Supplementary material for 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

The appendix is organized as follows;

- Appendix A: gives more implementation details about the experiments that we conduct.
- Appendix B: illustrates and explain the different variants of eP-ALM.
- Appendix C: presents more ablation studies on image-text and video-text tasks.
- Appendix D: shows some qualitative results.
- Appendix E: discusses the limitation of the proposed approach.

A. Implementation Details

We use OPT-2.7B in our final model. We extract the [CLS] tokens of the last 6 layers of perceptual encoders and prepend them, after 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. We finetune with the classical cross-entropy loss used to train the original OPT for VQA and Captioning. We use the AdamW optimizer with a lr of 1e-5 warmed up to 2e-5 then decreased to 1e-6 using a cosine scheduler. For **Adapters**, we use sequential Adapters after self attentions and feedforward layers with a downsample factor of 8 and ReLU activation. For **Soft Prompt**, we implement it as a linear embedding layer that takes numbers from 0 to the length of the prompt (here 10). We experiment also with adding an MLP after the prompts as done with other approaches [8]. We use the prompt with MLP for most of the experiments as we find that it gives slightly better results. The soft prompt and adapters are trained with a fixed lr of 1e-5. **eP-ALM**$_{pt}$**-L** is trained with a light-weight prompt (only trainable tokens without MLP), starting learning rate of 2e-4 and a fixed learning rate of 1e-3 for the prompt with a total batch size of 16.

VQA/GQA: we use a special token for VQA ($' < /a >'$) to separate the question from the answer. 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 1.8 days (as the perceptual encoder is frozen, saving its output tokens can save a lot of training time). For Few-shot experiments, we train longer (for 64 epochs) with higher starting learning rate (1e-4 warmed up to 2e-4 and decreased to 1e-5). Those marked by by a * are trained for 100 epochs as in PromptFuse [5].

Image Captioning we train for 8 epochs with a batch size of 64 and an image resolution of 224.

Video QA: we sample randomly 8 frames of resolution 224x224 for each video and train for 25 epochs with a batch size of 32. For Zero-Shot experiments, we train only for 4 epochs with starting learning rate of 1e-4. We use only the spatial self attention of TimeSformer to train on VQA v2.

Video Captioning: we sample randomly 16 frames of resolution 224x224 for each video and train for 25 epochs with a batch size of 64.

Audio Captioning we train for 30 epochs with frequency and time masking of 24 and 96 respectively. The mel bins is 128 and the audio length is 1024. Batch size 32. For Deep Prompt, we inject new soft prompt in all the 32 blocks of OPT (each with length 10).

B. eP-ALM Variants

We detail the different variants proposed in this paper (here we consider ViT-B/16 and OPT-350M for simplicity). These variants are illustrated in Fig.1:

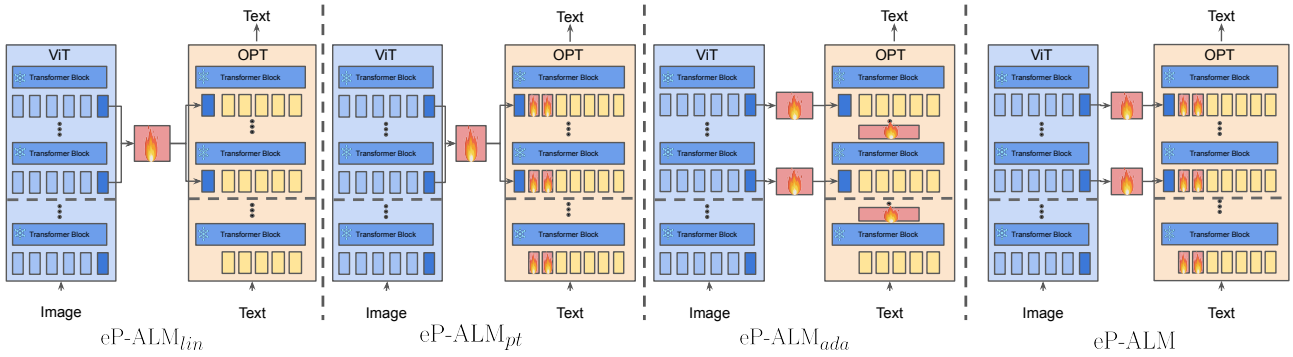

Figure 1: Illustration of the different variants of eP-ALM; eP-ALM$_{lin}$ is the most efficient variant that only trains the linear projection layer, eP-ALM$_{pt}$ adds trainable Soft Prompts (*i.e* Prompt Tuning), and eP-ALM$_{ada}$ replaces the Soft Prompt in eP-ALM (last figure) with trainable Adapters. All models extract the [CLS] tokens from the last layers of ViT and prepend/replace them in the last layers of OPT.

