CROSS-MODAL ALIGNMENT REGULARIZATION: ENHANCING LANGUAGE MODELS WITH VISION MODEL REPRESENTATIONS

Yulu Gan*
CSAIL & CBMM, MIT

Kaiya Ivy Zhao*
CSAIL, MIT

Phillip Isola
CSAIL, MIT

ABSTRACT

Cross-modal distillation has emerged as a critical technique for leveraging strengths across different modalities. However, existing methods have not enabled performance benefits between models trained on different modal data. In this work, we introduce a cross-modal alignment regularization (CMAR) term into language model training, aligning its representations with those of a vision model at specific layers. Our experiments demonstrate that our method enhances language model performance across various downstream tasks, in both pre-training and fine-tuning settings. Specifically, in the pre-training setting, we observe accuracy improvements of 1.01% on the Language Modeling Broadened to Account for Discourse Aspects (LAMBADA) dataset and 1.49% on the Causal Reasoning (COPA) dataset. Our method also proves effective in the fine-tuning setting, boosting performance by 1.20% on LAMBADA and 2.00% on COPA, indicating that a vision model can substantially enhance language model performance. CMAR provides a simple yet effective strategy to consistently enhance language model performance through representation alignment with vision models, which opens new avenues for improving model performance through direct cross-modal representation alignment.

1 INTRODUCTION

Human effortlessly combines multiple sensory inputs to form a unified understanding of the world, a capability that continues to be a challenge in AI. