Foundation Models for History Compression in Reinforcement Learning

Anonymous Author(s)
Affiliation
Address
email

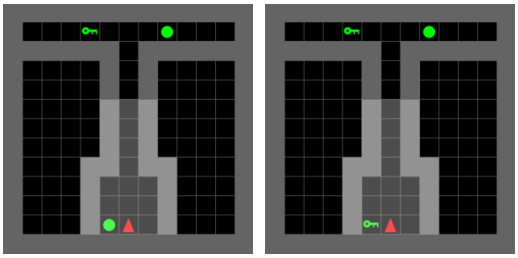
Abstract

Agents interacting under partial observability require access to past observations via a memory mechanism in order to approximate the true state of the environment. Recent work suggests that leveraging language as abstraction provides benefits for creating a representation of past events. History Compression via Language Models (HELM) leverages a pretrained Language Model (LM) for representing the past. It relies on a randomized attention mechanism to translate environment observations to token embeddings. In this work, we show that the representations resulting from this attention mechanism can collapse under certain conditions. This causes blindness of the agent to subtle changes in the environment that may be crucial in solving a certain task. We propose a solution to this problem consisting of two parts. First, we improve upon HELM by substituting the attention mechanism with a feature-wise centering-and-scaling operation. Second, we take a step toward semantic history compression by leveraging foundation models, such as CLIP, to encode observations, which further improves performance. By combining foundation models, our agent is able to solve the challenging MiniGrid-Memory environment. Surprisingly, however, our experiments suggest that this is not due to the semantic enrichment of the representation presented to the LM, but rather due to the discriminative power provided by CLIP.

1 Introduction

In Reinforcement Learning (RL) an agent interacts with an environment and learns from feedback provided in the form of a reward function. RL agents that are deployed in the real world often have to cope with partial observability. Therefore, the capability to approximate the true state of their surrounding environment by virtue of an agent's perception is crucial (Åström, 1964; Kaelbling et al., 1998). To this end, many agents employ a memory mechanism to track events that occurred in the past. In this memory, it is much more efficient to store abstract representations of the past rather than every detail the agent encountered. Thus, memory mechanisms such as LSTM (Hochreiter & Schmidhuber, 1997) or Transformer (Vaswani et al., 2017) compress sequences of high-dimensional observations.

We propose to use foundation models (FM; Bommasani et al., 2021) as memory mechanism. Since FMs come pretrained, they allow for a much higher sample efficiency than memory mechanisms trained from scratch. Moreover, FMs have demonstrated remarkable few-shot capabilities (Brown et al., 2020) and can learn abstract symbolic rules and perform reasoning (Petroni et al., 2019; Talmor et al., 2020; Kassner et al., 2020). In this work, we aim to leverage these properties, which FMs learned on large-scale text corpora, for RL. We believe that language is well suited as domain for a memory mechanism, since it appears to be an essential component in forming meaningful abstractions. In humans, the ability to abstract is heavily influenced by the exposure to language in early childhood (Waxman & Markow, 1995). Recently, there has been a surge of complex models that combine



β	Hellinger distance δ
1	0.00036 ± 1.88e-9
10	0.0036 ± 1.84e-7
100	0.035 ± 2.09e-5
1000	0.091 ± 0.01
1e4	0.093 ± 0.06

Figure 1: **Left:** Instances of memory environment. The agent only observes the white shaded region, and must navigate to the object seen in the starting room. **Right:** Hellinger distance (Hellinger, 1909) of softmax distributions over token embeddings for the two observations of the Memory environment on the left. Mean over 1000 different initializations of $\boldsymbol{P}$ of FH is shown. For very high values of $\beta > 1$e4, the softmax converges to a one-hot encoding and δ directly corresponds to the probability of mapping to a single, but different token embedding. Therefore, with higher values of β, there is only a 10% chance of avoiding collapse to the same token embedding.

multiple modalities, in particular images and text, as in CLIP (Radford et al., 2021). Since many RL environments rely on visual inputs, we use CLIP in our proposed memory mechanism to map visual inputs meaningfully into the language domain.

Prior work has illustrated that foundation models pretrained on language can efficiently compress sequences of observations and facilitate agent learning in partially observable RL environments (HELM, Paischer et al., 2022). The key challenge hereby is to map image-based observations to language representations. HELM tackles this problem using a randomly chosen mapping called FrozenHopfield (FH) that is — due to its randomness — inherently unable to form meaningful abstractions. In fact, we show that under certain conditions, the representations are prone to collapse rendering the agent unable to distinguish between different inputs (see Figure Fig. 1). We build upon HELM by (i) substituting FH with a feature-wise centering-and-scaling operation and (ii) incorporating a CLIP image encoder (Radford et al., 2021), which is pretrained in a multimodal fashion on web data consisting of images and text. We term the resulting new method HELMv2. Using HELMv2, we are able to distinguish even minute differences in the input when necessary, which drastically enhances downstream performance.

We demonstrate the effectiveness of HELMv2 on a diverse set of partially observable environments. Concretely, we train on 2D MiniGrid (Chevalier-Boisvert et al., 2018), and 3D MiniWorld environments (Chevalier-Boisvert, 2018). HELMv2 yields significant improvements over HELM in all environments. Further, we conduct ablation studies, which show that the improvements are not only due to the CLIP image encoder but also due to the replacement of the FH. Finally, we construct a mapping that successfully conveys the semantics extracted by CLIP to the LM. Surprisingly, however, using this mapping does not further improve the results in the selected environments.

2 Methods

HELM has demonstrated that pretrained LMs are well suited for compressing past observations that are randomly mapped to language tokens, which is realized by FH. The FH mechanism consists of a random matrix $\boldsymbol{P}$ and an attention mechanism over pretrained token embeddings (VocabAttn). More formally, let $\boldsymbol{E} = (\boldsymbol{e}_1, \dots, \boldsymbol{e}_k)^\top \in \mathbb{R}^{k \times m}$ be the token embedding matrix of the pretrained LM consisting of k embeddings $\boldsymbol{e}_i \in \mathbb{R}^m$. At every timestep t, we obtain inputs $\boldsymbol{x}_t \in \mathbb{R}^m$ for the LM from observations $\boldsymbol{o}_t \in \mathbb{R}^n$ via the FH mechanism by

$$\boldsymbol{x}_t^\top = \sigma(\beta \boldsymbol{o}_t^\top \boldsymbol{P}^\top \boldsymbol{E}^\top)\boldsymbol{E}, \tag{1}$$

where σ is the softmax function and $\boldsymbol{P} \in \mathbb{R}^{m \times n}$ has entries sampled from $\mathcal{N}(0, n/m)$. The resulting $\boldsymbol{x}_t$ lies in the convex hull of the token embeddings of the LM. The parameter β is a scaling factor that controls the dispersion of $\boldsymbol{x}_t$ within that convex hull.

HELMcs Our aim is to avoid representation collapse caused by the FH mechanism. We illustrate in Fig. 1 that FH is prone to representation collapse even for higher values of β, especially if observations

tend to be visually similar. To sidestep this issue, we substitute the attention mechanism in FH with a feature-wise centering-and-scaling operation. Let $\mathcal{B}$ be a batch of observations $\boldsymbol{o}_t$ and let $\boldsymbol{\mu}_{\mathcal{B}} \in \mathbb{R}^n$ be its mean feature vector and $\boldsymbol{\sigma}_{\mathcal{B}} \in \mathbb{R}^n$ the vector of standard deviations both taken over $\mathcal{B}$. Likewise let $\boldsymbol{\mu}_E \in \mathbb{R}^m$ and $\boldsymbol{\sigma}_E \in \mathbb{R}^m$ be means and standard deviations of the token embeddings $\boldsymbol{E}$. Then we compute $\boldsymbol{x}_t$ as

$$\boldsymbol{x}_t = \operatorname{diag}(\boldsymbol{\sigma}_E)\boldsymbol{P}\operatorname{diag}(\boldsymbol{\sigma}_{\mathcal{B}})^{-1}(\boldsymbol{o}_t - \boldsymbol{\mu}_{\mathcal{B}}) + \boldsymbol{\mu}_E, \tag{2}$$

where diag takes a vector and constructs a diagonal matrix from it. We refer to this setting as HELMcs.

