. Structured Decoding with Cross-Attention and Query Array

To adapt PERCEIVER-VL to different vision-and-language tasks with structured output space, we give a query array q of arbitrary length Q (see ‘Query Array’ in Fig. 1), to decoder cross-attention and apply a task-specific head (a fully-connected layer) to the cross-attention output. We use a decoder with a single cross-attention [27]. For multi-task learning, we simply concatenate the query array for different tasks. In the following, we describe decoder queries for two vision-and-language pretraining objectives. See appendix for the query consturctions for downstream tasks.

3.4.1 Vision-and-Language Pretraining

We use two popular objectives in vision-and-language domain for PERCEIVER-VL pretraining: Vision-Text Matching (VTM) and Masked Language Modeling (MLM). To create the final query for the VTM and MLM tasks, we concatenate the queries for the two tasks, as illustrated in Fig. 1.

Vision-Text Matching (VTM) asks a model to distinguish whether a given pair of visual input (image or video) and text input matches or not. We create an unmatched pair by replacing its visual input with a randomly selected negative one, with 50% probability. We create the VTM query with a learnable embedding ($Q = 1$), illustrated as [CLS] in Fig. 1. We apply a linear VTM head to the corresponding decoder output and perform binary classification.

Masked Language Modeling (MLM) asks a model to infer masked text inputs in a given context. Following [14], we randomly mask 15% of the input text tokens. We create the MLM query by adding a positional embedding and a mask embedding ($Q = L^T$). The mask embedding is a learned embedding of a binary indicator variable, where 1 indicates the masked text token. Note that we do not feed the token embeddings to the decoder, i.e., we do not provide the text input. In doing so, we encourage the encoder output z_e to have a compact representation that contains enough information for MLM. We apply a linear MLM head to the corresponding decoder output and use cross-entropy loss.

3.5. Mixed-Stream Architecture for Cross-Modal Retrieval

In Fig. 2, we show two widely used architectures used in cross-modal retrieval tasks: (a) single-stream [14, 33] and (b) multi-stream [59, 4]. The single-stream architecture models the multi-modal similarity score s^{VL} with multiple layers of encoder, whereas the multi-stream encoder models the multi-modal similarity simply with a dot product between single-modality encodings z_e^V, z_e^L from separate encoders. In many real-world applications, multi-stream architectures are widely used for retrieval for their high efficiency. This is because multi-stream architectures allow us to cache pre-computed visual encodings z_e^V and simply compute dot products with text query z_e^L during inference. In contrast, single-stream architectures tend to achieve higher accuracy but require expensive computation, where a joint input array goes through multiple encoder layers. We propose to use a ‘mixed-stream’ architecture (Fig. 2 (c)) that takes the best of both worlds. Note that a similar idea has been proposed for text retrieval [26]. As shown in Fig. 6, our mixed-stream architecture achieves a good accuracy-latency tradeoff.

4. Experiment Setup

We pretrain PERCEIVER-VL on a combination of video-text and image-text datasets, then finetune it on a set of downstream benchmarks for evaluation. Below, we explain the details of our training and evaluation setup.

4.1. Architecture Details

Model Details. For the PERCEIVER-VL encoder, we use $k = 3$ blocks of 1 cross-attention and $l = 3$ self-attentions, totaling 3 cross-attention layers and 12 self-attention layers. The decoder has 1 cross-attention layer. We follow BERT_{BASE} [14] and ViT-B/32 [16] to use a hidden size of 768 and 12 attention heads. We follow ViT-B/32 [16] to use image (and video frame) size 384 and patch size 32. We use PyTorch [56] to implement our model in experiments.

LayerDrop on Cross-Attention. We set a probability of $p^{LD} = 0.5$ to apply dropout to the cross-attention layers during vision-and-language pretraining. Note that we do not apply dropout to the first cross-attention, to ensure that input signal always goes into the latent array. We analyze the effect of using LayerDrop during pretraining, finetuning, and inference, as shown in Table 3.

Modality Aggregation. By default, we map the multi-modal inputs to the latent space by creating an input array based on the concatenation of visual and textual inputs (namely *Joint* encoding). We also explore two other ways of combining the two modalities: encoding each modality serially with separate cross-attentions, then applying self-attentions (*Separate* encoding); encoding each modality serially with separate cross-attentions with self-attentions between them (*Separate+* encoding). We illustrate these aggregation strategies in Fig. 3. In our ablation study, we found that the three methods perform comparably, where the *Joint* encoding has the least computation. Therefore, we adopt the *Joint* encoding as our default modality aggregation method. See appendix for the detailed experiments.

4.2. Weight Initialization from Vision Transformers

To compare with recent methods that use pretrained visual backbone models, we experiment with initializing weights of PERCEIVER-VL with two popular models: ViT-B/32 [16] and CLIP (ViT-B/16) [59]. As these models have 12 self-attention layers, we insert 3 cross-attention layers after every 4 self-attention layers (before 1st/5th/9th).

Two-stage training. Since transformer models do not have cross-attention layers, the cross-attention weights could not *warm-start*. To stabilize training, after initializing the PERCEIVER-VL weights from CLIP parameters, we first train only the cross-attention layers, while freezing all other modules. After initial convergence (e.g., 1 epoch in MSRVTT [80]), we train the whole model jointly. In our experiment, this two-stage training strategy achieves better weight transfer than single-stage training (see appendix).

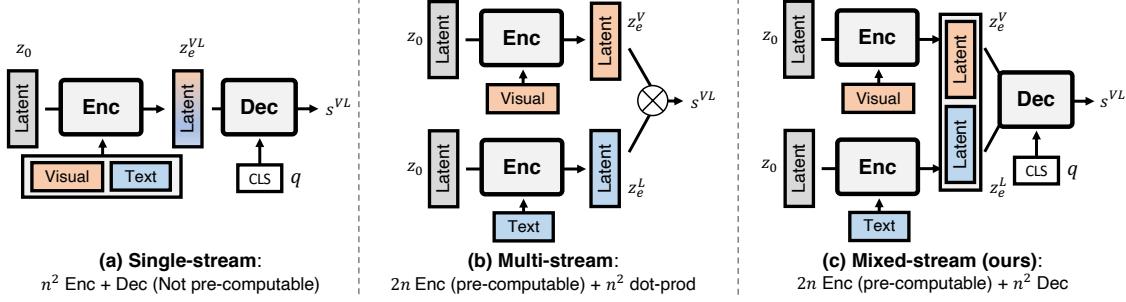


Figure 2. PERCEIVER-VL architectural variants for retrieval task (Sec. 3.5). (a) **single-stream**: encoder jointly processes concatenated multi-modal inputs, followed by decoder. (b) **multi-stream**: encoder separately processes single-modal inputs, followed by dot products. (c) **mixed-stream (ours)**: encoder separately processes single-modal inputs, followed by decoder processing the concatenation of encodings. We included the computation complexity of each architecture under its title. Note that most computation happens in the encoder (*Enc*). Because single-stream encoder does not allow pre-computation and also requires largest computation, it achieves the lowest efficiency during inference, while multi-stream and mixed-stream architectures have similar efficiency (see Sec. 5.2.3).

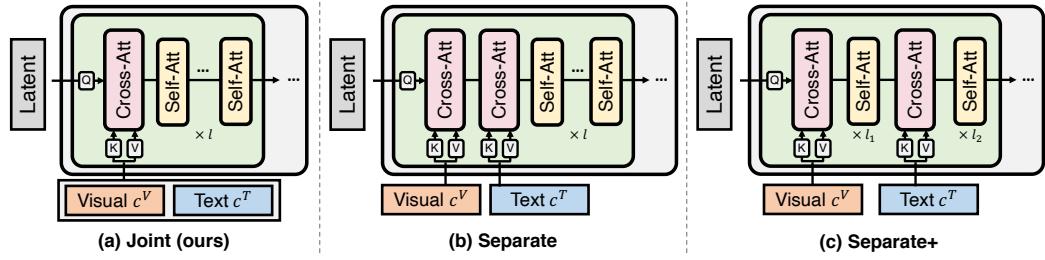


Figure 3. Different multi-modal encoding schemes. (a) **Joint**: Embeddings of each modality are concatenated and jointly encoded with a single cross-attention to the latent space. (b) **Separate**: using separate cross-attention for mapping each modality to the latent space. (c) **Separate+**: using self-attention after each single-modal cross-attention.

4.3. Pretraining Setups

Video-Text and Image-Text Datasets. We follow Frozen-in-Time [4] to pretrain PERCEIVER-VL on both video-text and image-text data. For image-text dataset, we use Conceptual Captions (CC) [64], which consists of 3M image-text pairs. For video-text dataset, we use Webvid [4], which consists of 2.5M video-text pairs, with an average video length of 18 seconds.

