eP-ALM: Efficient Perceptual Augmentation of Language Models

Mustafa Shukor[1] Corentin Dancette[1] Matthieu Cord[1,2]

[1]Sorbonne University [2]Valeo.ai

{firstname.lastname}@sorbonne-universite.fr

Abstract

*Large Language Models (LLMs) have so far impressed the world, with unprecedented capabilities that emerge in models at large scales. On the vision side, transformer models (*i.e., *ViT) are following the same trend, achieving the best performance on challenging benchmarks. With the abundance of such unimodal models, a natural question arises; do we need also to follow this trend to tackle multimodal tasks? In this work, we propose to rather direct effort to efficient adaptations of existing models, and propose to augment Language Models with perception. Existing approaches for adapting pretrained models for vision-language tasks still rely on several key components that hinder their efficiency. In particular, they still train a large number of parameters, rely on large multimodal pretraining, use encoders (*e.g., *CLIP) trained on huge image-text datasets, and add significant inference overhead. In addition, most of these approaches have focused on Zero-Shot and In Context Learning, with little to no effort on direct finetuning. We investigate the minimal computational effort needed to adapt unimodal models for multimodal tasks and propose a new challenging setup, alongside different approaches, that efficiently adapts unimodal pretrained models. We show that by freezing more than 99% of total parameters, training only one linear projection layer, and prepending only one trainable token, our approach (dubbed eP-ALM) significantly outperforms other baselines on VQA and Captioning across Image, Video, and Audio modalities, following the proposed setup. The code is available here: https://github.com/mshukor/eP-ALM.*

1. Introduction

Going large scale has led to outstanding performances that consistently improve across tasks, modalities, and domains on current benchmarks. Most of the progress so far has been in the vision and language domains. For Computer Vision, the ViT family [20] starts from the tiny model with 5M parameters to the enormous ViT-e [14] with 4B parameters and the largest ViT-22B with 22B parameters [19]. More captivating, are the scales of Large Language Models (LLMs), such as the BLOOM [70] and OPT [102] families, ranging from hundreds of millions of parameters to 175B, in addition to other models that go beyond 100B [6, 77, 18] up to 1T parameters [26]. These huge scales come with a need for very large pretraining datasets and long training times.

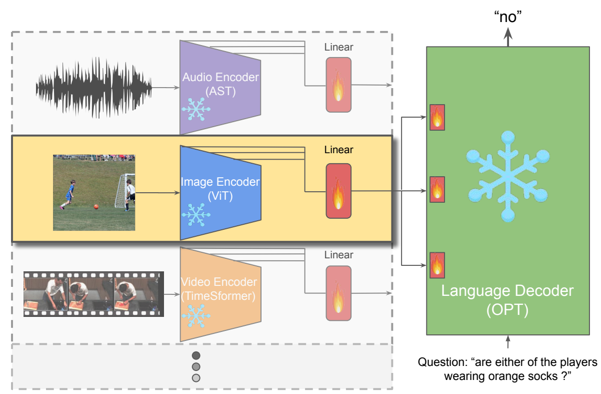

Figure 1: Illustration of eP-ALM to adapt unimodal models for multimodal tasks. The Language Model (Decoder) is augmented with perceptual context to steer its text generation. To condition the decoder on a given modality, the [CLS] tokens are extracted from several layers of a modality-specific encoder and then linearly projected before concatenation at different levels of the language decoder. Only unimodal models are used, and all pretrained modules are kept frozen.

The current prevalent paradigm to solve multimodal tasks, in particular, Vision-Language tasks is to leverage pretrained models, and then further train end-to-end [54, 73, 75, 89, 14] on large image-text datasets. However, the training cost is huge and unaffordable for much of the community, as these approaches always train all model parameters, even after initialization, on a huge amount of data.

With the abundance of unimodal models, a natural question arises;

Do we need also to follow this trend to tackle multimodal *tasks? or rather direct effort to efficient adaptations of existing models?*

Drawing inspiration from the recent work in Augmented Language Models (ALMs) [65], in this paper, we advocate for adapting pretrained LMs to solve multimodal tasks. Specifically, by augmenting LMs with perceptual encoders.

Several approaches have deviated from the end-to-end-training paradigm by freezing some pretrained modules and training only the adaptation parameters, such as, additional cross-attention [2], vision encoder [85] and Adapters [22].

Even though these approaches have taken a big step towards more parameter-efficient models, there are still many costly components that hinder their adoption by the large community, such as the training and inference memory and time cost.

In this work we argue that current approaches are far from optimal and it is possible to find more efficient approaches, in terms of the number of trainable parameters, training data, and compute, to adapt pretrained unimodal models for multimodal tasks. A better alignment of visual and language representations might help to devise extremely efficient adaptation approaches.

To investigate this hypothesis, we go a step further to efficiently leverage LLMs, and propose (1) a new technique to adapt unimodal models by freezing more than 99 % (up to 99.94%) of their parameters, alongside (2) a minimal and challenging setup to adapt pretrained unimodal models for Image/Video/Audio-Language tasks (*e.g.*, VQA [32, 92], Image and Audio Captioning [13, 47]). In this setup, we favor unimodal-only models, avoiding multimodal pretraining or massively trained multimodal encoders, and considering the typical LLMs architecture as the backbone. All that while freezing as much as possible of model parameters. The approach is illustrated in Fig.1.

Specifically, we adopt the publicly released OPT model [102] and unimodal encoders (*e.g.*, ViT, TimeSformer [5], AST [30]), which are kept frozen. We finetune directly the adaptation parameters on publicly available benchmarks of downstream tasks such as for VQA, GQA, Image Captioning, Video QA, Video Captioning, and Audio Captioning.

Based on this setup we investigate different design choices and propose very efficient approaches backed by the following interesting findings:

- Training a single linear layer directly on downstream multimodal datasets, and following the same setup, outperforms other work on Image/Video/Audio-Language tasks. With a few additional trainable parameters and a single learned prepended token, we can significantly improve the performance, while respecting a budget of 1% of trainable parameters, and keeping almost the same inference cost.
- Our approach enjoys better generalization (OOD, Zero-Shot) and is data-efficient (training on 1% of the data achieves 80% of performances) with better few-shot results than other approaches.
- While reaching good performance with small to mid-scale language models (*i.e*, 350M-2.7B) the improvement still increases by jointly scaling both vision and language models. When scaling both models, we can still outperform other approaches with only 0.06% of trainable parameters.
- Existing approaches do not behave well on the proposed challenging setup, without large multi-modal pretraining.