HELMv2 We obtain another setting by encoding the observations with the ResNet-50 CLIP image encoder, which has the same output dimension as the embeddings $\boldsymbol{e}_i$ used by TransformerXL (TrXL, Dai et al., 2019), which we use as LM, thus eliminating the need for the mapping $\boldsymbol{P}$. That is, we let $\boldsymbol{z}_t = \mathrm{CLIP}(\boldsymbol{o}_t) \in \mathbb{R}^m$ and construct $\mathcal{B}$ from $\boldsymbol{z}_t$. Consequently, $\boldsymbol{\mu}_{\mathcal{B}} \in \mathbb{R}^m$ and $\boldsymbol{\sigma}_{\mathcal{B}} \in \mathbb{R}^m$. We compute the LM inputs as

$$\boldsymbol{x}_t = \operatorname{diag}(\boldsymbol{\sigma}_E)\operatorname{diag}(\boldsymbol{\sigma}_{\mathcal{B}})^{-1}(\boldsymbol{z}_t - \boldsymbol{\mu}_{\mathcal{B}}) + \boldsymbol{\mu}_E. \tag{3}$$

3 Experimental Results

We investigate the limitations of HELM on the MiniGrid-MemoryS11-v0 environment (Memory, Fig. 1, left). The task for the agent is to remember the object in the room it spawns in, after navigating through a corridor. The corridor ends in a T-junction showing two objects in each direction, one of which is equivalent to the object in the starting room. The agent then has to choose the direction towards the object in the starting room. If the agent chooses the wrong direction the episode ends yielding no reward. However, the objects only slightly differ in shape (green ball vs green key). We demonstrate that FH collapses to the same representation for both objects by measuring the distances between the softmax distribution $\sigma(\cdot)$ over token embeddings (Fig. 1, right). Indeed, we observe that even for high values of β there is very little chance that HELM can discriminate between the two objects. Moreover, we take a closer look at the resulting representations for the two different observations and measure their cosine similaritiy and ℓ_2-distance. Compared to HELM and HELMcs, HELMv2 improves separability of the two observations by a large margin (see Table 1). Based on these findings we add a higher value of $\beta = 1000$ to the gridsearch for HELM. Due to their limitations, HELM, and HELMcs do not achieve better performance than randomly choosing a path at the end of the corridor (see Fig. 2). Contrary, HELMv2 is able to distinguish and memorize the objects and solve this task.

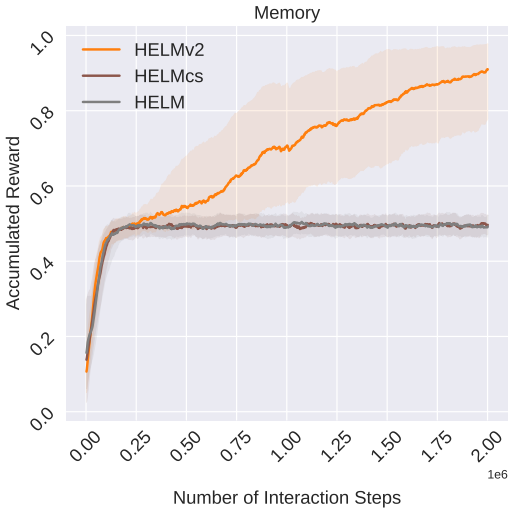


Figure 2: IQM and 95% bootstrapped CIs across 15 seeds for Memory.

Additionally, we compare HELMv2 to HELM on a set of six diverse partially observable environments. Particularly, we select the same MiniGrid environments as Paischer et al. (2022). Moreover, we

Method	Cosine Similarity (↓)	ℓ_2-distances (↑)
HELM, $\beta = 1$	0.99 ± 1.68e-7	0.04 ± 0.01
HELM, $\beta = 10$	0.99 ± 5.65e-6	0.36 ± 0.12
HELM, $\beta = 100$	0.99 ± 8.2e-4	4.76 ± 2.92
HELM, $\beta = 1000$	0.97 ± 0.08	47.49 ± 88.33
HELMcs (ours)	0.99 ± 3.65e-5	11.04 ± 0.24
HELMv2 (ours)	$\mathbf{0.83 \pm 0.04}$	$\mathbf{296.18 \pm 61.55}$

Table 1: Average cosine similarities and ℓ_2-distances for HELM, HELMcs, and HELMv2. HELMv2 can better discriminate between the two different objects.

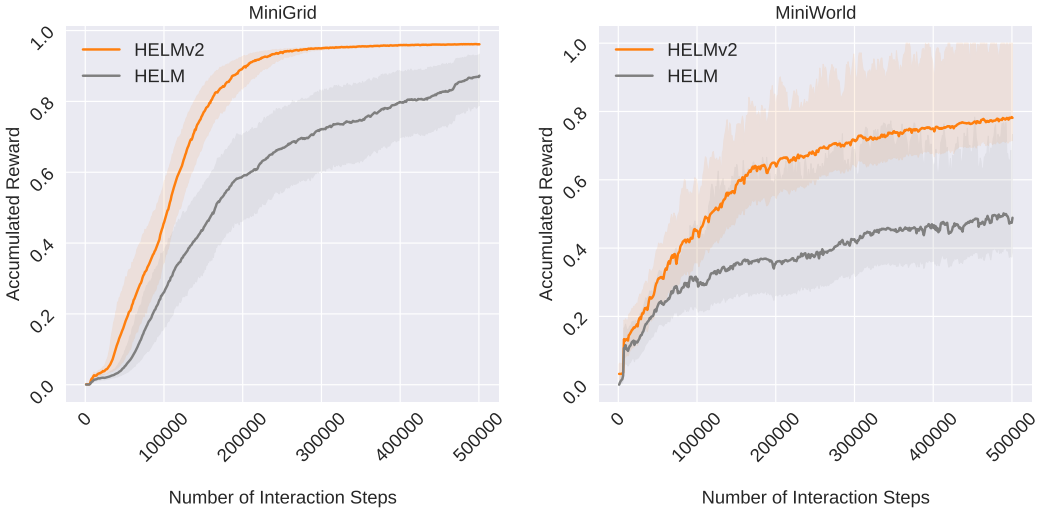


Figure 3: 30 seeds for 6 MiniGrid environments (**left**), and 8 more 3D MiniWorld environments (**right**). HELMv2 consistently outperforms HELM.

add eight 3D environments from the MiniWorld benchmark suite (Chevalier-Boisvert, 2018). For more details about our selected MiniWorld environments see Appendix A.1. Results are shown in Fig. 3 for MiniGrid (left) and MiniWorld (right). HELMv2 significantly outperforms HELM on both benchmark suites ($p = 4.59$e-19, $p = 2.25$e-5, for MiniGrid, and MiniWorld, respectively).

Due to the drastic improvement of HELMv2 over HELM we conduct ablation studies to isolate the effect of its components. In this regard, we add an additional setting, namely HELMclip. HELMclip uses the same ResNet-50 CLIP image encoder instead of the random projection while keeping the VocabAttn. We show the performance of HELMclip compared to all other settings on one MiniGrid (RedBlueDoors), and one MiniWorld (TMaze) environment in Fig. 4. We find that HELMclip improves upon HELMcs, which indicates that the main factor for decreased performance is the random projection matrix in HELM. Further, substituting VocabAttn with our feature-wise centering and scaling operation enhances performance even more.

Moreover, we perform an ablation where we substitute the CLIP image encoder with a ResNet (He et al., 2016, HELMv2-RN34-IN), and a Vision Transformer (Dosovitskiy et al., 2021, HELMv2-ViT-L16-IN), pretrained in a supervised manner on the popular Imagenet dataset (Deng et al., 2009). Additionally, we compare to a Vision Transformer version of CLIP (HELMv2-ViT-B/16). First, we again measure ℓ_2-distances and cosine similarities to quantify the ability of the different vision encoders to separate the two observations (see Table 2). The HELMv2-ViT-B/16 variant exhibits the lowest cosine similarity between both observations. Next, we investigate whether the improved separability of the two objects correlates with downstream performance. The learning curves for the Memory environment are shown in Fig. 5. There is no significant difference in the final performance

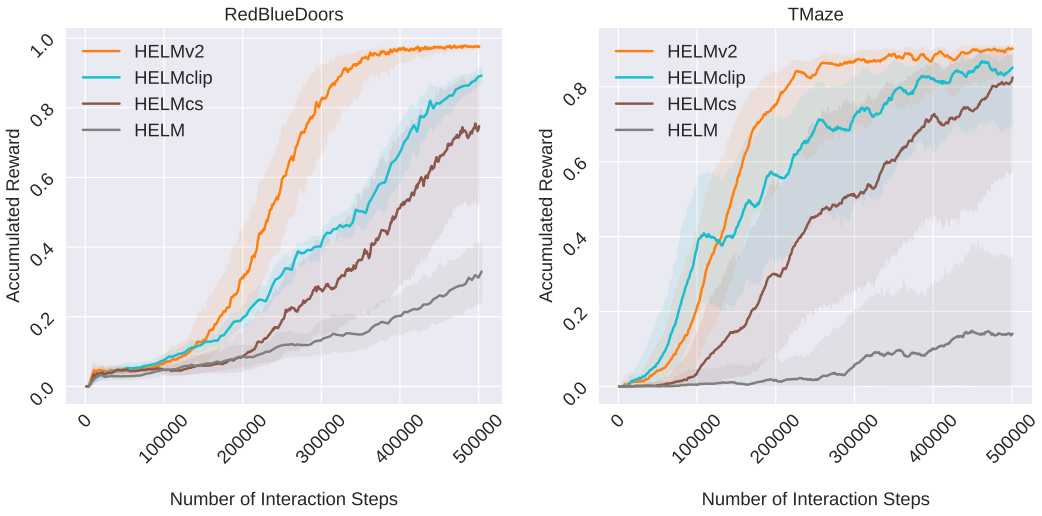


Figure 4: Mean IQM and 95% bootstrapped CIs across 30 seeds for RedBlueDoors (left), and TMaze (right) environments. HELMv2 consistently outperforms HELMclip on both environments.

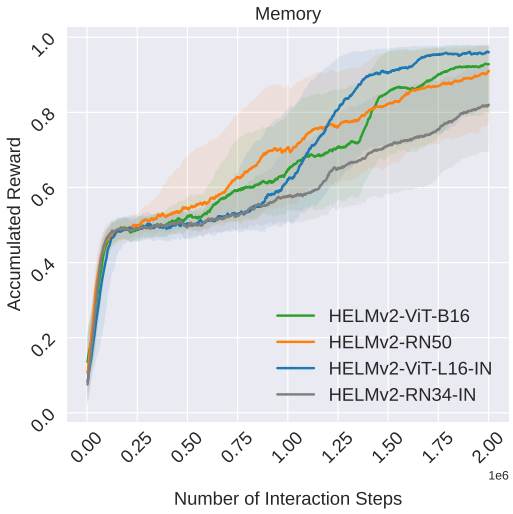


Figure 5: IQM and 95% bootstrapped CIs across 10 seeds for Memory environment.

of the different methods. It seems that for the Memory task it is important to distinguish the objects, but better separation does not lead to better performance.

Method	Cosine Similarity (↓)	ℓ_2-distances (↑)
HELMv2-RN50	0.83 ± 0.04	$\mathbf{296.18} \pm \mathbf{61.55}$
HELMv2-ViT-B/16	$\mathbf{0.73} \pm \mathbf{0.06}$	245.9 ± 49.28
HELMv2-ViT-IN	0.83 ± 0.04	280 ± 60.06
HELMv2-RN34-IN	0.76 ± 0.04	213.75 ± 45.8

Table 2: Cosine Similarity and ℓ_2-distances for different vision encoders in HELMv2.

All our methods are trained with Proximal Policy Optimization (PPO, Schulman et al., 2017). We evaluate all methods via the interquartile mean (IQM) and 95% bootstrapped confidence intervals (CIs) as suggested by Agarwal et al. (2021). To test for statistical significance, we perform a Wilcoxon

test (Wilcoxon, 1945) at the end of training. We follow the architectural design of Paischer et al. (2022) and perform the same hyperparameter search across all environments.

4 Investigation on Semantic History Compression

We want to pave the way toward a semantic history compression for RL, that is, to preserve the semantic concepts of an observation when mapping it into the LM space. In this regard, we provide compelling evidence that it is indeed feasible to create such a semantic mapping between CLIP and the LM (see Fig. 6). When using this semantic mapping in the two selected minimalistic RL environments, however, we do not observe a significant improvement over HELMv2.

We can compute such a mapping from the vocabularies of the CLIP language encoder $\mathcal{V}_{\text{CLIP}}$ and the LM encoder $\mathcal{V}_{\text{LM}}$ alone. First, we identify the overlap between the two vocabularies $\mathcal{V}_{OV} = \mathcal{V}_{\text{CLIP}} \cap \mathcal{V}_{\text{LM}}$ with size $l = |\mathcal{V}_{OV}|$. Next, we embed $\mathcal{V}_{OV}$ in the CLIP output space and the LM input space yielding embedding matrices $\boldsymbol{F} = (\boldsymbol{f}_1, \dots, \boldsymbol{f}_l) \in \mathbb{R}^{n \times l}$ and $\boldsymbol{E}_l \in \mathbb{R}^{m \times l}$, respectively. The mapping matrix $\boldsymbol{W}$ can be computed using any linear model. A common choice for creating a mapping between monolingual embedding spaces is the Procrustes method (Artetxe et al., 2018; Hoshen & Wolf, 2018; Smith et al., 2017; Lample et al., 2018; Zhang et al., 2016; Xing et al., 2015; Minixhofer et al., 2022). Algorithm 1 outlines an algorithm for computing such a semantic mapping. We refer to this setting as SHELM (for semantic HELM).

Algorithm 1 Associating CLIP output space with LM input space

Require: CLIP language encoder CLIP_{LM}, Vocabulary of CLIP $\mathcal{V}_{\text{CLIP}}$, Language Model Embedding Layer $\text{LM}(v_i)$, Language Model vocabulary $\mathcal{V}_{\text{LM}}$

$\mathcal{V}_{\text{OV}} \leftarrow \mathcal{V}_{\text{CLIP}} \cap \mathcal{V}_{\text{LM}}$ ▷ Search for overlapping vocabulary

$\boldsymbol{f}_i = \text{CLIP}_{LM}(v_i) \quad \text{for} \quad v_i \in \mathcal{V}_{\text{OV}}$ ▷ Embed tokens in CLIP output space

$\boldsymbol{e}_i = \text{LM}(v_i) \quad \text{for} \quad v_i \in \mathcal{V}_{\text{OV}}$ ▷ Embed tokens in LM input space

$\boldsymbol{W} \leftarrow \texttt{create_mapping}(\boldsymbol{F}, \boldsymbol{E})$ ▷ Compute mapping between embeddings

We compare four different linear mappings:

- **Linear:** Ordinary Least Squares, as in (Mikolov et al., 2013)
- **Ridge:** Least Squares with Thikonov regularization
- **Procrustes:** Least Squares with orthogonality constraint Schönemann (1966); Gower & Dijksterhuis (2005))
- **RobProc:** The Robust Procrustes algorithm, that iteratively refines the Procrustes method based on its predicion error $\|\boldsymbol{F}\boldsymbol{W} - \boldsymbol{E}\|_F^2$[1].

We perform a 5-fold cross validation and measure the accuracy, considering each token as its own class. Table 3 shows the average train and test accuracy for the different linear mapping methods for the RN50 CLIP backbone and TrXL. We observe a strong overfitting effect for Linear, while the two Procrustes generalize best. We show results for the other mapping methods in Table 6.

	Linear	Ridge	Procrustes	Robust Procrustes
RN50	0.786/0.0013	0.571/0.281	0.64/**0.304**	0.65/0.303

Table 3: Train/Test accuracy for different linear models optimized for mapping CLIP tokens to the TrXL embedding space. Average over 5-fold cross validation is shown. The Procrustes method yields the best mapping between the two embedding spaces.

Due to the alignment of the CLIP image encoder and the CLIP language encoder, this mapping can be used to project images to the LM space while preserving its semantics. We visualize sample images and the closest tokens in the LM space in Fig. 6. Further, we conduct a quantitative analysis using

[1]For more details, we refer the reader to Groenen et al. (2005)

publicly available image captioning datasets. In this regard, we use a pre-existing image/caption dataset to quantify how well the semantics of an image are preserved after mapping to the LM space. We draw a random subset of 1000 image-caption pairs of the popular MSCOCO dataset (Lin et al., 2014). The subset is filtered to contain only image-text pairs where the captions contain at least 5 tokens of our computed vocabulary overlap above. For preprocessing of the captions we remove stop words and apply stemming. Next, we propagate the image through various CLIP backbones, to obtain an image embedding and map it to the LM space using our pre-computed mappings. Finally, we rank tokens in the LM space based on their cosine similarity to the mapped image. Based on the obtained ranking we compute the Mean Reciprocal Rank (MRR, Craswell, 2009) and the Normalized Discounted Cumulative Gain (NDCG, Järvelin & Kekäläinen, 2002). We also include a comparison to token ranking in the CLIP space as a baseline to determine how much information is lost after applying our mapping. Table 4 shows the NDCG for the RN50 CLIP backbone, mapping to the embedding space of TrXL. We show more results for the different available vision backbones of CLIP for NDCG and MRR in Table 7, and Table 8 in the appendix, respectively.

	Linear	Ridge	Procrustes	RobProc	CLIP
RN50	0.252±0.038	0.285±0.062	0.325±0.082	**0.326±0.082**	0.421±0.099

Table 4: Normalized Discounted Cumulative Gain (NDCG) for ranked tokens in the LM embedding space given an image as input and applying our mapping. Image-caption pairs are drawn from the MSCOCO dataset. Ranking tokens in the CLIP space provides an upper bound.

Fig. 6 shows promising results for natural images from the MSCOCO dataset and their corresponding token rankings. Compared to all other methods, SHELM can better identify tokens prevalent in captions provided by humans (see Table 5).

	MRR (↑)	NDCG (↑)
Random	0.016±0.058	0.243±0.029
HELMclip, $\beta = 1$	9.1e-4±9.2e-4	0.221±0.023
HELMclip, $\beta = 10$	9.4e-4±9.5e-4	0.221±0.023
HELMclip, $\beta = 100$	0.017±0.061	0.242±0.034
HELMclip, $\beta = 1e3$	0.018±0.056	0.246±0.034
HELMclip, $\beta = 1e4$	0.018±0.056	0.246±0.035
HELMv2	0.018±0.057	0.244±0.031
HELMv2+RandOrtho	0.019±0.068	0.245±0.033
SHELM+VocabAttn, $\beta = 1$	9.1e-4±9.2e-4	0.221±0.023
SHELM+VocabAttn, $\beta = 10$	0.001±0.001	0.222±0.023
SHELM+VocabAttn, $\beta = 100$	0.228±0.357	0.303±0.086
SHELM+VocabAttn, $\beta = 1e3$	0.217±0.368	0.297±0.083
SHELM+VocabAttn, $\beta = 1e4$	0.209±0.364	0.295±0.082
SHELM	**0.264±0.353**	**0.326±0.082**

Table 5: Mean Reciprocal Rank (MRR) and Normalized Discounted Cumulative Gain (NDCG) for tokens ranked in the LM embedding space for different ablation setups.

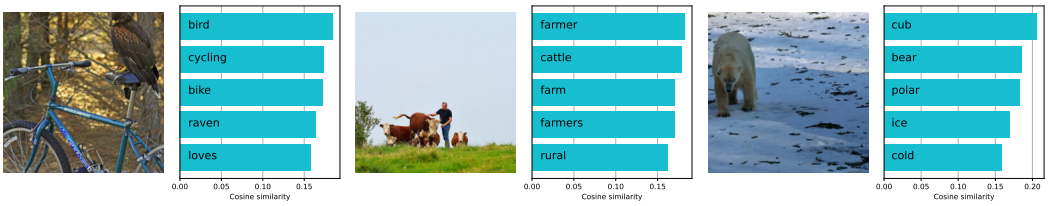


Figure 6: Top 5 closest tokens in the LM embedding space after applying the linear mapping to an image embedded by CLIP.

Figure 7: IQM and 95% bootstrapped CIs across 30 seeds on Memory (**left**), RedBlueDoors (**middle**), and TMaze (**right**) environments. HELMv2+RandOrtho performs on-par with HELMv2 and SHELM on RedBlueDoors and TMaze.

We show preliminary results of SHELM on the RedBlueDoors and the TMaze environments in Fig. 7, where SHELM yields performance on-par with HELMv2. Additionally, on the RedBlueDoors and TMaze environments, we compare with HELMv2+RandOrtho, which samples random orthogonal matrices from the Haar distribution (Stewart, 1980). Surprisingly, we observe no difference in performance between the different methods. We believe this is due to the fact that our selected environments are artificial 2D and 3D scenes for which the CLIP image encoder is unable to sufficiently extract semantically meaningful features. This is in line with findings of Fan et al. (2022) who minimally finetune CLIP for the Minedojo environment. This might change in real world environments with observations close to the CLIP pre-training domain. We aim to investigate this question in future work.

Finally, we conduct an additional analysis on how much the different methods preserve the semantics of natural images. In this regard, we perform the same quantitative analysis as above for the different methods presented in this work. We show the MRR and NDCG for our MSCOCO subset of image-caption pairs in Table 5. Methods such as HELMv2, HELMclip, or HELMv2+RandOrtho yield on-par or worse results compared to a random ranking. This is due to the fact, that for lower values of β the VocabAttn tends to output the mean over token embeddings. In this case, the resulting ranking is equal across images, and thus, results in lower MRR and NDCG. Also, substituting the RobProc mapping with a random mapping drawn from the Haar distribution (HELMv2+RandOrtho) leads to random rankings. Contrary, SHELM preserves more of the semantics resulting in much better rankings. Adding VocabAttn after the RobProc mapping again results in loss of information and a worse ranking.

5 Related Work

Reinforcement Learning with incomplete state information necessitates compression of past events. A variety of prior works has used History Compression to tackle Credit Assignment (Arjona-Medina et al., 2019; Patil et al., 2022; Widrich et al., 2021; Holzleitner et al., 2021), and partial observability (Hausknecht & Stone, 2015; Vinyals et al., 2019; Berner et al., 2019; Pleines et al., 2022). The question of what information to store given a stream of observations was investigated in prior works (Schmidhuber, 1992; Zenke et al., 2017; Kirkpatrick et al., 2016; Schwarz et al., 2018; Ruvolo & Eaton, 2013). Typical choices for network architectures that are capable of compressing a stream of data are LSTM (Hochreiter & Schmidhuber, 1997), and Transformer (Vaswani et al., 2017).

Foundation models (Bommasani et al., 2021), such as GPT-3 (Brown et al., 2020), demonstrated remarkable few-shot capabilities. Pretraining on large-scale text corpora enables such complex models to abstract relevant knowledge for solving math-word problems (Griffith & Kalita, 2019), symbolic math problems (Noorbakhsh et al., 2021), and even university-level math exercises (Drori et al., 2021). As shown by Petroni et al. (2019); Talmor et al. (2020); Kassner et al. (2020), pretrained LMs can learn abstract symbolic rules and perform reasoning. Recently, there has been a surge of complex models trained on large scale data that combines multiple modalities, such as images and text, as in CLIP (Radford et al., 2021). Moreover, vision FMs have been demonstrated to be well adaptable to foreign domains (Adler et al., 2020; Evci et al., 2022).

The Transformer architecture has blossomed in the realm of Natural Language Processing. Many works have focused on scaling pretraining of Transformers to create LMs that generate realistic human-like text (Devlin et al., 2019; Radford & Narasimhan, 2018; Radford et al., 2019; Brown et al., 2020; Thoppilan et al., 2022; Rae et al., 2021; Zhang et al., 2022; Hoffmann et al., 2022; Chowdhery et al., 2022; Dai et al., 2019). More recently, interest has sparked in leveraging the Transformer for decision making (Parisotto et al., 2020; Sukhbaatar et al., 2021; Chen et al., 2021; Janner et al., 2021; Zheng et al., 2022; Melo, 2022). Moreover, it has been found that pretrained LMs are very well suited for creating abstractions of a sequence of visual observations (Paischer et al., 2022).

Language provides useful abstractions for Reinforcement Learning. Hill et al. (2021) illustrate that language abstraction enables compositional generalization and intrinsic motivation in embodied environments. LMs have been used for exploration in text-based environments (Yao et al., 2020), or in visual environments that provide a language oracle (Mu et al., 2022). Multimodal models pretrained with language supervision provide abstract embedding spaces that can be used for visual exploration (Tam et al., 2022). Further, language pretraining has been leveraged to initialize policies in text-based environments (Li et al., 2022), and for sequence modeling in the offline RL setup (Reid et al., 2022). Moreover, language has been used for reward shaping (Wang et al., 2019; Bahdanau et al., 2019; Goyal et al., 2019). We leverage pretrained multimodal models and map them into a semantically meaningful region of the embedding space of a LM.

A plethora of works have focused on conditioning LMs on the popular CLIP model. Alayrac et al. (2022) use cross-attention combined with the Perceiver architecture (Jaegle et al., 2021) trained on vast web data to combine a frozen pretrained CLIP (Radford et al., 2021) with a generative language model. Yu et al. (2022) adds a contrastive loss between CLIP and a LM encoder for image captioning. Similarly, (Mokady et al., 2021) trains a mapping model from the CLIP output space to a LM input space for image captioning. In contrast, (Su et al., 2022) proposes to condition the decoding step of a finetuned LM on image captioning on a frozen CLIP model. Finally, (Tewel et al., 2021) optimizes the context cache during inference to mazimize a similarity score computed with CLIP for zero-shot image captioning. In contrast to these methods, our mapping can be efficiently computed offline, and allows conditioning a LM on visual information without any additional optimization procedure. Zeng et al. (2022) use language as communication interface between various foundation models to sole certain tasks. Further, CLIP has demonstrated its robustness and versatility (Bommasani et al., 2021) across various tasks in the RL setup (Ostapenko et al., 2022; Parisi et al., 2022). We aim to investigate the effect of promising recent improvements upon the CLIP model, such as CLOOB (Fürst et al., 2021), more closely in our framework as well.

6 Conclusion

Reinforcement Learning with incomplete state information requires an agent to remember past events. Recent work had illustrated that large pretrained LMs are suitable for creating abstractions of past events (Paischer et al., 2022). HELM randomly maps observations to language tokens before compressing the compounding information via a LM. We presented certain conditions under which HELM is incapable of remembering information that is essential to solve a given task. Further, we proposed an incremental improvement over HELM comprising a feature-wise centering-and-scaling operation and a pretrained vision encoder, which we call HELMv2. On the challenging MiniGrid-Memory environment HELMv2 overcomes the limitations of HELM and successfully solves the task. Furthermore, HELMv2 yields significant improvements over HELM on minimalistic 2D and 3D environments.

Additionally, we pave the way toward a semantic history compression mechanism, that does not require training in the RL context. By leveraging multi-modal pretrained foundation models we demonstrated how to obtain a semantic mapping between a vision encoder and a LM. First results on MiniGrid and MiniWorld environments, however, showed that a semantic alignment between CLIP and LM does not significantly improve performance. Inherently, the semantic mapping is limited by the ability of the vision encoder to extract semantically meaningful features of the corresponding simulations. We surmise that this restriction is the reason why we did not observe improvements. In the future, we aim at improving the semantic mapping between the foundation models and demonstrating that it can provide benefits for more realistic scenes.

References

Adler, T., Brandstetter, J., Widrich, M., Mayr, A., Kreil, D. P., Kopp, M., Klambauer, G., and Hochreiter, S. Cross-Domain Few-Shot Learning by Representation Fusion. *CoRR*, abs/2010.06498, 2020. arXiv: 2010.06498.

Agarwal, R., Schwarzer, M., Castro, P. S., Courville, A. C., and Bellemare, M. G. Deep Reinforcement Learning at the Edge of the Statistical Precipice. In Ranzato, M., Beygelzimer, A., Dauphin, Y. N., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 29304–29320, 2021.

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., and Simonyan, K. Flamingo: a Visual Language Model for Few-Shot Learning. *CoRR*, abs/2204.14198, 2022. doi: 10.48550/arXiv.2204.14198. arXiv: 2204.14198.

Arjona-Medina, J. A., Gillhofer, M., Widrich, M., Unterthiner, T., Brandstetter, J., and Hochreiter, S. RUDDER: Return Decomposition for Delayed Rewards. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché Buc, F., Fox, E. B., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 13544–13555, 2019.

Artetxe, M., Labaka, G., and Agirre, E. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 789–798, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1073.

Bahdanau, D., Hill, F., Leike, J., Hughes, E., Hosseini, S. A., Kohli, P., and Grefenstette, E. Learning to Understand Goal Specifications by Modelling Reward. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

Berner, C., Brockman, G., Chan, B., Cheung, V., Debiak, P., Dennison, C., Farhi, D., Fischer, Q., Hashme, S., Hesse, C., Józefowicz, R., Gray, S., Olsson, C., Pachocki, J., Petrov, M., Pinto, H. P. d. O., Raiman, J., Salimans, T., Schlatter, J., Schneider, J., Sidor, S., Sutskever, I., Tang, J., Wolski, F., and Zhang, S. Dota 2 with Large Scale Deep Reinforcement Learning. *CoRR*, abs/1912.06680, 2019.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., Arx, S. v., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N. S., Chen, A. S., Creel, K., Davis, J. Q., Demszky, D., Donahue, C., Doumbouya, M., Durmus, E., Ermon, S., Etchemendy, J., Ethayarajh, K., Fei-Fei, L., Finn, C., Gale, T., Gillespie, L., Goel, K., Goodman, N. D., Grossman, S., Guha, N., Hashimoto, T., Henderson, P., Hewitt, J., Ho, D. E., Hong, J., Hsu, K., Huang, J., Icard, T., Jain, S., Jurafsky, D., Kalluri, P., Karamcheti, S., Keeling, G., Khani, F., Khattab, O., Koh, P. W., Krass, M. S., Krishna, R., Kuditipudi, R., and al, e. On the Opportunities and Risks of Foundation Models. *CoRR*, abs/2108.07258, 2021. arXiv: 2108.07258.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020.

Chen, L., Lu, K., Rajeswaran, A., Lee, K., Grover, A., Laskin, M., Abbeel, P., Srinivas, A., and Mordatch, I. Decision Transformer: Reinforcement Learning via Sequence Modeling. In Ranzato, M., Beygelzimer, A., Dauphin, Y. N., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 15084–15097, 2021.

Chevalier-Boisvert, M. MiniWorld: Minimalistic 3D Environment for RL & Robotics Research, 2018. Publication Title: GitHub repository.

Chevalier-Boisvert, M., Willems, L., and Pal, S. Minimalistic Gridworld Environment for OpenAI Gym, 2018. Publication Title: GitHub repository.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., Reif, E., Du, N., Hutchinson, B., Pope, R., Bradbury, J., Austin, J., Isard, M., Gur-Ari, G., Yin, P., Duke, T., Levskaya, A., Ghemawat, S., Dev, S., Michalewski, H., Garcia, X., Misra, V., Robinson, K., Fedus, L., Zhou, D., Ippolito, D., Luan, D., Lim, H., Zoph, B., Spiridonov, A., Sepassi, R., Dohan, D., Agrawal, S., Omernick, M., Dai, A. M., Pillai, T. S., Pellat, M., Lewkowycz, A., Moreira, E., Child, R., Polozov, O., Lee, K., Zhou, Z., Wang, X., Saeta, B., Diaz, M., Firat, O., Catasta, M., Wei, J., Meier-Hellstern, K., Eck, D., Dean, J., Petrov, S., and Fiedel, N. PaLM: Scaling Language Modeling with Pathways. *CoRR*, abs/2204.02311, 2022. doi: 10.48550/arXiv.2204.02311. arXiv: 2204.02311.

Craswell, N. Mean Reciprocal Rank. In LIU, L. and ÖZSU, M. T. (eds.), *Encyclopedia of Database Systems*, pp. 1703–1703. Springer US, Boston, MA, 2009. doi: 10.1007/978-0-387-39940-9_488.

Dai, Z., Yang, Z., Yang, Y., Carbonell, J. G., Le, Q. V., and Salakhutdinov, R. Transformer-XL: Attentive Language Models beyond a Fixed-Length Context. In Korhonen, A., Traum, D. R., and Màrquez, L. (eds.), *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pp. 2978–2988. Association for Computational Linguistics, 2019. doi: 10.18653/v1/p19-1285.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 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)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

Drori, I., Tran, S., Wang, R., Cheng, N., Liu, K., Tang, L., Ke, E., Singh, N., Patti, T. L., Lynch, J., Shporer, A., Verma, N., Wu, E., and Strang, G. A Neural Network Solves and Generates Mathematics Problems by Program Synthesis: Calculus, Differential Equations, Linear Algebra, and More. *CoRR*, abs/2112.15594, 2021.

Evci, U., Dumoulin, V., Larochelle, H., and Mozer, M. C. Head2Toe: Utilizing Intermediate Representations for Better Transfer Learning. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvári, C., Niu, G., and Sabato, S. (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 6009–6033. PMLR, 2022.

Fan, L., Wang, G., Jiang, Y., Mandlekar, A., Yang, Y., Zhu, H., Tang, A., Huang, D.-A., Zhu, Y., and Anandkumar, A. MineDojo: Building Open-Ended Embodied Agents with Internet-Scale Knowledge. *CoRR*, abs/2206.08853, 2022. doi: 10.48550/arXiv.2206.08853. arXiv: 2206.08853.

Fürst, A., Rumetshofer, E., Tran, V., Ramsauer, H., Tang, F., Lehner, J., Kreil, D. P., Kopp, M., Klambauer, G., Bitto-Nemling, A., and Hochreiter, S. CLOOB: Modern Hopfield Networks with InfoLOOB Outperform CLIP. *CoRR*, abs/2110.11316, 2021. arXiv: 2110.11316.

Gower, J. and Dijksterhuis, G. Procrustes Problems. *Procrustes Problems, Oxford Statistical Science Series*, Vol. 30, January 2005. doi: 10.1093/acprof:oso/9780198510581.001.0001. ISBN: 9780198510581.

Goyal, P., Niekum, S., and Mooney, R. J. Using Natural Language for Reward Shaping in Reinforcement Learning. In Kraus, S. (ed.), *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pp. 2385–2391. ijcai.org, 2019. doi: 10.24963/ijcai.2019/331.

Griffith, K. and Kalita, J. Solving Arithmetic Word Problems Automatically Using Transformer and Unambiguous Representations. *CoRR*, abs/1912.00871, 2019.

Groenen, P. J. F., Giaquinto, P., and Kiers, H. A. L. An improved majorization algorithm for robust procrustes analysis. *Springer US*, pp. 151–158, 2005.

Hausknecht, M. J. and Stone, P. Deep Recurrent Q-Learning for Partially Observable MDPs. In *2015 AAAI Fall Symposia, Arlington, Virginia, USA, November 12-14, 2015*, pp. 29–37. AAAI Press, 2015.

He, K., Zhang, X., Ren, S., and Sun, J. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 770–778. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.90.

Hellinger, E. Neue Begründung der Theorie quadratischer Formen von unendlichvielen Veränderlichen. *Journal für die reine und angewandte Mathematik*, 1909(136):210–271, 1909. doi: doi:10.1515/crll.1909.136.210.

Hill, F., Tieleman, O., Glehn, T. v., Wong, N., Merzic, H., and Clark, S. Grounded Language Learning Fast and Slow. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

Hochreiter, S. and Schmidhuber, J. Long Short-Term Memory. *Neural Comput.*, 9(8):1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. d. L., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., Driessche, G. v. d., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., Rae, J. W., Vinyals, O., and Sifre, L. Training Compute-Optimal Large Language Models. *CoRR*, abs/2203.15556, 2022. doi: 10.48550/arXiv.2203.15556. arXiv: 2203.15556.

Holzleitner, M., Gruber, L., Arjona-Medina, J. A., Brandstetter, J., and Hochreiter, S. Convergence Proof for Actor-Critic Methods Applied to PPO and RUDDER. *Trans. Large Scale Data Knowl. Centered Syst.*, 48:105–130, 2021. doi: 10.1007/978-3-662-63519-3_5.

Hoshen, Y. and Wolf, L. Non-Adversarial Unsupervised Word Translation. In Riloff, E., Chiang, D., Hockenmaier, J., and Tsujii, J. (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pp. 469–478. Association for Computational Linguistics, 2018. doi: 10.18653/v1/d18-1043.

Jaegle, A., Gimeno, F., Brock, A., Vinyals, O., Zisserman, A., and Carreira, J. Perceiver: General Perception with Iterative Attention. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4651–4664. PMLR, 2021.

Janner, M., Li, Q., and Levine, S. Offline Reinforcement Learning as One Big Sequence Modeling Problem. In Ranzato, M., Beygelzimer, A., Dauphin, Y. N., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 1273–1286, 2021.

Järvelin, K. and Kekäläinen, J. Cumulated Gain-Based Evaluation of IR Techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, October 2002. ISSN 1046-8188. doi: 10.1145/582415.582418. Place: New York, NY, USA Publisher: Association for Computing Machinery.

Kaelbling, L. P., Littman, M. L., and Cassandra, A. R. Planning and Acting in Partially Observable Stochastic Domains. *Artificial Intelligence*, 101:99–134, 1998.

Kassner, N., Krojer, B., and Schütze, H. Are Pretrained Language Models Symbolic Reasoners over Knowledge? In Fernández, R. and Linzen, T. (eds.), *Proceedings of the 24th Conference on Computational Natural Language Learning, CoNLL 2020, Online, November 19-20, 2020*, pp. 552–564. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.conll-1.45.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N. C., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., and Hadsell, R. Overcoming catastrophic forgetting in neural networks. *CoRR*, abs/1612.00796, 2016.

Lample, G., Conneau, A., Ranzato, M., Denoyer, L., and Jégou, H. Word translation without parallel data. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

Li, S., Puig, X., Paxton, C., Du, Y., Wang, C., Fan, L., Chen, T., Huang, D.-A., Akyürek, E., Anandkumar, A., Andreas, J., Mordatch, I., Torralba, A., and Zhu, Y. Pre-Trained Language Models for Interactive Decision-Making. *CoRR*, abs/2202.01771, 2022. arXiv: 2202.01771.

Lin, T.-Y., Maire, M., Belongie, S. J., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft COCO: Common Objects in Context. In Fleet, D. J., Pajdla, T., Schiele, B., and Tuytelaars, T. (eds.), *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pp. 740–755. Springer, 2014. doi: 10.1007/978-3-319-10602-1_48.

Melo, L. C. Transformers are Meta-Reinforcement Learners. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvári, C., Niu, G., and Sabato, S. (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 15340–15359. PMLR, 2022.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. Distributed Representations of Words and Phrases and their Compositionality. In Burges, C. J., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

Minixhofer, B., Paischer, F., and Rekabsaz, N. WECHSEL: Effective initialization of subword embeddings for cross-lingual transfer of monolingual language models. In Carpuat, M., Marneffe, M.-C. d., and Ruíz, I. V. M. (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pp. 3992–4006. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.naacl-main.293.

Mokady, R., Hertz, A., and Bermano, A. H. ClipCap: CLIP Prefix for Image Captioning. *CoRR*, abs/2111.09734, 2021. arXiv: 2111.09734.

Mu, J., Zhong, V., Raileanu, R., Jiang, M., Goodman, N. D., Rocktäschel, T., and Grefenstette, E. Improving Intrinsic Exploration with Language Abstractions. *CoRR*, abs/2202.08938, 2022. arXiv: 2202.08938.

Noorbakhsh, K., Sulaiman, M., Sharifi, M., Roy, K., and Jamshidi, P. Pretrained Language Models are Symbolic Mathematics Solvers too! *CoRR*, abs/2110.03501, 2021.

Ostapenko, O., Lesort, T., Rodríguez, P., Arefin, M. R., Douillard, A., Rish, I., and Charlin, L. Foundational Models for Continual Learning: An Empirical Study of Latent Replay. *arXiv preprint arXiv:2205.00329*, 2022.

Paischer, F., Adler, T., Patil, V. P., Bitto-Nemling, A., Holzleitner, M., Lehner, S., Eghbal-zadeh, H., and Hochreiter, S. History Compression via Language Models in Reinforcement Learning. *CoRR*, abs/2205.12258, 2022. doi: 10.48550/arXiv.2205.12258. arXiv: 2205.12258.

Parisi, S., Rajeswaran, A., Purushwalkam, S., and Gupta, A. The Unsurprising Effectiveness of Pre-Trained Vision Models for Control. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvári, C., Niu, G., and Sabato, S. (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 17359–17371. PMLR, 2022.

Parisotto, E., Song, F., Rae, J., Pascanu, R., Gulcehre, C., Jayakumar, S., Jaderberg, M., Kaufman, R. L., Clark, A., Noury, S., Botvinick, M., Heess, N., and Hadsell, R. Stabilizing Transformers for Reinforcement Learning. In *International Conference on Machine Learning*, pp. 7487–7498. PMLR, November 2020.

Patil, V. P., Hofmarcher, M., Dinu, M.-C., Dorfer, M., Blies, P. M., Brandstetter, J., Arjona-Medina, J. A., and Hochreiter, S. Align-RUDDER: Learning From Few Demonstrations by Reward Redistribution. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvári, C., Niu, G., and Sabato, S. (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 17531–17572. PMLR, 2022.

Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P. S. H., Bakhtin, A., Wu, Y., and Miller, A. H. Language Models as Knowledge Bases? In Inui, K., Jiang, J., Ng, V., and Wan, X. (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pp. 2463–2473. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1250.

Pleines, M., Pallasch, M., Zimmer, F., and Preuss, M. Generalization, Mayhems and Limits in Recurrent Proximal Policy Optimization. *CoRR*, abs/2205.11104, 2022. doi: 10.48550/arXiv.2205.11104. arXiv: 2205.11104.

Radford, A. and Narasimhan, K. Improving Language Understanding by Generative Pre-Training. 2018.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning Transferable Visual Models From Natural Language Supervision. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Rae, J. W., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, H. F., Aslanides, J., Henderson, S., Ring, R., Young, S., Rutherford, E., Hennigan, T., Menick, J., Cassirer, A., Powell, R., Driessche, G. v. d., Hendricks, L. A., Rauh, M., Huang, P.-S., Glaese, A., Welbl, J., Dathathri, S., Huang, S., Uesato, J., Mellor, J., Higgins, I., Creswell, A., McAleese, N., Wu, A., Elsen, E., Jayakumar, S. M., Buchatskaya, E., Budden, D., Sutherland, E., Simonyan, K., Paganini, M., Sifre, L., Martens, L., Li, X. L., Kuncoro, A., Nematzadeh, A., Gribovskaya, E., Donato, D., Lazaridou, A., Mensch, A., Lespiau, J.-B., Tsimpoukelli, M., Grigorev, N., Fritz, D., Sottiaux, T., Pajarskas, M., Pohlen, T., Gong, Z., Toyama, D., d'Autume, C. d. M., Li, Y., Terzi, T., Mikulik, V., Babuschkin, I., Clark, A., Casas, D. d. L., Guy, A., Jones, C., Bradbury, J., Johnson, M., Hechtman, B. A., Weidinger, L., Gabriel, I., Isaac, W. S., Lockhart, E., Osindero, S., Rimell, L., Dyer, C., Vinyals, O., Ayoub, K., Stanway, J., Bennett, L., Hassabis, D., Kavukcuoglu, K., and Irving, G. Scaling Language Models: Methods, Analysis & Insights from Training Gopher. *CoRR*, abs/2112.11446, 2021. arXiv: 2112.11446.

Reid, M., Yamada, Y., and Gu, S. S. Can Wikipedia Help Offline Reinforcement Learning? *CoRR*, abs/2201.12122, 2022. arXiv: 2201.12122.

Ruvolo, P. and Eaton, E. ELLA: An Efficient Lifelong Learning Algorithm. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, volume 28 of *JMLR Workshop and Conference Proceedings*, pp. 507–515. JMLR.org, 2013.

Schmidhuber, J. Learning Complex, Extended Sequences Using the Principle of History Compression. *Neural Comput.*, 4(2):234–242, 1992. doi: 10.1162/neco.1992.4.2.234.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal Policy Optimization Algorithms. *CoRR*, abs/1707.06347, 2017. arXiv: 1707.06347.

Schwarz, J., Czarnecki, W., Luketina, J., Grabska-Barwinska, A., Teh, Y. W., Pascanu, R., and Hadsell, R. Progress & Compress: A scalable framework for continual learning. In Dy, J. G. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4535–4544. PMLR, 2018.

Schönemann, P. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1): 1–10, 1966.

Smith, S. L., Turban, D. H. P., Hamblin, S., and Hammerla, N. Y. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

Stewart, G. W. The Efficient Generation of Random Orthogonal Matrices with an Application to Condition Estimators. *SIAM Journal on Numerical Analysis*, 17(3):403–409, 1980. doi: 10.1137/0717034. _eprint: https://doi.org/10.1137/0717034.

Su, Y., Lan, T., Liu, Y., Liu, F., Yogatama, D., Wang, Y., Kong, L., and Collier, N. Language Models Can See: Plugging Visual Controls in Text Generation. *CoRR*, abs/2205.02655, 2022. doi: 10.48550/arXiv.2205.02655. arXiv: 2205.02655.

Sukhbaatar, S., Ju, D., Poff, S., Roller, S., Szlam, A., Weston, J., and Fan, A. Not All Memories are Created Equal: Learning to Forget by Expiring. In *International Conference on Machine Learning*, pp. 9902–9912. PMLR, July 2021.

Talmor, A., Elazar, Y., Goldberg, Y., and Berant, J. oLMpics - On what Language Model Pre-training Captures. *Trans. Assoc. Comput. Linguistics*, 8:743–758, 2020.

Tam, A. C., Rabinowitz, N. C., Lampinen, A. K., Roy, N. A., Chan, S. C. Y., Strouse, D. J., Wang, J. X., Banino, A., and Hill, F. Semantic Exploration from Language Abstractions and Pretrained Representations. *CoRR*, abs/2204.05080, 2022. doi: 10.48550/arXiv.2204.05080. arXiv: 2204.05080.

Tewel, Y., Shalev, Y., Schwartz, I., and Wolf, L. Zero-Shot Image-to-Text Generation for Visual-Semantic Arithmetic. *arXiv preprint arXiv:2111.14447*, 2021.

Thoppilan, R., Freitas, D. D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H.-T., Jin, A., Bos, T., Baker, L., Du, Y., Li, Y., Lee, H., Zheng, H. S., Ghafouri, A., Menegali, M., Huang, Y., Krikun, M., Lepikhin, D., Qin, J., Chen, D., Xu, Y., Chen, Z., Roberts, A., Bosma, M., Zhou, Y., Chang, C.-C., Krivokon, I., Rusch, W., Pickett, M., Meier-Hellstern, K. S., Morris, M. R., Doshi, T., Santos, R. D., Duke, T., Soraker, J., Zevenbergen, B., Prabhakaran, V., Diaz, M., Hutchinson, B., Olson, K., Molina, A., Hoffman-John, E., Lee, J., Aroyo, L., Rajakumar, R., Butryna, A., Lamm, M., Kuzmina, V., Fenton, J., Cohen, A., Bernstein, R., Kurzweil, R., Aguera-Arcas, B., Cui, C., Croak, M., Chi, E. H., and Le, Q. LaMDA: Language Models for Dialog Applications. *CoRR*, abs/2201.08239, 2022. arXiv: 2201.08239.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., Choi, D. H., Powell, R., Ewalds, T., Georgiev, P., Oh, J., Horgan, D., Kroiss, M., Danihelka, I., Huang, A., Sifre, L., Cai, T., Agapiou, J. P., Jaderberg, M., Vezhnevets, A. S., Leblond, R., Pohlen, T., Dalibard, V., Budden, D., Sulsky, Y., Molloy, J., Paine, T. L., Gülçehre, C., Wang, Z., Pfaff, T., Wu, Y., Ring, R., Yogatama, D., Wünsch, D., McKinney, K., Smith, O., Schaul, T., Lillicrap, T. P., Kavukcuoglu, K., Hassabis, D., Apps, C., and Silver, D. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nat.*, 575(7782):350–354, 2019. doi: 10.1038/s41586-019-1724-z.

Wang, X., Huang, Q., Celikyilmaz, A., Gao, J., Shen, D., Wang, Y.-F., Wang, W. Y., and Zhang, L. Reinforced Cross-Modal Matching and Self-Supervised Imitation Learning for Vision-Language Navigation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 6629–6638. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00679.

Waxman, S. R. and Markow, D. Words as Invitations to Form Categories: Evidence from 12- to 13-Month-Old Infants. *Cognitive Psychology*, 29:257–302, 1995.

Widrich, M., Hofmarcher, M., Patil, V. P., Bitto-Nemling, A., and Hochreiter, S. Modern Hopfield Networks for Return Decomposition for Delayed Rewards. In *Deep RL Workshop NeurIPS 2021*, 2021.

Wilcoxon, F. Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, 1(6):80–83, 1945. ISSN 00994987. Publisher: [International Biometric Society, Wiley].

Xing, C., Wang, D., Liu, C., and Lin, Y. Normalized Word Embedding and Orthogonal Transform for Bilingual Word Translation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1006–1011, Denver, Colorado, May 2015. Association for Computational Linguistics. doi: 10.3115/v1/N15-1104.