Training Details. We use Adam optimizer [34] with a learning rate 1e-5 and weight decay 0.001. We set 200k training steps for joint Video-Text & Image-Text pretraining. We use the batch size of 4096 with gradient accumulation on 4 RTX 2080Ti GPUs for 14 days.

4.4. Downstream Tasks

After pretraining, we evaluate PERCEIVER-VL on various vision-and-language benchmarks, covering cross-modal retrieval and visual question answering for both video-text and image-text datasets.

4.4.1 Video-Text Tasks

For video retrieval, we use MSRVTT [80], LSMDC [62], DiDeMo [2], ActivityNet Captions [36]. For video question answering, we use TGIF-QA [29] and MSRVTT-QA.

Dataset Details. MSRVTT contains 10K web video clips, with 20 captions for each clip. LSMDC contains 118,081 short clips from 202 movies with each clip containing one caption. DiDeMo contains 10,464 videos with 40,543 temporally localized sentences. ActivityNet Captions have 20k videos with 3.65 temporally localized sentences per video on average, resulting in 100k sentences in total. Videos have an average duration of 180 seconds. For DiDeMo and ActivityNet Captions, we follow previous work [50, 41] to use paragraph-to-video retrieval, where we concatenate all sentences from the same video as a single text query for retrieval. TGIF-QA contains 165K QA pairs from 72K animated GIF videos. We follow [41] to evaluate our model on three TGIF-QA tasks: action, transition, and frame. MSRVTT-QA contains 10K videos with 243K open-ended questions collected from MSRVTT videos.

Training Details. We use Adam optimizer [34] with a learning rate 1e-5 and weight decay 0.001. We use 16

frames for ActivityNet Captions and 8 frames for other tasks. We use frame size 384×384 , and maximum text length 40 for all tasks.

4.4.2 Image-Text Tasks

For image retrieval, we use Flickr30k [57]. For visual question answering, we use VQAv2 [3] and NLVR² [67].

Dataset Details. VQAv2 contains 204,721 images from COCO [49], with a minimum of 3 questions per image and 10 grounded answers. NLVR² contains 107,292 examples of sentences grounded in image pairs. Flickr30k dataset has 31,000 images collected from Flickr each with 5 sentences.

Training Details. We use Adam optimizer with a learning rate of 1e-4 and weight decay of 0.001. We use image size 384×384 and maximum text length 40 for all tasks.

5. Results and Analysis

We first compare PERCEIVER-VL with recent methods in video-text / image-text benchmarks, where it achieves the highest efficiency, while maintaining competitive performance (Sec. 5.1). Then we analyze the efficiency of PERCEIVER-VL in detail (Sec. 5.2). In appendix, we also present ablation studies of different architectural components and training strategies for PERCEIVER-VL.

5.1. Comparison to State-of-the-Art

In Table 1, we compare PERCEIVER-VL with the state-of-the-art video-text models on 4 text-to-video retrieval (MSRVTT, DiDeMo, LSMDC, ActivityNet) and 2 video question answering (TGIF, MSRVTT-QA) benchmarks. In Table 2, we compare PERCEIVER-VL with state-of-the-art image-text models on text-to-image retrieval (Flickr30k) and 2 visual question answering (VQAv2, NLVR²) benchmarks. The closest baseline of our model is Frozen-in-Time [4], as it is pretrained on the same pretraining dataset (CC/Webvid) and handles both images and videos in a single architecture. PERCEIVER-VL achieves competitive performance across the board, for both image-based and video-based tasks, while maintaining significantly higher efficiency. PERCEIVER-VL has the lowest GFLOPs and inference time (see the rightmost columns of the tables).

As some recent video retrieval models adopt CLIP [59] trained on 400M image-text pairs from the Web, we also provide experiments with a CLIP variant by inserting randomly initialized cross-attentions inside the CLIP visual encoder. The use of CLIP significantly improves the retrieval performance (e.g., 32.6→45.9 on MSRVTT R@1). There is a certain gap between our model and the baselines, because CLIP self-attention layers are trained to handle image

patches, rather than compact latent spaces. Thus, we gray out the CLIP-based results in Table 1 to highlight the fact that our models are not directly comparable to transformer-based models. We expect that a better weight initialization (e.g., from a PERCEIVER architecture trained on 400M image-text pairs) would further improve the performance of our models.

5.2. Efficiency Analysis

5.2.1 Scaling Input Array

In Fig. 4, we compare the computations of PERCEIVER-VL, ViLT-B/32 [33], and Frozen-in-Time [4] for video inputs of different scales, by varying the number of frames (*left*) and the frame size (*right*). All three models have 12 self-attention layers with hidden size 768. Powered by efficient cross-attention-based encoding, PERCEIVER-VL shows a remarkably better scalability (lower GFLOPs) than ViLT-B/32 and Frozen-in-Time in both plots.

5.2.2 Scaling Latent Array

We study the effect of varying the latent array size N to explore whether we can further improve the efficiency of PERCEIVER-VL. In Fig. 5, we show the effect of varying the latent array sizes during finetuning in terms of computation and downstream performance on MSRVTT. We use $N=128$ during pretraining. When scaling up or down the latent array for a pretrained model, we simply initialize a new latent array where we empirically find that it gives similar performance compared to interpolating the pretrained latent array. We can see that the GFLOPs scales linearly with N , while the retrieval performance remains reasonably well in three different pretraining setups (e.g., CC+Webvid PT: 24.0 → 24.6 → 26.8 → 27.1 with latent array size: 32 → 64 → 128 → 256).

5.2.3 Mixed-Stream Architecture for Retrieval

In Fig. 6 we compare different retrieval architecture variants discussed in Sec. 3.5 and Fig. 2, in terms of accuracy and inference time on MSRVTT val split. The single-stream architecture achieves the highest R@1 (27.2), but also takes the longest inference time. The multi-stream architecture achieves the lowest R@1 (26.0), with the shortest inference time. Our mixed-stream architecture achieves a good accuracy-latency tradeoff, with R@1 (26.8) close to single-stream architecture, while running significantly faster.

5.2.4 LayerDrop to Encoder Cross-Attentions

In Table 3 we analyze the effect of applying LayerDrop (LD) [17] to cross-attention layers in encoder, as discussed in Sec. 3.3 on MSRVTT retrieval. We use the mixed-stream architecture as the default setting. First, we observe that LD

Model	Pretraining Datasets	Visual Backbone	Retrieval \uparrow				QA Acc. \uparrow		GFLOPs \downarrow	Time (ms) \downarrow
			MSR	DDM	LSM	ACT	TGIF	MSR		
Models using other input modalities (<i>e.g.</i> , audio)										
HERO [45]	TV/HT100M	ResNet152+Slowfast [22, 18]	20.5	-	-	-	-	-	935.2	2200.0
MMT [19]	HT100M	S3D+VGG+DenseNet161 [78, 23, 24]	26.6	-	12.9	-	-	-	-	-
AVLNET [63]	HT100M	ResNet152+ResNeXt [22, 77]	27.1	-	17.0	-	-	-	153.4	2000.0
Models with CLIP initialization										
Hunyuan(+DSL) [55]	-	CLIP (ViT-B/16)	55.0	52.1	29.7	57.3	-	-	2022.8	-
CLIP2TV [20]	-	CLIP (ViT-B/16)	49.3	45.5	-	44.1	-	-	2212.3	-
DRL [75]	-	CLIP (ViT-B/32)	48.8	49.0	26.5	46.2	-	-	511.0	320.0
CAMoE(+DSL) [10]	-	CLIP (ViT-B/32)	47.3	-	25.9	-	-	-	399.7	-
MDMMT-2 [37]	-	CLIP (ViT-B/32)	48.5	-	26.9	-	-	-	-	-
Ours ^{N=32, Multi}	CC/Webvid	CLIP (ViT-B/16)	45.9	-	-	-	-	-	80.0	80.0
HT100M [53]	HT100M	ResNet152+ResNeXt [22, 77]	14.9	-	7.1	-	-	-	164.3	1100.0
ClipBERT [41]	COCO/CC	ResNet50 [31]	22.0	20.4	-	21.3	60.3	37.4	340.0	700.0
Frozen-in-Time [4]	CC/Webvid	Timesformer-B/16 [6]	31.0	31.0	15.0	-	-	-	89.0	260.0
Ours ^{N=128, Mixed}	CC/Webvid	ViT-B/32 [16]	32.6	30.5	15.8	33.9	69.2	43.2	43.9	72.0

Table 1. Finetuning performance on text-to-video retrieval and video question answering benchmarks. We report R@1 for text-to-video retrieval tasks (see appendix for R@5/R@10) and report QA accuracy on the FrameQA task. *GFLOPs* shows the inference cost on a single sample, and *Time (ms)* indicates the average inference time across all samples on MSRVTT val split. For a fair comparison, we gray out 1) the models that use input modalities other than video and text (*e.g.*, audio) and 2) the models that use CLIP visual encoder [59] (the cross-attention layers of PERCEIVER-VL cannot be initialized with CLIP parameters and trained from scratch; see the discussion in Sec. 5.1). *MSR*=MSRVTT, *DDM*=DiDeMo, *LSM*=LSMDC, *ACT*=ActivityNet, *TGIF*=TGIF-QA. ^{N=128} means latent size N=128. ^{Multi} and ^{Mixed} mean multi-stream and mixed-stream respectively.

Model	Pretraining Datasets	Visual Backbone	Retrieval \uparrow		QA Accuracy \uparrow		GFLOPs \downarrow	Time (ms) \downarrow
			Flickr30k	VQAv2	NLVR ²	NLVR ²		
Models using additional object tag inputs								
VinVL-Base [85]	COCO/CC/SBU/Flickr/OI*	Faster-RCNN [85]	-	75.95	83.08	949.9	1023.3	800.0
OSCAR-Base [48]	COCO/CC/SBU/Flickr*	Faster-RCNN [1]	-	73.16	78.36	956.4	1000.0	
UNITER-Base [9]	COCO/CC/SBU/VG	Faster-RCNN [1]	72.5	72.70	75.80	949.9	1000.0	
ViLT-B/32 [33]	COCO/CC/SBU/VG	ViT-B/32 [16]	64.4	71.26	76.13	55.9	32.0	
Ours ^{N=128}	COCO/CC/SBU/VG	ViT-B/32 [16]	62.4	71.62	75.53	30.5	18.0	
LXMERT [69]	COCO/VG*	Faster-RCNN [1]	-	72.42	74.50	952.0	1100.0	
VisualBERT [46]	COCO	Faster-RCNN [1]	-	70.80	67.00	425.0	1000.0	
Pixel-BERT-R50 [25]	COCO/VG	ResNet50 [22]	53.4	71.35	72.40	136.8	150.0	
Ours ^{N=128}	COCO/VG	ViT-B/32 [16]	61.7	70.45	74.87	30.5	18.0	
Frozen-in-Time [4]	CC/Webvid	Timesformer-B/16 [6]	61.0	-	-	63.9	70.0	
Ours ^{N=64}	CC/Webvid	ViT-B/32 [16]	61.0	70.12	74.52	17.0	8.0	
Ours ^{N=128}	CC/Webvid	ViT-B/32 [16]	61.8	70.91	75.44	30.5	18.0	

Table 2. Finetuning performance on text-to-image retrieval and visual question answering benchmarks. For NLVR², we show Test-P accuracy. For Flickr30k, we show text-to-image retrieval R@1 (see appendix for R@5/R@10). Note that for brevity, we only show the image or video source datasets for *Pretraining Datasets*; the datasets that added additional text annotations are not included in the column (we use * to highlight them). For example, LXMERT is trained with image-text datasets COCO and VG, as well as the three QA datasets based on COCO and VG images, *i.e.*, VQAv2, VGQA and GQA. We also gray out models that use additional object tags in the first block and are not comparable to our model. *GFLOPs* shows the inference cost on a single sample, *Time (ms)* indicates the average inference time over all samples in VQAv2 minival split; For a fair comparison, we gray out models that are pretrained with more data. ^{N=128} means latent size N=128.

acts as a regularizer, as we see LD improves the MSRVTT accuracy in the first block, while increasing p^{LD} too high 0.5 \rightarrow 0.7 does not help the performance (28.8 \rightarrow 26.6). The last row in the bottom block achieves the best accuracy (27.1), with LD during both pretraining and finetuning. Second, removing cross-attention layers without LD during finetuning hurts performance (see 26.1 \rightarrow 24.0 in the mid-

dle block). Lastly, with LD during finetuning, the latency of the inference time can be reduced by 19.4% (72.0 ms \rightarrow 58.0 ms), with minimal accuracy drop (see 27.1 \rightarrow 26.3 in the bottom block). This indicates that, with a LD-finetuned model, we can control its latency on demand at the inference time by varying the number of cross-attention layers, without storing checkpoints of multiple models.