2. Related Work

Vision-Language Models (VLMs). Previously, vision-language tasks have been solved with models heavily customized for the particular task at hand [46, 25, 7, 43, 40]. The success in Self Supervised Learning [91, 33, 35, 81, 8] and the importance of good initialization have pushed researchers to transfer these ideas to VLMs and started Vision-Language Pretraining (VLP) on large scale video-text [27, 55, 87], image-text datasets in general domains [60, 15, 73, 54, 50, 53], as well as specific domains, such as Cooking [74], Medical Images [66] and Event Extraction [56]. VLP is a step to move away from the burden of customization by having one pretrained model, exploited for several downstream tasks. Recently, we have witnessed impressive work that go a step further towards more unification, by unifying the model, the training objective, and input-output format [88, 89, 14, 61]. All these models train most of the model parameters, even after initialization, which becomes more and more costly with the current trend in scaling data, model size, and compute [98, 14]. Another approach for VLM is to exploit existing pretrained models by keeping them frozen and training only the adaptation parameters [2, 22, 57]. This work advocates for the latter favoring training efficiency in terms of memory and time.

Adapting Language Models. Large Language Models (LLMs) [6, 77, 37, 18, 102, 70] have impressed the world in this last few years, showing unprecedented performance on a myriad of NLP tasks. Scaling LLMs to hundreds of billions of parameters has been motivated by the capabilities that surprisingly emerge [90] at this scale and lead to sudden jumps of relevant metrics on hard downstream tasks [78, 68, 36]. This generalization ability pushed researchers to start adapting these models for other modalities [85, 2], tasks [101, 82, 97, 34] and domains [76]. Currently, most of the focus is concentrated on exploiting LLMs for vision-language tasks, such as Flamingo [2] which trains 10B parameters to adapt a frozen 70B parameter language model, and other successful efficient techniques that are based on vision-conditioned prompt tuning (Frozen [85], PromptFuse [57], LiMBeR [64]) and adapters (MAGMA [22]). This work has demonstrated good performance, showing that it is possible to devise very efficient approaches to adapt existing language models [86, 37]. On the video side, little work has been proposed, mostly based on Adapters [96, 72]. The closest to our approach is PromptFuse [57] which finetunes directly for VQA; however, they use encoder-decoder language models and train a soft prompt that is prepended to the input.

Efficient Learning. Parameter-Efficient learning is an interesting line of research that consists of adapting pretrained models using very few trainable parameters. Prompt Tuning [52] is one such approach that appends a few learnable tokens, or Soft Prompts to contextualize the input and steer the output of the frozen model toward the desired task. Other approaches use Adapters [38, 3], which are trainable MLP, consisting of 2 linear projection layers with activation in between and inserted inside the model to adapt the self-attention and feedforward layers. Many other approaches have been proposed in the context of NLP such as LoRa [39], Bitfit [99], Hyperformer [62], Compacters [45] and $(IA)^3$ [58]. These approaches have been successfully adapted to other modalities such as image [42, 12], image-text [80, 79, 103], with very little work on video [67] and Audio [49].

Another line of research is Data-Efficient techniques, where the objective is to attain similar performance by significantly reducing the training datasets. Recently, some efforts have been proposed for vision [83, 84], language [21] and vision-language [73, 16, 11], which mostly focus on designing better training objectives [73]. However, little work has been done to investigate the connection between parameter efficiency and data efficiency, which is considered in this work.

3. Framework

To solve multimodal tasks, we propose to augment pretrained LLMs with perception through unimodal perceptual encoders (Fig.1). We detail our approach in the following.

3.1. eP-ALM

We augment a pretrained LM with perception through several modality-specific encoders. The encoders interact with LM through linearly projected, modality-specific [CLS] tokens. To ease the adaptation, we leverage some parameter-efficient techniques, such as Prompt Tuning. In this section, we detail the design principles of our approach, which is illustrated in Fig.2.

Language Model (LM) We adopt OPT models [102], which are autoregressive language decoders consisting of

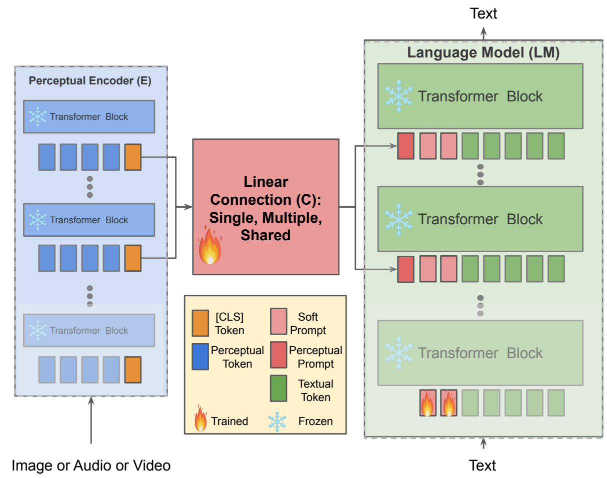

Figure 2: Illustration of the adaptation mechanism in eP-ALM. The perceptual input (image/video/audio) is fed to the perceptual encoder **E** (*e.g.*, ViT) and the corresponding text to the LM (*e.g.*, OPT), which then generates a text conditioned on the perceptual input. The multimodal interaction is done via the $[CLS]$ tokens acting as Perceptual Prompt, and are extracted from the last layers of the encoder, then injected in the last layers of LM, after passing by the Linear Connection C. The previous $[CLS]$ token is replaced by the new one coming from a deeper layer, keeping the number of tokens fixed. The first layers (grayed) of each model are kept intact without any modality interaction. We ease the adaptation with a Soft Prompt that is prepended to the input of LM.

Self-Attention and Feed Forward layers. They are trained with *next token prediction* objective on 180B tokens mostly in English and gathered from different datasets [28, 4]. The authors released a family of models with different scales, starting from 125M up to 175B model size. Besides being open source and trained on English data, the different model sizes allow us to readily investigate the effect of scale, and help to devise new approaches with affordable model sizes.

Perceptual Encoders We favor only unimodal models. For images, we use the vanilla ViT model [20] which consists of Self Attention and FeedForward layers and is pretrained for image classification on ImageNet [69]. For Video, we use TimeSformer [5] that consists of a ViT-like model augmented with temporal attention and pretrained on kinetics [9]. For Audio, we adopt AST [30], a vanilla adaptation of ViT to digest spectrograms, that is pretrained on AudioSet [29]. Even though we consider only these 3 encoders, the extension of the approach to other types of encoders and modalities is straightforward.

Perceptual Prompt Injection. LMs are usually controlled via different textual prompts, such as questions and instructions. Here, the LM is controlled by both the text and the perceptual encoders. Specifically, the projected perceptual tokens are prepended to the textual tokens. Naively using all visual tokens, adds significant computation costs during training and inference, due to the quadratic complexity of attention layers with the number of tokens. This becomes more apparent with LLMs. To mitigate this, we consider only the [CLS] token of the perceptual encoders and prepend it to the text tokens. This increases the total number of tokens by 1 which maintains almost the same inference speed.