Yao, S., Rao, R., Hausknecht, M. J., and Narasimhan, K. Keep CALM and Explore: Language Models for Action Generation in Text-based Games. In Webber, B., Cohn, T., He, Y., and Liu, Y. (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pp. 8736–8754. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.704.

Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., and Wu, Y. CoCa: Contrastive Captioners are Image-Text Foundation Models. *CoRR*, abs/2205.01917, 2022. doi: 10.48550/arXiv.2205.01917. arXiv: 2205.01917.

Zeng, A., Wong, A., Welker, S., Choromanski, K., Tombari, F., Purohit, A., Ryoo, M. S., Sindhwani, V., Lee, J., Vanhoucke, V., and Florence, P. Socratic Models: Composing Zero-Shot Multimodal Reasoning with Language. *CoRR*, abs/2204.00598, 2022. doi: 10.48550/arXiv.2204.00598. arXiv: 2204.00598.

Zenke, F., Poole, B., and Ganguli, S. Continual Learning Through Synaptic Intelligence. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3987–3995. PMLR, 2017.

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., Mihaylov, T., Ott, M., Shleifer, S., Shuster, K., Simig, D., Koura, P. S., Sridhar, A., Wang, T., and Zettlemoyer, L. OPT: Open Pre-trained Transformer Language Models. *CoRR*, abs/2205.01068, 2022. doi: 10.48550/arXiv.2205.01068. arXiv: 2205.01068.

Zhang, Y., Gaddy, D., Barzilay, R., and Jaakkola, T. Ten Pairs to Tag – Multilingual POS Tagging via Coarse Mapping between Embeddings. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1307–1317, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1156.

Zheng, Q., Zhang, A., and Grover, A. Online Decision Transformer. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvári, C., Niu, G., and Sabato, S. (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 27042–27059. PMLR, 2022.

Åström, K. J. Optimal control of Markov processes with incomplete state information. *Journal of Mathematical Analysis and Applications*, 10:174–205, 1964.

A Supplementary Material

A.1 Environments

We choose 8 diverse 3D environments of the MiniWorld benchmark suite:

- **TMaze:** The agent must navigate towards an object that is randomly placed at either end of a T-junction.
- **YMaze:** Same as TMaze, but with a Y-junction.
- **MazeS3Fast:** A procedurally generated maze in which the agent needs to find a goal object.
- **FourRooms:** The agent must reach a red box that is located in one of four interconnected rooms.
- **PickupObjs:** Several objects are placed in a large room and must be collected by the agent.
- **CollectHealth:** The agent spawns in a room filled with acid and must collect medikits in order to survive as long as possible.
- **PutNext:** Several boxes of various colors and sizes are placed in a big room. The agent must put a red box next to a yellow one.
- **Sign:** The agent spawns in a U-shaped maze containing various objects of different colors. One side of the maze contains a sign which displays a color in written form. The aim is to collect all objects in the corresponding color.

We neglect the OneRoom and the Hallway environments, since those are easily solved by all our methods. Further, we neglect the Sidewalk environment since it is essentially the same task as Hallway with a different background. For a more detailed description of the MiniGrid environments we refer the reader to Paischer et al. (2022).

A.2 Mapping between Language Embedding Spaces

We show additional results for computing linear mappings between CLIP and TrXL for different image backbones. In Table 6 we find that generally there is a tendency that the mapping improves for bigger backbones. Again, the Procrustes variants consistently exhibit the highest average accuracy on the test set across 5 folds.

	Linear	Ridge	Procrustes	RobProc
RN50	0.786/0.0013	0.571/0.281	0.64/**0.304**	0.65/0.303
RN101	0.561/0.013	0.56/0.277	0.621/0.313	0.642/**0.314**
RN50x4	0.628/0.003	0.625/0.27	0.678/**0.329**	0.691/0.326
RN50x16	0.694/0.002	0.692/0.272	0.722/0.345	0.732/**0.348**
RN50x64	0.798/0.001	0.79/0.274	0.742/**0.356**	0.745/0.353
ViT-B/32	0.576/0.008	0.574/0.295	0.595/**0.321**	0.612/0.317
ViT-B/16	0.585/0.004	0.583/0.309	0.614/**0.342**	0.631/0.337
ViT-L/14	0.705/0.002	0.677/0.314	0.669/**0.368**	0.677/0.367
ViT-L/14*	0.705/0.003	0.677/0.321	0.666/**0.37**	0.676/0.367

Table 6: Train/Test accuracy for different linear models optimized for mapping CLIP tokens to the TrXL embedding space. Average over 5-fold cross validation is shown. Centering and scaling is used for pre-processing of the token embeddings. For the Procrustes mappings, zero-padding is applied if the CLIP embedding is of lower dimensionality than the TrXL embeddings. The Procrustes methods yield the best mapping between the two embedding spaces. ViT-L/14* received images resized to 336 pixels as input during pretraining. Linear heavily overfits to the training set. Larger backbones lead to a better test accuracy.

A.3 Mapping Images to the Language Space

We complement our findings on the ranking experiment in Section 4 with results for additional vision backbones of CLIP. Table 7 shows the NDCG, while Table 8 shows the MRR for the different image

encoders. Again, we observe the trend towards increased MRR and NDCG for larger backbones. We also include token rankings in the CLIP space which serve as a point of reference on how much information is lost by the mapping to the LM space.

	Linear	Ridge	Procrustes	RobProc	CLIP
RN50	0.252±0.038	0.285±0.062	0.325±0.082	**0.326±0.082**	0.421±0.099
RN101	0.26±0.042	0.285±0.06	0.344±0.089	**0.345±0.09**	0.523±0.099
RN50x4	0.255±0.037	0.275±0.049	0.345±0.089	**0.347±0.089**	0.42±0.096
RN50x16	0.253±0.037	0.272±0.052	0.34±0.089	**0.343±0.089**	0.419±0.099
RN50x64	0.257±0.042	0.265±0.047	0.342±0.091	**0.343±0.091**	0.429±0.101
ViT-B/32	0.255±0.038	0.269±0.052	0.35±0.088	**0.35±0.089**	0.433±0.102
ViT-B/16	0.251±0.039	0.278±0.056	0.357±0.091	**0.356±0.091**	0.435±0.0.099
ViT-L/14	0.252±0.037	0.271±0.049	0.35±0.093	**0.351±0.093**	0.43±0.098
ViT-L/14*	0.255±0.041	0.271±0.05	0.356±0.095	**0.357±0.097**	0.439±0.099

Table 7: Normalized Discounted Cumulative Gain (NDCG) for ranked tokens in the LM embedding space given an image as input and applying our mapping. Image-caption pairs are drawn from the MSCOCO dataset. Ranking tokens in the CLIP space provides an upper bound. ViT-L/14* receives images resized to 336 pixels as input.

	Linear	Ridge	Procrustes	RobProc	CLIP
RN50	0.034±0.113	0.115±0.231	0.26±0.35	**0.264±0.353**	0.514±0.4
RN101	0.054±0.144	0.102±0.21	0.315±0.366	**0.318±0.37**	0.532±0.393
RN50x4	0.03±0.086	0.077±0.179	0.322±0.37	**0.33±0.374**	0.524±0.391
RN50x16	0.023±0.059	0.078±0.185	0.308±0.372	**0.322±0.379**	0.52±0.4
RN50x64	0.05±0.153	0.062±0.17	0.311±0.376	**0.314±0.375**	0.532±0.0.396
ViT-B/32	0.032±0.095	0.073±0.189	0.329±0.369	**0.332±0.372**	0.558±0.405
ViT-B/16	0.036±0.127	0.102±0.23	0.371±0.384	**0.372±0.39**	0.554±0.4
ViT-L/14	0.036±0.107	0.073±0.182	0.329±0.381	**0.337±0.382**	0.552±0.4
ViT-L/14*	0.043±0.129	0.069±0.164	0.352±0.386	**0.362±0.392**	0.575±0.393

Table 8: Mean Reciprocal Rank (MRR) for ranked tokens in the LM embedding space given an image as input and applying our mapping. On average, the first token appearing in an image caption is approximately ranked third. Image-caption pairs are drawn from the MSCOCO dataset. MRR in the CLIP output space serves as upper bound. On average CLIP ranks the first token appearing in a corresponding caption on second position. ViT-L/14* receives images resized to 336 pixels as input.