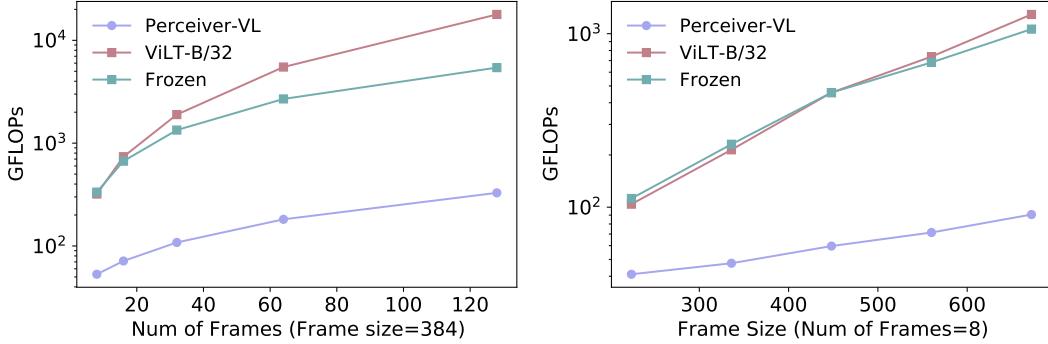


Figure 4. Input scaling comparison between PERCEIVER-VL, ViLT-B/32 [33], and Frozen-in-Time [4] for video inputs with different the number of frames (*left*) and frame size (*right*). Note that the GFLOPs are illustrated in log-scale. On both plots, PERCEIVER-VL shows remarkably lower computation compared to the other two models.

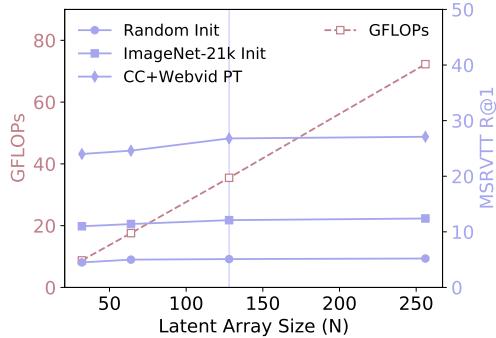


Figure 5. The efficiency-accuracy tradeoff of using different latent array size N during finetuning on MSRVTT. During pretraining, we use the latent array size $N = 128$ (blue vertical line). We use mixed-stream architecture by default.

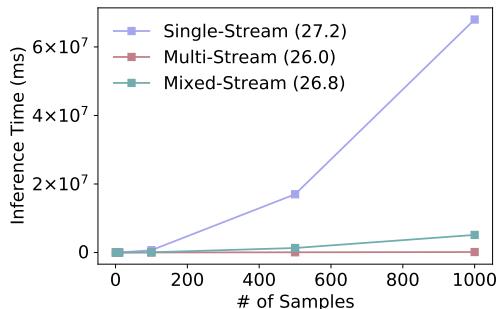


Figure 6. Comparison of retrieval architectures with different number of MSRVTT video-text pairs. The inference time is calculated from the total retrieval time.

6. Conclusion

In this work, we present PERCEIVER-VL, a vision-and-language framework that efficiently handles high-dimensional multi-modal inputs such as long videos and text. The efficiency of PERCEIVER-VL comes from linear complexity based on iterative cross-attention, LayerDrop on cross-attention layers, and a mixed-stream architecture for cross-modal retrieval. Experiments on diverse vision-

Pretraining	Finetuning	Inference	# Cross-attentions in encoder	MSRVTT R@1	Time (ms)
			3		
1 ~ 3 (0.5)	3	3	26.4	72.0	72.0
1 ~ 3 (0.7)	3	3	26.8	72.0	72.0
1 ~ 3 (0.5)	1	1	26.6	72.0	72.0
1 ~ 3 (0.5)	3	3	26.1	58.0	72.0
1 ~ 3 (0.5)	3	3	26.8	72.0	72.0
1 ~ 3 (0.5)	3	1	24.0	58.0	72.0
1 ~ 3 (0.5)	1 ~ 3 (0.5)	1	26.3	58.0	72.0
1 ~ 3 (0.5)	1 ~ 3 (0.5)	3	27.1	72.0	72.0

Table 3. Accuracy and inference time on MSRVTT retrieval with varied number of cross-attentions in PERCEIVER-VL mixed-stream encoder. We include the layer dropout probability p^{LD} in brackets if used. Note that PERCEIVER-VL has 3 cross-attention layers in the encoder, and we do not apply dropout to the first cross-attention in the encoder ($p^{LD} = 0$) to ensure that the latent array always receives signals from the input.

and-language benchmarks show that our framework has a remarkably higher efficiency than state-of-the-art models, while achieving competitive or better performance. Moreover, we comprehensively analyze the efficiency of our framework, including measuring the scalability in terms of input and latent array size, reducing latency by dropping cross-attention layers, comparing architecture variants, and an ablation study on model training details. It would be an interesting future work to further explore efficient vision-and-language modeling with even more diverse tasks.

