

# Appendix: Towards Multimodal-augmented Pre-trained Language Models via Self-balanced Expectation-Maximization Iteration

Anonymous Authors

## 1 MORE TRAINING DETAILS

The hyperparameter  $\alpha$  is set to 0.5 to achieve the optimal performance in experiments. The temperature coefficient  $\tau$  is set to 0.2. We employ Adam as the optimizer with a weight decay of 0.01 and tune all models for 3 epochs. In this work, we utilize CLIP (ViT-B/32)<sup>1</sup> [3], AudioCLIP (Full-Trained)<sup>2</sup> [2] and CLIP-ViP (base-patch32)<sup>3</sup> [4] text encoders to obtain representations of image, audio, and video modalities, respectively. The hidden sizes for image, audio, and video representations are set to 512, 1024, and 512. In addition, we set the training batch size of 32 on the GLUE benchmark and CommonGen dataset, 8 on the CSQA dataset, and 12 on the SQuAD v2.0 datasets. We use grid search to determine the optimal hyperparameters mentioned above. For VQA tasks, we use our MASE to replace the original PLMs and fine-tune it according to settings similar to ConceptBert [1]. We use Spearman’s correlation as the metric on STS-B and the remaining GLUE tasks using accuracy as the metric. All experiments are conducted on 8 RTX 4090 GPUs.

## 2 THEORETICAL ANALYSIS OF OUR MODAL PROXY

In this section, we analyze why our modal proxies are a good bridge for multimodal information transmission. We analyze from the perspective of information theory that using our multimodal proxies for training is implicitly transmitting information from other modal sources. Maximizing the mutual information between our multimodal proxies and labels is essentially maximizing the mutual information between real modal data and labels. Specifically, we provide the following theorem:

**THEOREM 1.** *Given a multimodal contrastive pre-training model  $\Theta$ , constructed on a large-scale dataset that integrates modalities  $k$  and  $t$ , suppose  $Z_k$  and  $Z_t$  are the embeddings produced by  $\Theta$  for modalities  $k$  and  $t$  respectively, and  $Y$  is a set of labels for a specific task. It is proposed that the optimization of mutual information between  $Z_t$  and  $Y$  inherently optimizes the mutual information between  $Z_k$  and  $Y$ , hence:*

$$\max \mathbf{I}(Z_t, Y) \equiv \max \mathbf{I}(Z_k, Y), \quad (1)$$

where  $\mathbf{I}$  denotes the mutual information.

*Proof:* Given the model  $\Theta$  is trained on a comprehensive dataset comprising multimodal pairs  $(k, t)$ , embeddings  $Z_k$  and  $Z_t$  are expected to encode similar informational contents about the labels  $Y$ . Assuming high alignment between these embeddings due to the shared training objective, we observe that:

$$p(Z_k | Z_t) \approx 1 \quad \text{and} \quad p(Z_t | Z_k) \approx 1, \quad (2)$$

suggesting a nearly deterministic inferential reciprocity between  $Z_k$  and  $Z_t$ . According to the definition of mutual information, we

have:

$$I(Z_k, Y) = \sum_{z_k, y} p(z_k, y) \log \left( \frac{p(z_k, y)}{p(z_k)p(y)} \right); \quad (3)$$
$$I(Z_t, Y) = \sum_{z_t, y} p(z_t, y) \log \left( \frac{p(z_t, y)}{p(z_t)p(y)} \right).$$

To formalize this, we utilize the chain rule of mutual information:

$$I(Z_k; Y|Z_t) = I(Z_k; Y) - I(Z_k; Z_t; Y), \quad (4)$$

where  $I(Z_k; Z_t; Y)$  represents the mutual information among  $Z_k$ ,  $Z_t$ , and  $Y$ . Given the strong mutual information between  $Z_k$  and  $Z_t$ ,  $I(Z_k; Y|Z_t)$  approximates to 0, indicating that:

$$I(Z_k; Y) \approx I(Z_t; Y). \quad (5)$$

Thus, maximizing  $\mathbf{I}(Z_t, Y)$  leads to the maximization of  $\mathbf{I}(Z_k, Y)$  and vice versa, under the assumption of adequate training and representative data modalities.

This theorem implies that, in contexts where embeddings are derived from jointly trained multimodal data via MC-PTMs, optimizing the mutual information for one modality’s embedding about the labels effectively optimizes it for the other. In other words,  $Z_t$  is a well-implicit modal proxies for modality  $k$ . We can utilize  $Z_t$  to efficiently transfer multimodal knowledge into PLMs to augment cognitive processing and understanding.

**Table 1: We use RoBERTa-base as the base model to evaluate the impact of EM iteration times on CSQA and SQuAD v2.**

Iterations steps	CSQA		SQuAD v2	
	Acc.	F1-score	Acc.	F1-score
1	85.5	75.8	80.3	83.6
2	86.7	76.3	81.1	84.4
5	87.2	76.8	81.9	85.0
10	<b>87.7</b>	<b>77.8</b>	<b>82.8</b>	<b>85.7</b>
15	87.5	<b>77.8</b>	82.4	85.6

## 3 EFFECT OF EM ITERATION STEPS

We use RoBERTa-base as the backbone network to evaluate the impact of EM algorithm iterations on the performance of our probabilistic framework. We present the results of quantitative analysis as shown in Table 1. We can observe that when the number of iteration steps is between [1,10], the model performance continuously improves with the increase of iteration steps. This indicates that the bidirectional optimization of modal information injection and information balance estimation is mutually reinforcing. Our iterative algorithm can promote bidirectional optimization and improve the performance of PLMs. In addition, the results in Table 1 also indicate that when the number of iterations is 10, our probability framework can be fully optimized to achieve optimal performance.

<sup>1</sup><https://github.com/openai/CLIP>

<sup>2</sup><https://github.com/AndreyGuzhov/AudioCLIP>

<sup>3</sup><https://github.com/microsoft/XPretrain/tree/main/CLIP-ViP>

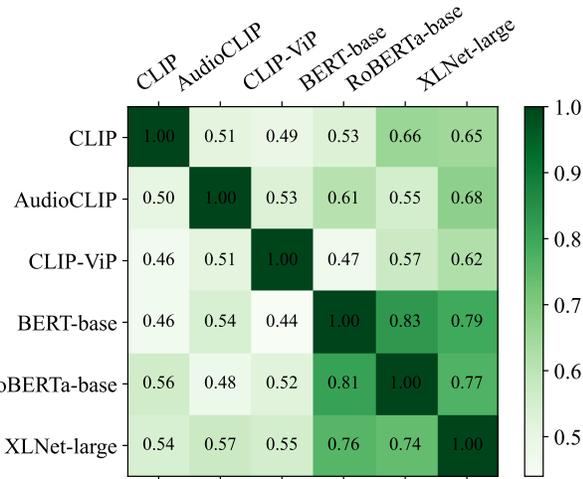


Figure 1: The overlap of correct predictions between each pair of models in the CSQA dataset.

#### 4 DIFFERENT PRE-TRAINED MODELS BEHAVE DIFFERENTLY.

To further demonstrate the effectiveness of our model proxies, we use the overlapping situation correctly predicted by the model to evaluate the output features of different encoders. We mathematically define the concept of overlap in correct predictions between two models  $\mathcal{M}_1$  and  $\mathcal{M}_2$  as:

$$\mathcal{O}(\mathcal{M}_1, \mathcal{M}_2) = \frac{|\mathcal{S}_{\mathcal{M}_1} \cap \mathcal{S}_{\mathcal{M}_2}|}{|\mathcal{S}_{\mathcal{M}_1}|}, \quad (6)$$

where,  $\mathcal{S}_{\mathcal{M}}$  denotes the set of predictions made by model  $\mathcal{M}$ . We obtain the model overlap coefficients  $\mathcal{O}$  among different models (i.e., CLIP, AudioCLIP, CLIP-ViP, BERT, RoBERTa and XLNet) on the CSQA dataset in Figure 1. We can observe a high degree of overlap among PLMs i.e., BERT-base, RoBERTa-base, and XLNet-large), while the overlap between MC-PTM i.e., CLIP, AudioCLIP, and CLIP-ViP) and other models are significantly smaller. This difference empirically explains the significant performance gain obtained by integrating multimodal proxies.

Text Encoder	CSQA		SQuAD v2	
	Acc.	F1	Acc.	F1
Gaussian Noise (0M)	82.1	70.3	76.8	79.6
BERT-base (110M)	84.0	73.9	78.5	82.1
RoBERTa-large (355M)	84.3	74.3	79.0	82.2
T5-3b (1500M)	84.9	74.5	80.2	83.5
CLIP-base (63M)	<b>86.7</b>	<b>76.6</b>	<b>81.8</b>	<b>85.0</b>

Table 2: Performance comparison of different text encoders in our approach on CSQA and SQuAD v2 datasets. We employ RoBERTa-base as the base model. "CLIP-base" represents using the text encoders of CLIP, AudioCLIP and CLIP-ViP.

#### 5 DIFFERENT MULTIMODAL PROXY EXTRACTION METHODS

To investigate the effectiveness of our multimodal semantic proxies, we use PLMs trained on pure text corpus instead of encoders in the MC-PTMs (i.e., CLIP-based models). We utilize BERT-base, RoBERTa-large, and T5-3b, along with random Gaussian noise as alternatives to CLIP-based text encoders to conduct experiments for assessing the significance of multimodal representation. The results are shown in Table 2, indicating that our method achieved the most significant performance gain. This experiment indicates that the performance gain brought by our method is not only related to the text information provided by the text encoder but also to the implicit multimodal information provided by MC-PTMs. That is to say, the semantics we inject into PLMs do contain additional modal information that text features do not possess. In addition, this also demonstrates the significant role of additional modal knowledge in enhancing the expression and reasoning capabilities of PLM.

Base Model	Methods	Param.	Latency	Speedup
BERT-base	+None	110M	13.1ms	1×
	+MPB	118M	18.3ms	1.4×
	+VOKEN	121M	22.6ms	1.7×
	+iACE	568M	42.9ms	3.3×
	+MASE	135M	23.9ms	1.8×
RoBERTa-base	+None	355M	24.1ms	1x
	+MPB	363M	28.9ms	1.2×
	+VOKEN	367M	32.7ms	1.4×
	+iACE	738M	90.8ms	3.8×
	+MASE	383M	34.2ms	1.4×

Table 3: Runtime analysis. We calculate the model parameter size and average inference time (latency) for each sample using different methods on the SST-2 dataset.

#### 6 RUNTIME ANALYSIS

To further analyze the runtime efficiency of our method, we conduct runtime analysis on different methods with the results shown in Table 3. It can be observed that our method only introduces subtle computational overhead compared to the baseline method. Note that our multimodal proxy extraction module is frozen at runtime, therefore no additional learnable parameters will be introduced. Furthermore, although our method involves iterative optimization, the encoders of PLMs only need to extract multimodal features once (referring to the algorithm in the manuscript).

#### 7 MORE EXPERIMENTS ON DIFFERENT BASE MODELS.

We use XLNet as the baseline model to further test our method, and the results are shown in Table 4. It can be seen that our method still exhibits excellent performance on larger base models.

233  
234  
235  
236  
237  
238  
239  
240  
241  
242  
243  
244  
245  
246  
247  
248  
249  
250  
251  
252  
253  
254  
255  
256  
257  
258  
259  
260  
261  
262  
263  
264  
265  
266  
267  
268  
269  
270  
271  
272  
273  
274  
275  
276  
277  
278  
279  
280  
281  
282  
283  
284  
285  
286  
287  
288  
289  
290

291  
292  
293  
294  
295  
296  
297  
298  
299  
300  
301  
302  
303  
304  
305  
306  
307  
308  
309  
310  
311  
312  
313  
314  
315  
316  
317  
318  
319  
320  
321  
322  
323  
324  
325  
326  
327  
328  
329  
330  
331  
332  
333  
334  
335  
336  
337  
338  
339  
340  
341  
342  
343  
344  
345  
346  
347  
348



**Question:** How many giraffes are in the photo?  
**MPB:** four  
**MASE:** one



**Question:** What color is the window frame?  
**MPB:** red  
**MASE:** black



**Question:** What action are these two doing?  
**MPB:** running  
**MASE:** brushing teeth



**Question:** Is the baby holding a toothbrush?  
**MPB:** no  
**MASE:** yes



**Question:** What is the young man carrying?  
**MPB:** bag  
**MASE:** duffle bag



**Question:** What is behind and to the right of the bench?  
**MPB:** dog  
**MASE:** tree

Figure 2: Visualize evaluation results on the VQA 2.0 validation set.

Base Model	Methods	Modality	Average
XLNet-large	+None	T	85.22
	+MPB	T+I+A+V	87.39
	+MASE	T+I	89.25
	+MASE	T+I+A	89.93
	+MASE	T+I+A+V	<b>90.61</b>

Table 4: Experiments on different base models (i.e., XLNet) on the GLUE benchmark.

### 8 CASE STUDY OF OUR MASE

We present a visualization example of our method and baseline MPB in a cross-modal QA task (i.e., VQA 2.0 benchmark) in Figure 2. Furthermore, we provide some qualitative analysis examples for the QQP, QNLI, CSQA and SQuADv2 datasets as shown in Table 5. The effectiveness of our method in NLU and QA tasks can be intuitively seen in the following examples.

### REFERENCES

- [1] François Gardères, Maryam Ziaefard, Baptiste Abeloos, and Freddy Lécué. 2020. ConceptBert: Concept-Aware Representation for Visual Question Answering. In *Findings*. <https://api.semanticscholar.org/CorpusID:226284018>
- [2] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas R. Dengel. 2021. Audioclip: Extending Clip to Image, Text and Audio. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2021)*, 976–980.
- [3] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*. PMLR, 8748–8763.
- [4] Hongwei Xue, Yuchong Sun, Bei Liu, Jianlong Fu, Rui Song, Houqiang Li, and Jiebo Luo. 2022. CLIP-ViP: Adapting Pre-trained Image-Text Model to Video-Language Alignment. In *International Conference on Learning Representations*.

QQP	Sentence 1	Why is the EPA held in such low esteem by a large proportion of Libertarians and the Right?
	Sentence 2	Why is the EPA (Environmental Protection Agency) held in such low esteem by the American Conservatives?
	Ground Truth	(1) Equivalent
	BERT-base	(0) Not Equivalent
	BERT-base + MASE	(1) Equivalent
QNLI	Sentence 1	What area in modern-day Canada received Huguenot immigrants?
	Sentence 2	They also spread beyond Europe to the Dutch Cape Colony in South Africa, the Dutch East Indies, the Caribbean, and several of the English colonies of North America, and Quebec, where they were accepted and allowed to worship freely.
	Ground Truth	entailment
	BERT-base	not_entailment
	BERT-base + MASE	entailment
CSQA	Question	Janet was watching the film because she liked what? (A) erection (B) laughter (C) being entertained (D) fear (E) bordem
	Ground Truth	C
	BERT-base	B
	BERT-base + MASE	C
SQuAD v2	Question	What is the area called where two plates move apart?
	Ground Truth	"answers": [{"text": "divergent boundaries", "answer_start": 295}]
	BERT-base	"answers": [{"text": "asthenosphere", "121"}]
	BERT-base + MASE	"answers": [{"text": "divergent boundaries", "answer_start": 295}]

**Table 5: Case studies of our method on different datasets.**