Connecting Models with Cross-Modal Hierarchical Linear layers. When freezing the perceptual encoders and language models, the minimal number of trainable parameters are those that amount to connecting these two models while adjusting the embedding dimensions in case of a mismatch. Therefore, we base our approach on this constraint and train only one linear projection layer (*single connection*, Fig.2) to connect both models. To exploit the hierarchical representation encoded in pretrained models, instead of taking only the [CLS] token of the last output layer, we take the [CLS] tokens from several layers of the perceptual model, and we inject these tokens into several layers of the LM (*shared connection*). The tokens coming from early layers are injected earlier and are then replaced by those coming from deeper layers. We favor only the deeper layers (*e.g.*, the last 6 layers of the ViT-B/16, and the last 12 layers of OPT-350M) where the representations are more abstract and less modality-specific. Moreover, using the same linear projection at different representation levels might not help to capture the particularity of such a hierarchy, to this end, we also experiment with different linear layers for each level (*multiple connections*).

Multimodal Adaptation with Parameter-Efficient Techniques. We explore several parameter-efficient techniques to ease the adaptation to multimodal tasks. The main technique we use is *Prompt Tuning* [52]: it consists of prepending trainable tokens or Soft Prompts to the textual tokens input of the LM. This gives useful context to steer the model output. Contrary to hard prompts that are manually engineered, this provides a more flexible and easier approach for task-dependant contextualization. For the sake of efficiency, we prepend only 10 learnable tokens. We also experiment *Adapters* [38] as detailed later. The approach can be formalized as follows (better read with Fig.2):

$$\begin{aligned} [CLS]_i &= \mathrm{C}(\mathbf{E_i}(X)), & i &= N_E/2, ..., N_E, \\ t_j &= \mathrm{LM_j}([CLS]_i, p_{j-1}, t_{j-1}), & j &= N_L/2, ..., N_L, \end{aligned} \tag{1}$$

where $[CLS]_i$ is the perceptual token of the input X extracted from the layer i of the perceptual encoder ($\mathbf{E_i}$) with N_E layers. $[CLS]_i$ is projected using the linear connection C and prepended, alongside the Soft Prompt p to the embeddings of the textual tokens t_{j-1} coming from previous layer in the LM ($\mathrm{LM_{j-1}}$). This operation is repeated each 2 layers in the LM (with N_L layers).

3.2. Efficiency-driven Training Framework Setup

Current approaches still rely on many costly components that hinder their adoption by the large community. Specifically; they **(1)** still train a lot of parameters (*e.g.* vision encoders [85] and adapters [22] with ∼325M params/5.11%), **(2)** still maintain the multimodal pretraining with image-text pair datasets on top of the unimodal pretraining [64, 22, 85], **(3)** leverage multimodal encoders such as CLIP, pretrained on 400M image-text pairs [22, 64], **(4)** add significant computation overhead during inference, due to the long visual prompt, especially when evaluating with In Context Learning (ICL), that becomes common with LLMs [22, 64]. In this work, we propose a new setup to adapt unimodal models for multimodal downstream tasks. The setup is more challenging and is motivated by the quest for the least effort needed to exploit pretrained models. The setup is the following:

- Training only adaptation parameters (*e.g.*, Soft Prompt, linear connection), while keeping as much as possible of pretrained parameters frozen (parameter efficient).
- Avoiding multimodal pretraining and finetuning directly on downstream multimodal datasets (data/compute efficient).
- Using only pretrained unimodal models, and avoid using multimodal encoders pretrained on huge datasets (data efficient).
- Keeping fast inference (*e.g.*, 1 additional token), by avoiding long prompts, and using additional heavy modules (compute efficient).
- Using decoder-only language models (*e.g.*, OPT), the current architecture adopted by LLMs (due to its pretraining efficiency and open-ended generation capacity).

Specifically, we train only the linear connection and the soft prompt directly on the downstream multimodal tasks. This amounts to less than 1% of trainable parameters that we can push further to 0.06% with big models.

The Pretrain Zero-shot Setup. The focus of this work is direct finetuning on target datasets. However, the proposed mechanism (Sec.3.1) can be adapted straightforwardly to the pretrain-zeroshot setup. In the appendix, we show that eP-ALM outperforms previous work and it is competitive with recent SoTA following the zero-shot evaluation.

4. Experiments

Implementation details. We use OPT-2.7B in our main model, eP-ALM, and we experiment in Section 4.2 with OPT models of various sizes. We extract the [CLS] tokens of the last 6 layers of perceptual encoders and prepend them, after a linear projection, to the text tokens of the last 12 layers of the OPT. Note that we replace the previous [CLS] with the new one to keep the same number of tokens.

For VQA and VideoQA, we cast the problem as open-ended generation and compute the accuracy after a strict comparison between the output text (without truncation) and the ground truth one. Note that this setting is more challenging compared to classification-based VQA and not in favor of our approach as the model might generate semantically correct answers but using different words. We use a special token ('') to separate the question from the answer. For captioning, we report the widely adopted CIDEr and BLUE@4 scores. We finetune with the classical cross-entropy loss used to train the original OPT for VQA and Captioning tasks. We use the AdamW optimizer with a learning rate (lr) of 1e-5 warmed up to 2e-5 then decreased to 1e-6 using a cosine scheduler. We train for 8 epochs with a batch size of 64 (128 for GQA) and an image resolution of 224. Training our approach with OPT-2.7B for VQA v2 can be done on a single V100 GPU 32GB for few hours. More details are given in the appendix. We find the method sensitive to the text decoding approach (Tab. 8). Following other work, we use greedy decoding with beam search for the main results (Sec. 4.1), and multinomial/random sampling for the ablation study (Sec. 4.2).

eP-ALM Variants. Our main model, **eP-ALM** (illustrated in Figure 2), has multiple linear connections; specific learned linear layers for each [CLS] token injected in the model. In addition to Prompt Tuning. We also test variants of this model: **eP-ALM**$_{ada}$ (eP-ALM with Adapters instead of Soft Prompts), **eP-ALM**$_{lin}$ (trains a shared linear connection with all [CLS] tokens, and no prompt tuning) and **eP-ALM**$_{pt}$ (lin + Soft Prompt). For Adapters, we follow other work [22] and add sequentially one adapter module after self-attention and feedforward layers in all the blocks of OPT. While this might give better results, it adds a significant number of trainable parameters.

4.1. Main Results

In this section, we present the main comparison with other approaches. We present the results for the image modality in Section 4.1.1, the video modality in Section 4.1.2, and the audio modality in Section 4.1.3.

4.1.1 Image-Text Results

We use a frozen ViT-B/16 pretrained on ImageNet1K as the image encoder. We consider the following image-text benchmarks; VQA v2 [32], GQA [41] and COCO Caption [13]. We use Karpathy splits for VQA v2 and COCO, unless specified otherwise.

Baselines. As we are the first to propose this setup, to have a fair comparison, we reimplemented some of the existing approaches and use the same vision (ViT-ImageNet) and language (OPT) models for all:
1) $B_{PromptFuse}$; which is equivalent to PromptFuse [57] and uses Prompt Tuning (N=10). We add a linear projection for the last [CLS] token. The [CLS] token is prepended to the input of the LM. Note that we could not avoid adding a trained linear projection as there is a mismatch between the dimensions of the vision and language model.
2) B_{MAGMA}; which is equivalent to MAGMA [22] and uses Adapters. We prepend the [CLS] token to the input of LM after linear projection. Note that, we consider only the [CLS] token as we find it better than prepending all image tokens (eP-ALM$^*_{MAGMA}$). We also find that training the ViT degrades the performance, thus we keep it frozen in favor of their approach.
3) B_{LimBEr}; which is equivalent to LimBEr [64] and only trains the linear projection to project visual tokens and prepend them to the input text. Similarly, we only consider the [CLS] token as it gives better accuracy.

Comparison to Other Work. Based on our study (Sec. 4.2), we use ViT-B/16 and OPT-2.7B in our main model and in our replication of other approaches. In Table 1 we compare with other work on VQA v2, GQA, and COCO Caption. We significantly outperform other approaches with at least +10 points on VQA v2, +9 points on GQA and we double the scores on COCO Caption. eP-ALM$_{pt}$-L with OPT-6.7B and ViT-L gives the best scores while training only 0.06% of model parameters.

Note that for COCO Caption, other works give very low scores (thus we did not report them).

Few-shot Results: Are Parameter-Efficient Models also Data-Efficient? In this section, we investigate how data-efficient our model can be. To this end, we train on a very small portion (randomly sampled) from the VQA training set and evaluate on the validation set. Table 2, shows the superiority of our approach over other baselines. Interestingly, we can achieve 80% (41.9 vs 52.77) of the performance when training on 1% of the data. This validates the approach on low resources scenarios and shows that, in addition to being parameter-efficient, our model is also data-efficient.

Method	VQA v2		GQA		COCO	
	Val	Test	Val	Test	B@4	CIDEr
PromptFuse† [57]	34.1†	–	–	–	–	–
B_{LimBEr}	34.1	33.5	30.81	29.4	–	–
$B_{PromptFuse}$	40.4	39.5	33.74	31.51	15.05	48.26
B_{MAGMA}	32.2	31.8	30.98	28.93	–	–
eP-ALM$_{pt}$	48.8	47.8	43.8	40.3	27.52	91.92
eP-ALM	**50.7/53.3†**	**50.2**	**45.0**	**40.4**	**29.47**	**97.22**
eP-ALM$_{pt}$-L*	54.58/54.47†	54.47	46.86	42.7	31.24	107.0

Table 1: Comparison with other work after direct finetuning on VQA v2, GQA, and COCO Caption. eP-ALM significantly outperforms other approaches. eP-ALM uses ViT-B/16 and OPT-2.7B. eP-ALM-L uses OPT-6.7B and ViT-L/16. †: use standard split.*: trained more than 8 epochs. We use greedy decoding with beam=1.

Method	Train. data % (# of shots)	VQA v2
PromptFuse* [57]	0.12% (512)	29.40
B_{LimBEr}	1% (4.4K)	28.9
$B_{PromptFuse}$	1% (4.4K)	31.9
B_{MAGMA}	1% (4.4K)	34.5
eP-ALM$_{lin}$*	0.12% (512)	31.3
eP-ALM$_{pt}$	0.12% (512)	30.36
eP-ALM	0.12% (512)	35.54
eP-ALM	1% (4.4K)	41.9
eP-ALM	10% (44K)	47.4
eP-ALM	100% (443K)	**52.77**

Table 2: Few-shot Results on VQA v2 validation set (standard split). *: longer training.

Out of Distribution (OOD) Generalization: Do Parameter-Efficient Models Generalize Better? Here we investigate whether our parameter-efficient approach can perform well in OOD scenarios. To this end, we follow other approaches [1] and train our model on the training set of a given benchmark, and evaluate it on the validation set of another benchmark, without multimodal pretraining. We measure the performance gap, i.e. the accuracy difference between a model trained on a different benchmark and the same model trained on the target benchmark. Tab.3 shows that eP-ALM, that trains 0.06% of total parameters, is very competitive in terms of OOD accuracy with other baselines, that train all model parameters and pretrain on large amount of data. Specifically, we outperform VILBERT on VQAv2 by more than 2 points. Interestingly, the OOD-IID gap (Δ) [1] for eP-ALM, is at least 2 times lower compared to ALBEF [54] and VilBERT [60] (Note that, Δ should be considered alongside the accuracy, as a very poor model will have small Δ), this. This reveals that our parameter-efficient approach generalizes relatively well in OOD scenarios.

4.1.2 Video-Text Results

We investigate how much our approach generalizes to other modalities. To this end, we evaluate eP-ALM for Video QA on MSRVTT-QA [92] and MSVD-QA [92] and for Video

Method	Multimodal PT data	Trained param. (%)	Train data	Test data VQA v2	Test data GQA	Gap Δ
ALBEF [1]	14M	100%	VQA v2	–72.1	50.1	-21.8
			GQA	50.3	–64.2	-14.1
VILBERT [1]	3M	100%	VQA v2	–62.2	42.6	-20.4
			GQA	41.8	–65.3	-22.7
eP-ALM$_{pt}$-L *(nb3)	0	0.06%	VQA v2	–54.78	41.39	**-9.59**
			GQA	45.19	–47.19	**-5.8**

Table 3: Out-Of-Distribution Generalization on GQA and VQA v2 (standard split). The Gap shows the performance degradation when the model is trained on a different dataset.

Captioning on MSR-VTT [93]. For the video encoding, we use the TimeSformer-base[5] model pretrained on Kinetics-600 [10]. We use 8 and 16 224x224 frames for VQA and captioning respectively.

Comparison to other work to the best of our knowledge, FrozenBiLM [96] is the only parameter-efficient work proposing to adapt LMs for video-language tasks. It uses Adapters to adapt the frozen CLIP-ViT and Bidirectional LM for Video QA. We compare our approach to our re-implementation of this baseline; where we train only the Adapters and the linear projection layer to project the last [CLS] token and prepend it to the input text ones. The results in Tab. 4 show that eP-ALM outperforms this baseline by a significant margin. The reason why the latter does not give good results might be due to prepending the visual tokens to the input of OPT. We can reduce the number of parameters and slightly degrade the performance by using a shared linear connection (eP-ALM vs eP-ALM$_{pt}$).

Method	Trained param (%)	MSVD-QA Test	MSRVTT-QA Test	MSRVTT CIDEr	MSRVTT B@4
$B_{FrozenBiLM}$ (from [96])	3.72 %	14.58	6.33	-	-
eP-ALM	0.89 %	**38.64**	**36.16**	**47.31**	**38.51**
eP-ALM$_{pt}$	0.54 %	38.79	35.62	45.30	39.34

Table 4: Comparison with different approaches after direct finetuning on MSVD-QA, MSRVTT-QA, and MSRVTT Caption.

Zero-Shot Results To explore the generalization of our approach, we evaluate on Zero-Shot for VideoQA, where the model is trained on a dataset different from the target one. Table 5 shows a comparison with other approaches. eP-ALM (with TimeSformer as encoder), trained on VQA v2 (standard split) by considering the image as a video with a single frame, outperforms other approaches trained on significantly more data. Specifically, eP-ALM outperforms Flamingo-3B [2] on MSRVTT-QA by more than 2 points, and attains double the scores of FrozenBiLM [96]. Contrary to some of other approaches that cast the task as classification (similarity-based) [95] or constrained generation through masking, considering only a subset of answers (1k or 2k) [96, 100, 55], our approach is evaluated (with a character-wise comparison with the ground-truth) with unconstrained Open-ended Generation (OE Gen) and can generate answers with arbitrary lengths. This is more challenging and not in favor of our approach.

Method	Training data	Train. Param. (%)	OE Gen	MSRVTT-QA	MSVD-QA
JustAsk [95]	ActivityNet-QA	89.6%	✗	2.7	-
JustAsk [95]	HowToVQA69M	89.6%	✗	2.9	7.5
LAVENDER [55]	WebVid2.5M+CC3M	100%	✗	4.5	11.6
MERLOT Reserve [100]	YT-Temporal-1B	100%	✗	5.8	-
FrozenBiLM † [96]	400M-CLIP + VQA v2	2.9%	✗	6.9	12.6
Flamingo 3B [2]	M3W+ALIGN+VTP	40%	✓	11.0	**27.5**
eP-ALM	VQA v2	0.9%	✓	13.17	24.82
eP-ALM †	VQA v2	0.9%	✓	**14.54**	27.09

Table 5: Zero-Shot results on Video QA. OE Gen: unconstrained Open-Ended Generation. † evaluated on questions with top 1k answers.

4.1.3 Audio-Text Results

We investigate the generalization of our approach to the audio domain. The encoder is AST-base model [30] pre-trained for classification on AudioSet [29]. We evaluate on AudioCaps dataset [47], the largest benchmark for Audio Captioning. We train with mel spectrograms of 128 bins and frequency and time masking with a batch size of 8.

To the best of our knowledge, no prior work has been proposed to efficiently adapt LM for audio-text tasks, thus we compare with other end-to-end trained SoTA that takes only the audio signal as input. Tab. 6 shows that our approach is very competitive with previous work, showing the potential of efficient adaptation of LM for the audio modality.

Method	Trained	AudioCaps					
	param (%)	B@1	B@2	METEOR	CIDEr	SPICE	SPIDEr
Kim et al. [48]	100%	0.614	0.446	0.203	0.593	0.144	0.369
Koizumi et al. [51]	100%	0.638	0.458	0.199	0.603	0.139	0.371
Eren et al. [23]	100%	**0.710**	0.490	**0.290**	0.750	-	-
Xu et al. [94]	100%	0.655	0.476	0.229	0.660	0.168	0.414
Mei et al. [63]	100%	0.647	0.488	0.222	0.679	0.160	0.420
Gontier et al. [31]	100%	0.699	**0.523**	0.241	**0.753**	**0.176**	**0.465**
Liu et al. [59]	100%	0.671	0.498	0.232	0.667	0.172	0.420
eP-ALM (nb3)	0.90 %	66.08	47.57	22.69	63.61	16.29	-

Table 6: Comparison with other work for Audio Captioning on AudioCaps Test set.

4.1.4 Comparison with SoTA

To contextualize the work, we compare eP-ALM to other SoTA that trains large number of parameters and most often with large-scale pretraining. Tab. 7 shows a comparison with both zero-shot (ZS) and Finetuning (FT) setups. The performance of eP-ALM is generally higher than ZS scores and still below FT ones. However, the performance gap with FT models, is smaller with the audio and video modalities.

Dataset (Metric)	SoTA (ZS)	eP-ALM (FT)	SoTA (FT)
AudioCaps (CIDEr)	–	<u>63.6</u>	**66.7** (Liu et al. [59])
MSRVTT-QA (Acc)	17.4 (Flamingo80B [2])	<u>36.7</u>	**44.1** (OmniVL [87])
MSR-VTT (CIDEr)	–	<u>50.7</u>	**60** (MV-GPT [71])
COCO (CIDEr)	84.3 (Flamingo80B [2])	<u>107.0</u>	**145.3** (OFA [88])
VQAv2 (Acc)	<u>56.3</u> (Flamingo80B [2])	53.3	**84.3** (PaLI [14])
GQA (Acc)	29.3 (FewVLM [44])	<u>42.7</u>	**60.8** (VL-T5 [17])

Table 7: Comparison of eP-ALM with text generation-based SoTA that train significant number of parameters, including methods with large-scale pretraining. Best and next best scores are bolded and underlined respectively. FT: Finetuning. ZS: Zero-shot.

4.2. Ablation Study

In this section, we ablate different component of our work.

Comparison between different text generation methods. We find the approach sensitive to the text decoding strategy. In Tab. 8, we compare with different text decoding methods; multinomial/random sampling [24] and greedy decoding with beam-search (1 to 5 beams). Greedy decoding significantly outperform multinomial sampling, and increasing the number of beams leads to additional improvements, to the detriment of increasing inference cost.

Decoding Method	# of beams	VQA v2	GQA	COCO		MSVD-QA	MSRVTT-QA	MSRVTT	
		Test	Test	B@4	CIDEr	Test	Test	B@4	CIDEr
Multinomial	1	43.86	37.02	13.37	58.60	31.18	26.96	20.77	14.13
Greedy	1	54.47	42.70	31.24	107.0	38.64	36.16	47.31	38.51
Greedy	3	54.90	42.99	33.35	113.0	38.93	36.66	50.66	39.02
Greedy	5	54.92	42.95	33.86	115.54	–	–	–	–

Table 8: Comparison with different text generation mechanisms using eP-ALM-L.

In the following, we run some ablations for Image-Text tasks, mostly on VQA v2.

Comparison with different variants and baselines. We start by comparing the different variants to other work in Fig. 3. All models use OPT-350M and ViT-B/16. Other approaches lag significantly behind our model. B_{MAGMA} gives the best results (23.3% acc.) among them, followed by $B_{PromptFuse}$ (18.82% acc.) and finally B_{LimBEr} (10.75 % acc.). We also compare with another MAGMA baseline ($B_{MAGMA}{}^*$) that prepends all visual tokens to the input, and we find a significant degradation compared to passing only the [CLS] token. This reveals that prepending all visual tokens directly to the input hinders the adaptation.

We can notice a consistent improvement of eP-ALM when adding more trainable parameters. The most parameter-efficient model is eP-ALM$_{lin}$ which has 30.72%, while the best has 34.34% (with Adapters eP-ALM$_{ada}$). Interestingly, eP-ALM$_{lin}$ with only one linear layer succeeds to get good performance on this challenging setup, revealing that the

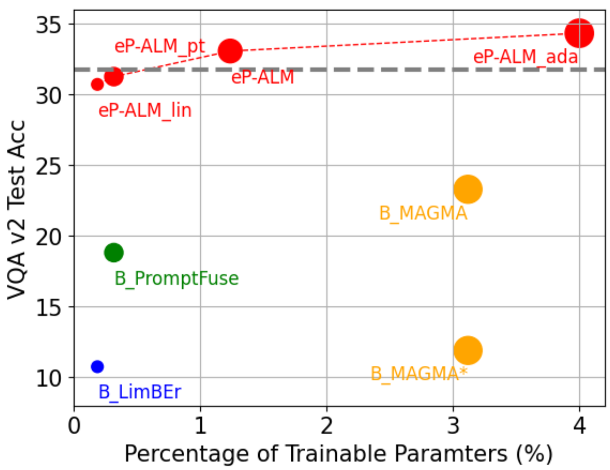

Figure 3: Comparison with other baselines on VQA v2. eP-ALM uses ViT-B/16 and OPT-350M. Our approach significantly outperforms other approaches. eP-ALM already surpasses the dashed "upper" baseline that trains all model parameters. Allocating more adaptation parameters help to increase the scores. Marker size indicates model size.

language and visual representation spaces are not very far. Other parameter-efficient techniques such as Prompt Tuning can help to get additional points (30.72 with eP-ALM$_{lin}$ vs 31.27 with eP-ALM$_{pt}$). Moreover, using different layers for each injected [CLS] token seems to give significant improvement (31.27 with eP-ALM$_{pt}$ vs 33.08 with eP-ALM).

Finally, we show that eP-ALM surpasses the "full finetuning" baseline (grayed line) that finetune all parameters by 1.27 points (31.79 vs 33.08). This reveals that training all weights of pretrained models on small datasets can reduce their generalization capability and degrades performance.

As a trade-off between performance and efficiency, we favor eP-ALM which we carry on for the following study.

Extraction and Injection Level of [CLS] Tokens. Here we investigate which [CLS] tokens to extract from the ViT and where is the best position to inject them inside the OPT model. Table 9 shows that extracting the last [CLS] tokens (from the last 6 layers) is better than using only the last one, as done in other approaches (Acc 30.53 vs 33.08). In addition, using all [CLS] tokens seems to degrade the performance. Moreover, prepending [CLS] tokens to all OPT layers degrades slightly (33.08 vs 32.15), and prepending to the input of OPT gives the worst results. This might indicate that it is easier to merge visual and textual tokens deeper in the model, where the representations are more abstract, compared to the first layers where we have more modality-specific features and higher representation mismatch.

[CLS] tokens		VQA v2
From ViT layers	To OPT layers	Test Acc.
12	12 to 23	30.53
6 to 12	12 to 23	33.08
1 to 12	12 to 23	31.17
6 to 12	1 to 23	32.15
12	1	18.82

Table 9: Ablation study. Extracting the [CLS] tokens from the last layers of ViT (layers 6 to 12) is better than taking only the last token (layer 12). Injecting the [CLS] tokens lately (layers 12-24) in the OPT is better than injecting them in all layers or only in the input.

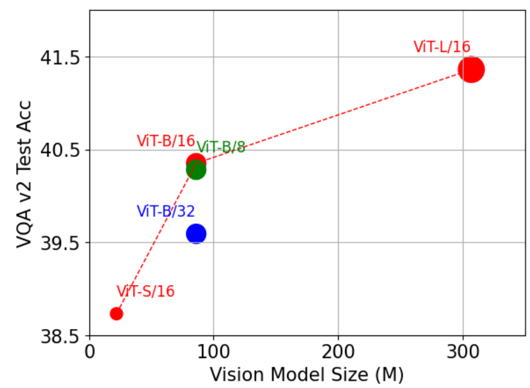

Figure 4: Scaling Vision Model; the score increases with the size of the ViT. Increasing the patch resolution beyond 16 does not help.

Scaling LM. An interesting question that we investigate is the impact of scaling the language model's parameters on our approach. Ideally, we would like to have an approach that efficiently exploits LLMs for other tasks and modalities, without having access to enormous computational resources. In Table 5, we show that the scores increase with the model size with the biggest jump being between OPT-350M (33.08 vs 37.29) and OPT-1.3B ($\sim \times 4$ the model size). The consistent improvement with scale shows the effectiveness of the approach when considering very big models.

Scaling Visual Model. We also study how the model behaves when scaling the visual encoder. In Figure 4, we can notice that the scores increase with the size of the ViT. The best is ViT-L/16 (41.36) and the worst is ViT-S/16 (Acc 38.73). However, the ViT resolution or the number of image patches/tokens does not seem to have a significant effect on the final performance after a resolution of 16.

Scaling Compute. Table 10 shows that our approach scales with compute, as training for more epochs leads to 4 points gain in VQA accuracy. Interestingly with OPT-6.7B and ViT-L (eP-ALM$_{pt}$-L), we achieve a score of 43.6 by

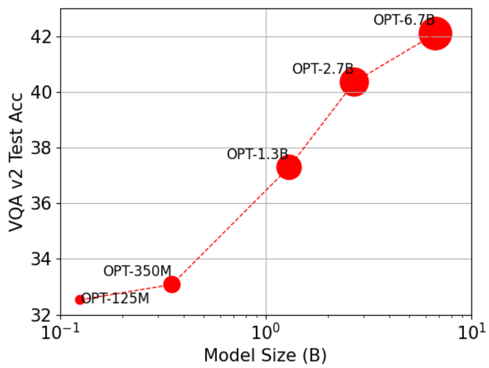

Figure 5: Scaling LM; we scale the OPT from 125M to 6.7B parameter. eP-ALM becomes more effective with scale. The biggest performance jump is when scaling beyond 1B parameters (1.3B).

training only **0.06%** of model parameters (∼4M params).

As a trade-off between performance and model size, we favor OPT-2.7B and use it for all other experiments.

Method	number of epochs	VQA v2 Val
eP-ALM	8	38.9
eP-ALM	32	42.9
eP-ALM$_{pt}$-L	8	42.5
eP-ALM$_{pt}$-L	32	43.6

Table 10: Scaling Compute. Evaluation on VQA v2 standard split.

5. Conclusion

In this work, we propose a new challenging setup to efficiently adapt unimodal models for multimodal tasks, which is centered around augmenting existing LMs with perception. Without multimodal pretraining, and with almost 4M trainable parameters consisting of a linear connection and a Soft Prompt, we can adapt a frozen 7B model and reach an accuracy of 54.5% on VQA v2, with unconstrained open-ended generation. We validate the effectiveness of the approach with Images, Video, and Audio modalities. This direct finetuning setup has several advantages; (a) training data/compute efficiency, (b) attains generally higher performance than pretrain-zeroshot setup, (c) easy to adapt to new tasks, modalities or other LLMs, where no costly pretraining is needed. However, the mechanism proposed in eP-ALM can be adapted in a straightforward manner to this setup.

Even though the results are still far from the state-of-the-art approaches that train most of the model parameters on much more data, the extremely small percentage of trainable parameters (0.06%) and the increasing scores with model size and compute make the work promising towards finding an intermediate point, between extremely efficient and extremely inefficient approaches, which is hopefully closer to the former.

The method has some limitations, which we illustrate in the appendix. In general, the model struggles to capture fine-grained details in the images, favors coherent generation over factual one, might hallucinate some objects not present in the image, and lacks common sense reasoning. Our approach inherits most of the limitations and biases of pretrained models, especially the LM, and training only a few adaptation parameters does not seem to avoid the transfer of these biases. Finally, the model is trained with next token prediction and is able to produce coherent text, however, it is still not clear how this paradigm can lead to real reasoning capabilities.

Acknowledgments This work was partly supported by ANR grant VISA DEEP (ANR-20-CHIA-0022), and HPC resources of IDRIS under the allocation 2022-[AD011013415] and 2023-[AD011013415R1] made by GENCI. The authors would like to thank Theophane Vallaeys for fruitful discussion.

References

[1] Aishwarya Agrawal, Ivana Kajić, Emanuele Bugliarello, Elnaz Davoodi, Anita Gergely, Phil Blunsom, and Aida Nematzadeh. Rethinking evaluation practices in visual question answering: A case study on out-of-distribution generalization. *arXiv preprint arXiv:2205.12191*, 2022. 6

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022. 2, 6, 7

[3] Ankur Bapna and Orhan Firat. Simple, scalable adaptation for neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1538–1548, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. 3

[4] Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 830–839, 2020. 3

[5] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4, 2021. 2, 3, 6

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 1, 2

[7] Remi Cadene, Hedi Ben-Younes, Matthieu Cord, and Nicolas Thome. Murel: Multimodal relational reasoning for

visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1989–1998, 2019. 2

[8] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021. 2

[9] Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*, 2018. 3

[10] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*, 2019. 6

[11] Jun Chen, Han Guo, Kai Yi, Boyang Li, and Mohamed Elhoseiny. Visualgpt: Data-efficient adaptation of pretrained language models for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18030–18040, 2022. 3

[12] Shoufa Chen, Chongjian GE, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. 3

[13] 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. 2, 5

[14] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022. 1, 2, 7

[15] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020. 2

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

[17] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In *International Conference on Machine Learning*, pages 1931–1942. PMLR, 2021. 7

[18] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022. 1, 2

[19] Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, et al. Scaling vision transformers to 22 billion parameters. *arXiv preprint arXiv:2302.05442*, 2023. 1

[20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 1, 3

[21] Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, et al. Glam: Efficient scaling of language models with mixture-of-experts. In *International Conference on Machine Learning*, pages 5547–5569. PMLR, 2022. 3

[22] Constantin Eichenberg, Sidney Black, Samuel Weinbach, Letitia Parcalabescu, and Anette Frank. Magma–multimodal augmentation of generative models through adapter-based finetuning. *arXiv preprint arXiv:2112.05253*, 2021. 2, 3, 4, 5

[23] Ayşegül Özkaya Eren and Mustafa Sert. Audio captioning based on combined audio and semantic embeddings. In *2020 IEEE International Symposium on Multimedia (ISM)*, pages 41–48, 2020. 7

[24] Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833*, 2018. 7

[25] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. From captions to visual concepts and back. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1473–1482, 2015. 2

[26] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022. 1

[27] Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. Violet: End-to-end video-language transformers with masked visual-token modeling. *arXiv preprint arXiv:2111.12681*, 2021. 2

[28] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020. 3

[29] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017. 3, 7

[30] Yuan Gong, Yu-An Chung, and James Glass. AST: Audio Spectrogram Transformer. In *Proc. Interspeech 2021*, pages 571–575, 2021. 2, 3, 7

[31] Félix Gontier, Romain Serizel, and Christophe Cerisara. Automated audio captioning by fine-tuning bart with audioset tags. In *DCASE 2021 - 6th Workshop on Detection and Classification of Acoustic Scenes and Events*, Virtual, Spain, Nov. 2021. 7

[32] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 2, 5

[33] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020. 2

[34] Yaru Hao, Haoyu Song, Li Dong, Shaohan Huang, Zewen Chi, Wenhui Wang, Shuming Ma, and Furu Wei. Language models are general-purpose interfaces. *arXiv preprint arXiv:2206.06336*, 2022. 2

[35] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 2

[36] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. 2

[37] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022. 2, 3

[38] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR, 2019. 3, 4

[39] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. 3

[40] Ronghang Hu, Anna Rohrbach, Trevor Darrell, and Kate Saenko. Language-conditioned graph networks for relational reasoning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10294–10303, 2019. 2

[41] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 5

[42] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIII*, page 709–727, Berlin, Heidelberg, 2022. Springer-Verlag. 3

[43] Huaizu Jiang, Ishan Misra, Marcus Rohrbach, Erik Learned-Miller, and Xinlei Chen. In defense of grid features for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10267–10276, 2020. 2

[44] Woojeong Jin, Yu Cheng, Yelong Shen, Weizhu Chen, and Xiang Ren. A good prompt is worth millions of parameters: Low-resource prompt-based learning for vision-language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2763–2775, Dublin, Ireland, May 2022. Association for Computational Linguistics. 7

[45] Rabeeh Karimi Mahabadi, James Henderson, and Sebastian Ruder. Compacter: Efficient low-rank hypercomplex adapter layers. *Advances in Neural Information Processing Systems*, 34:1022–1035, 2021. 3

[46] Lei Ke, Wenjie Pei, Ruiyu Li, Xiaoyong Shen, and Yu-Wing Tai. Reflective decoding network for image captioning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8888–8897, 2019. 2

[47] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. Audiocaps: Generating captions for audios in the wild. In *NAACL-HLT*, 2019. 2, 7

[48] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. AudioCaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 119–132, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. 7

[49] Ju-ho Kim, Jungwoo Heo, Hyun-seo Shin, Chan-yeong Lim, and Ha-Jin Yu. Integrated parameter-efficient tuning for general-purpose audio models. *arXiv preprint arXiv:2211.02227*, 2022. 3

[50] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021. 2

[51] Yuma Koizumi, Yasunori Ohishi, Daisuke Niizumi, Daiki Takeuchi, and Masahiro Yasuda. Audio captioning using pre-trained large-scale language model guided by audio-based similar caption retrieval. *arXiv preprint arXiv:2012.07331*, 2020. 7

[52] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, 2021. 3, 4

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

[54] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in Neural Information Processing Systems*, 34, 2021. 1, 2, 6