Acknowledgments

We thank the reviewers for their helpful comments. This work was supported by ARO Award W911NF2110220, DARPA KAIROS Grant FA8750-19-2-1004, ONR Grant N000141812871, and NSF-AI Engage Institute DRL-211263. The views, opinions, and/or findings contained in this article are those of the authors and not of the funding agency.

References

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

[2] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *ICCV*, pages 5803–5812, 2017.

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

[4] Max Bain, Arsha Nagrani, GÜl Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. *arXiv preprint arXiv:2104.00650*, 2021.

[5] Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

[6] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? *arXiv preprint arXiv:2102.05095*, 2021.

[7] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, pages 961–970, 2015.

[8] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

[9] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Learning universal image-text representations. In *ECCV*, 2020.

[10] Xing Cheng, Hezheng Lin, Xiangyu Wu, Fan Yang, and Dong Shen. Improving video-text retrieval by multi-stream corpus alignment and dual softmax loss. *arXiv preprint arXiv:2109.04290*, 2021.

[11] Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*, 2019.

[12] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In *ICML*, 2021.

[13] Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. Electra: Pre-training text encoders as discriminators rather than generators. In *ICLR*, 2020.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2018.

[15] Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. Unified language model pre-training for natural language understanding and generation. In *NeurIPS*, 2019.

[16] 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. In *ICLR*, 2021.

[17] Angela Fan, Edouard Grave, and Armand Joulin. Reducing transformer depth on demand with structured dropout. *arXiv preprint arXiv:1909.11556*, 2019.

[18] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, pages 6202–6211, 2019.

[19] Valentin Gabeur, Chen Sun, Kartek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *ECCV*, pages 214–229. Springer, 2020.

[20] Zijian Gao, Jingyu Liu, Sheng Chen, Dedan Chang, Hao Zhang, and Jinwei Yuan. Clip2tv: An empirical study on transformer-based methods for video-text retrieval. *arXiv preprint arXiv:2111.05610*, 2021.

[21] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. Convolutional sequence to sequence learning. In *ICML*, pages 1243–1252. PMLR, 2017.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[23] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *ICASSP*, pages 131–135. IEEE, 2017.

[24] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *ICCV*, pages 4700–4708, 2017.

[25] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*, 2020.

[26] Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring. *arXiv preprint arXiv:1905.01969*, 2019.

[27] Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Kopputla, Daniel Zoran, Andrew Brock, Evan Shelhamer, et al. Perceiver io: A general architecture for structured inputs & outputs. *arXiv preprint arXiv:2107.14795*, 2021.

[28] Andrew Jaegle, Felix Gimeno, Andrew Brock, Andrew Zisserman, Oriol Vinyals, and Joao Carreira. Perceiver: General perception with iterative attention. *arXiv preprint arXiv:2103.03206*, 2021.

[29] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *CVPR*, pages 2758–2766, 2017.

[30] Chao Jia, Yafei 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. *arXiv preprint arXiv:2102.05918*, 2021.

[31] Huaizu Jiang, Ishan Misra, Marcus Rohrbach, Erik Learned-Miller, and Xinlei Chen. In Defense of Grid Features for Visual Question Answering. In *CVPR*, 2020.

[32] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and Francois Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International Conference on Machine Learning*, pages 5156–5165. PMLR, 2020.

[33] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. *arXiv preprint arXiv:2102.03334*, 2021.

[34] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2014.

[35] Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. *arXiv preprint arXiv:2001.04451*, 2020.

[36] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *ICCV*, 2017.

[37] Alexander Kunitsyn, Maksim Kalashnikov, Maksim Dzabraev, and Andrei Ivaniuta. Mdmm2-2: Multidomain multi-modal transformer for video retrieval, one more step towards generalization. *arXiv preprint arXiv:2203.07086*, 2022.

[38] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. In *ICLR*, 2020.

[39] Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Sungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International Conference on Machine Learning*, pages 3744–3753. PMLR, 2019.

[40] Jie Lei, Tamara L Berg, and Mohit Bansal. Qvhighlights: Detecting moments and highlights in videos via natural language queries. In *NeurIPS*, 2021.

[41] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In *CVPR*, 2021.

[42] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. *arXiv preprint arXiv:1809.01696*, 2018.