eP-ALM$_{lin}$: we extract the [CLS] tokens from the last 6 layers of the frozen ViT and inject them in the last 12 layers of the frozen OPT. To reduce inference cost, each couple of layers (here 2), we replace the previous [CLS] with the new one (thus only increasing the number of tokens by 1 the whole process). All visual [CLS] tokens are projected by one trainable linear projection layer (shared) to fit their dimension to that of the OPT.

eP-ALM$_{pt}$: we augment eP-ALM$_{lin}$ with Prompt Tuning, which consists of prepending trainable tokens (*i.e*, soft prompt) to the input of the LM. This might help the model to adapt well to the new task by providing context to the text input. For the sake of efficiency, we prepend only 10 learnable tokens.

eP-ALM: while one linear projection is appealing, it might not be able to capture all the particularity of different [CLS] tokens. To overcome this, we use different projections for each [CLS], while keeping the soft prompt.

eP-ALM$_{ada}$: another alternative to Prompt Tuning are Adapters. We follow other approaches [2] and add sequentially one adapter module (downsample, activation then upsample) after self attention and feedforward layers in all the blocks of OPT. While this might give better results, it adds significant number of trainable parameters.

C. Ablation Study

Here we present additional ablation study.

C.1. Image-Text

Training All Parameters Here we investigate how much gain we can obtain by unfreezing the pretrained models. We experiment on VQA v2 with eP-ALM. Table 1 shows

Trainable Models		LM	VQA v2
VM	LM	size	test Acc.
✗	✗	350M	33.08
✗	✓	350M	35.44
✓	✓	350M	35.47

Table 1: Ablation study: we study how much gain we can obtain by also training the pretrained vision and language models. We see slight improvement by training the pretrained models.

that finetuning the pretrained models in our eP-ALM gives slight improvement, despite the large number of trainable parameters. Note that, we find that using very small learning rate (lr=1e-7) is the only option (while keeping an lr of 1e-5 for the connectors) to unfreeze these models without significant degradation.

C.2. Video-Text

Video Encoder: here we compare different encoders to process the videos. We compare the TimeSformer [1] that has both spatial and temporal attention and trained for video classification with a simple baseline, ViT trained on ImageNet, that ignores the temporal dynamics. For ViT, we take the average of [CLS] tokens of the processed frames while for TimeSformer we consider the one [CLS] token. Table 2 shows that using video-specific encoders gives significantly better results for video captioning. In addition, we find that using 16 frames instead of 8 gives slight improvement.

Injection and Extraction level of [CLS] tokens: here we show the importance of leveraging the hierarchical representation in both the video encoder and language model. Table 3 shows the results on MSVD-QA. We show that keeping the interaction between cross modal tokens to the last layers (layer 19 to 31) of the OPT leads to significantly better

Method	MSRVTT	
	CIDEr	B@4
ViT-B Avg.	17.96	12.77
ViT-B Avg. (16 f)	17.82	12.85
TimeSformer	20.11	13.53
TimeSformer (16 f)	20.58	14.12

Table 2: Ablation (Caption) MSRVTT Caption.

results. Extracting several tokens from different tokens of the TimeSformer gives slight improvement. However, using hierarchical video transformers [4, 6] might lead to better results. We noticed also that Adapters generally give better results than Prompt Tuning, this might be because when training on videos we sample randomly some frames, which prevent the model to overfit in case of small datasets.

Adaptation approach	[CLS] tokens from encoder layers	to OPT layers	MSVD-QA test Acc.
Soft Prompt	12	1	13.49
	12	1 to 31	27.16
	12	19 to 31	30.86
	6 to 12	19 to 31	31.18
Adapters	12	1	12.40
	12	1 to 31	34.86
	12	19 to 31	35.94

Table 3: Ablation study: we investigate the extraction and injection position of [CLS] tokens for Video QA.

C.3. Audio-Text

Time and Frequency Masking: following other approaches [3, 7] we train eP-ALM with time and frequency masking on AudioCaps. Table 4 shows that masking significantly help, however, using too much of masking hurt the performance.

Masking Window		AudioCaps	
Time	Frequency	CIDEr	B@4
256	64	33.94	10.21
192	48	35.67	10.40
96	24	37.14	11.37
0	0	36.01	10.23

Table 4: Ablation Study: time and frequency masking help for Audio Captioning.

D. Qualitative Results

We show some qualitative results of our eP-ALM model with OPT-2.7B in Fig. 2. For VQA, we can notice that our model is able to correctly answer the questions. Moreover, some of the answers are richer and more accurate than the manually labeled ground truth in the dataset. This also reveals that the exact matching evaluation protocol is not in favor of the open-ended generation produced by our model. Interestingly, it seems that the model learned the answering style in the training set (*i.e*, short and concise answers). For Captioning, the model can generate coherent sentences describing the image globally. However, it still misses some details in the image.

Figure 2: Qualitative results of eP-ALM: the model is able to generate accurate answers and coherent descriptions of the image.

E. Limitations

Even though we show appealing results for very efficient training, the method has several limitations, which we illustrate some of them in Fig. 3. For VQA, we can notice that the model is unable to capture finegrained details in the images (*e.g.*, number of colours and the zebra in the first 2 examples), which might be due to constraining the interaction with the vision model through the [CLS] tokens, that generally capture global information about the image. In case of hard questions, the model favor coherent generation of a relevant question followed by its correct answer, instead of answering the main question ("A: what color is the phone?? black" in example 3).

For Captioning, the model seems to favor outputting a coherent sentence, even though it is not entirely correct

("many" cows in a "crowded" city). Secondly, the model might hallucinate some objects that does not appear in the image ("apples" in the example 2). Finally, the model lacks common sense reasoning, making him unable to understand that elephants are not small and being far from the camera does not change this fact (example 3).

Our approach inherents most of the limitations and biases of pretrained models, especially the LM, and training only 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.

References

[1] 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

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

[4] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *CVPR*, 2022. 3

[5] 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. 1

[6] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211, 2022. 3

[7] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. Specaugment: A simple data augmentation method for automatic speech recognition. *Proc. Interspeech 2019*, pages 2613–2617, 2019. 3

[8] 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. 1

Figure 3: Illustration of some limitations of eP-ALM: the model struggles to capture finegrained details, favors coherence over factual responses, hallucinates some objects and lacks common sense reasoning. Ground truth answers are highlighted in green.

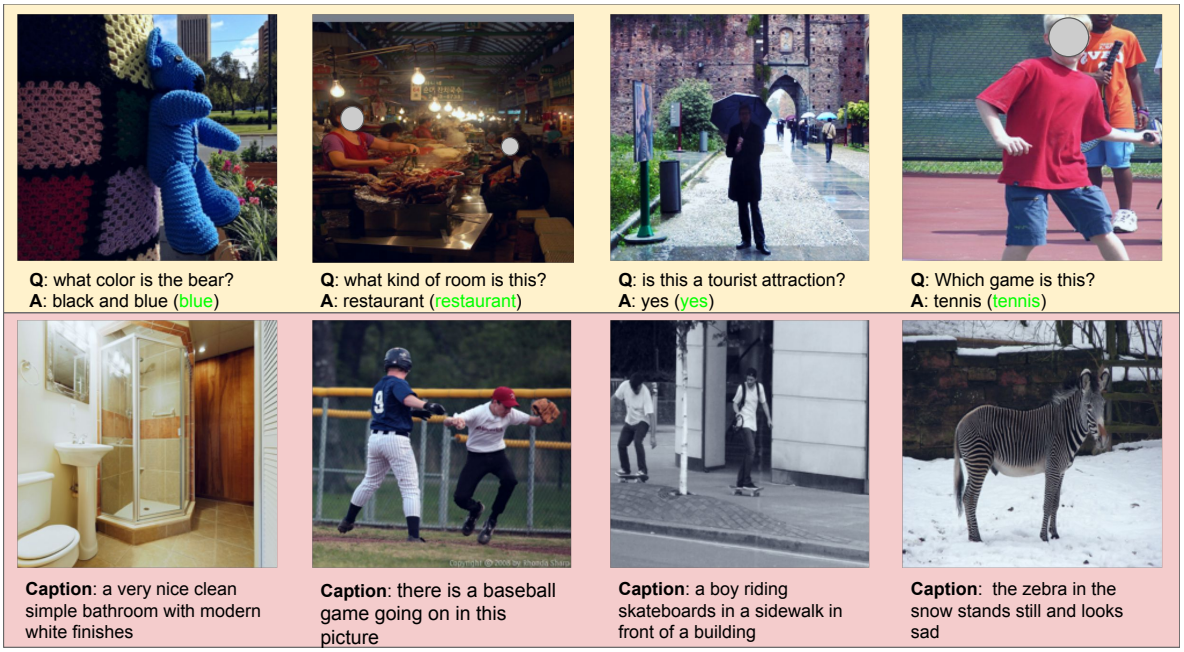

Figure 4: Qualitative results of eP-ALM: the model is able to generate accurate answers and coherent description of the image. Ground truth answers are highlighted in green.