[55] Linjie Li, Zhe Gan, Kevin Lin, Chung-Ching Lin, Zicheng Liu, Ce Liu, and Lijuan Wang. Lavender: Unifying video-language understanding as masked language modeling. *arXiv preprint arXiv:2206.07160*, 2022. 2, 7

[56] Manling Li, Ruochen Xu, Shuohang Wang, Luowei Zhou, Xudong Lin, Chenguang Zhu, Michael Zeng, Heng Ji, and Shih-Fu Chang. Clip-event: Connecting text and images with event structures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16420–16429, 2022. 2

[57] Sheng Liang, Mengjie Zhao, and Hinrich Schütze. Modular and parameter-efficient multimodal fusion with prompting. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2976–2985, 2022. 2, 3, 5, 6

[58] Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *arXiv preprint arXiv:2205.05638*, 2022. 3

[59] Xubo Liu, Xinhao Mei, Qiushi Huang, Jianyuan Sun, Jinzheng Zhao, Haohe Liu, Mark D Plumbley, Volkan Kilic, and Wenwu Wang. Leveraging pre-trained bert for audio captioning. In *2022 30th European Signal Processing Conference (EUSIPCO)*, pages 1145–1149. IEEE, 2022. 7

[60] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 2, 6

[61] Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. *arXiv preprint arXiv:2206.08916*, 2022. 2

[62] Rabeeh Karimi Mahabadi, Sebastian Ruder, Mostafa Dehghani, and James Henderson. Parameter-efficient multi-task fine-tuning for transformers via shared hypernetworks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 565–576, 2021. 3

[63] XINHAO MEI, XUBO LIU, QIUSHI HUANG, MARK DAVID PLUMBLEY, and WENWU WANG. Audio captioning transformer. In *Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE 2021)*. 7

[64] Jack Merullo, Louis Castricato, Carsten Eickhoff, and Ellie Pavlick. Linearly mapping from image to text space. *arXiv preprint arXiv:2209.15162*, 2022. 3, 4, 5

[65] Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ram Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, et al. Augmented language models: a survey. *arXiv preprint arXiv:2302.07842*, 2023. 2

[66] Jong Hak Moon, Hyungyung Lee, Woncheol Shin, Young-Hak Kim, and Edward Choi. Multi-modal understanding and generation for medical images and text via vision-language pre-training. *IEEE Journal of Biomedical and Health Informatics*, 26(12):6070–6080, 2022. 2

[67] Junting Pan, Ziyi Lin, Xiatian Zhu, Jing Shao, and Hongsheng Li. St-adapter: Parameter-efficient image-to-video transfer learning for action recognition. *arXiv preprint arXiv:2206.13559*, 2022. 3

[68] Roma Patel and Ellie Pavlick. Mapping language models to grounded conceptual spaces. In *International Conference on Learning Representations*, 2022. 2

[69] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 3

[70] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022. 1, 2

[71] Paul Hongsuck Seo, Arsha Nagrani, Anurag Arnab, and Cordelia Schmid. End-to-end generative pretraining for multimodal video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17959–17968, 2022. 7

[72] Erica K Shimomoto, Edison Marrese-Taylor, Hiroya Takamura, Ichiro Kobayashi, Hideki Nakayama, and Yusuke Miyao. Towards parameter-efficient integration of pre-trained language models in temporal video grounding. *arXiv preprint arXiv:2209.13359*, 2022. 3

[73] Mustafa Shukor, Guillaume Couairon, and Matthieu Cord. Efficient vision-language pretraining with visual concepts and hierarchical alignment. In *33rd British Machine Vision Conference (BMVC)*, 2022. 1, 2, 3

[74] Mustafa Shukor, Nicolas Thome, and Matthieu Cord. Structured vision-language pretraining for computational cooking. *arXiv preprint arXiv:2212.04267*, 2022. 2

[75] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15638–15650, 2022. 1

[76] Ishika Singh, Valts Blukis, Arsalan Mousavian, Ankit Goyal, Danfei Xu, Jonathan Tremblay, Dieter Fox, Jesse Thomason, and Animesh Garg. Progprompt: Generating situated robot task plans using large language models. *arXiv preprint arXiv:2209.11302*, 2022. 2

[77] Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, et al. Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model. *arXiv preprint arXiv:2201.11990*, 2022. 1, 2

[78] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022. 2

[79] Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. Lst: Ladder side-tuning for parameter and memory efficient transfer learning. In *Advances in Neural Information Processing Systems*. 3

[80] Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5227–5237, 2022. 3

[81] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? *Advances in Neural Information Processing Systems*, 33:6827–6839, 2020. 2

[82] Anthony Meng Huat Tiong, Junnan Li, Boyang Li, Silvio Savarese, and Steven CH Hoi. Plug-and-play vqa: Zero-shot vqa by conjoining large pretrained models with zero training. *arXiv preprint arXiv:2210.08773*, 2022. 2

[83] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021. 3

[84] Hugo Touvron, Matthieu Cord, and Hervé Jégou. Deit iii: Revenge of the vit. In *ECCV*, 2022. 3

[85] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212, 2021. 2, 3, 4

[86] Ben Wang and Aran Komatsuzaki. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. https://github.com/kingoflolz/mesh-transformer-jax, May 2021. 3

[87] Junke Wang, Dongdong Chen, Zuxuan Wu, Chong Luo, Luowei Zhou, Yucheng Zhao, Yujia Xie, Ce Liu, Yu-Gang Jiang, and Lu Yuan. Omnivl: One foundation model for image-language and video-language tasks. *arXiv preprint arXiv:2209.07526*, 2022. 2, 7

[88] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *arXiv preprint arXiv:2202.03052*, 2022. 2, 7

[89] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022. 1, 2

[90] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022. 2

[91] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018. 2

[92] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM International Conference on Multimedia*, MM '17, page 1645–1653, New York, NY, USA, 2017. Association for Computing Machinery. 2, 6

[93] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 6

[94] Xuenan Xu, Heinrich Dinkel, Mengyue Wu, Zeyu Xie, and Kai Yu. Investigating local and global information for automated audio captioning with transfer learning. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 905–909. IEEE, 2021. 7

[95] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just ask: Learning to answer questions from millions of narrated videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1686–1697, 2021. 6, 7

[96] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Zero-shot video question answering via frozen bidirectional language models. In *NeurIPS 2022-36th Conference on Neural Information Processing Systems*, 2022. 3, 6, 7

[97] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. An empirical study of gpt-3 for few-shot knowledge-based vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3081–3089, 2022. 2

[98] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 2

[99] Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1–9, 2022. 3

[100] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16375–16387, 2022. 7

[101] Andy Zeng, Adrian Wong, Stefan Welker, Krzysztof Choromanski, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, et al. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv preprint arXiv:2204.00598*, 2022. 2

[102] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022. 1, 2, 3

[103] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language

models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022. 3