[43] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. Tvr: A large-scale dataset for video-subtitle moment retrieval. In *ECCV*, 2020.

[44] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, Dixin Jiang, and Ming Zhou. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *AAAI*, pages 11336–11344, 2020.

[45] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. Hero: Hierarchical encoder for video+language omni-representation pre-training. In *EMNLP*, 2020.

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

[47] Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. *arXiv preprint arXiv:2012.15409*, 2020.

[48] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*, 2020.

[49] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014.

[50] Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. Use what you have: Video retrieval using representations from collaborative experts. In *BMVC*, 2020.

[51] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

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

[53] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *ICCV*, pages 2630–2640, 2019.

[54] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, pages 405–421. Springer, 2020.

[55] Shaobo Min, Weijie Kong, Rong-Cheng Tu, Dihong Gong, Chengfei Cai, Wenzhe Zhao, Chenyang Liu, Sixiao Zheng, Hongfa Wang, Zhifeng Li, et al. Hunyuan_tvr for text-video retrieval. *arXiv preprint arXiv:2204.03382*, 2022.

[56] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett, editors, *NeurIPS*, pages 8024–8035. Curran Associates, Inc., 2019.

[57] 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*, pages 2641–2649, 2015.

[58] Jiezhong Qiu, Hao Ma, Omer Levy, Scott Wen-tau Yih, Sinong Wang, and Jie Tang. Blockwise self-attention for long document understanding. *arXiv preprint arXiv:1911.02972*, 2019.

[59] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learn-

ing transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

[60] Jack W Rae, Anna Potapenko, Siddhant M Jayakumar, and Timothy P Lillicrap. Compressive transformers for long-range sequence modelling. *arXiv preprint arXiv:1911.05507*, 2019.

[61] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 2020.

[62] Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. A dataset for movie description. In *CVPR*, pages 3202–3212, 2015.

[63] Andrew Rouditchenko, Angie Boggust, David Harwath, Brian Chen, Dhiraj Joshi, Samuel Thomas, Kartik Audhkhasi, Hilde Kuehne, Rameswar Panda, Rogerio Feris, et al. Avlnet: Learning audio-visual language representations from instructional videos. *arXiv preprint arXiv:2006.09199*, 2020.

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

[65] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mass: Masked sequence to sequence pre-training for language generation. In *ICML*, 2019.

[66] Kenneth O Stanley. Compositional pattern producing networks: A novel abstraction of development. *Genetic programming and evolvable machines*, 8(2):131–162, 2007.

[67] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. In *ACL*, 2019.

[68] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *ICCV*, 2019.

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

[70] Hao Tan and Mohit Bansal. Vokenization: improving language understanding with contextualized, visual-grounded supervision. In *EMNLP*, 2020.

[71] Zineng Tang, Jaemin Cho, Hao Tan, and Mohit Bansal. Vid-lankd: Improving language understanding via video-distilled knowledge transfer. *NeurIPS*, 34, 2021.

[72] Zineng Tang, Jie Lei, and Mohit Bansal. Decembert: Learning from noisy instructional videos via dense captions and entropy minimization. In *NAACL-HLT*, pages 2415–2426, 2021.

[73] Yi Tay, Dara Bahri, Liu Yang, Donald Metzler, and Da-Cheng Juan. Sparse sinkhorn attention. In *International Conference on Machine Learning*, pages 9438–9447. PMLR, 2020.

[74] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

[75] Qiang Wang, Yanhao Zhang, Yun Zheng, Pan Pan, and Xian-Sheng Hua. Disentangled representation learning for text-video retrieval. *arXiv preprint arXiv:2203.07111*, 2022.

[76] Sinong Wang, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*, 2020.

[77] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *CVPR*, 2017.

[78] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *ECCV*, pages 305–321, 2018.

[79] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueling Zhuang. Video question answering via gradually refined attention over appearance and motion. In *ACM MM*, 2017.

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

[81] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. In *NeurIPS*, 2019.

[82] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueling Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *AAAI*, 2019.

[83] Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. Merlot reserve: Neural script knowledge through vision and language and sound. *arXiv preprint arXiv:2201.02639*, 2022.

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

[85] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *CVPR*, pages 5579–5588, 2021.

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

[87] Luowei Zhou, Chenliang Xu, and Jason J Corso. Towards automatic learning of procedures from web instructional videos. In *AAAI*, 2017.

[88] Linchao Zhu and Yi Yang. Actbert: Learning global-local video-text representations. In *CVPR*, 2020.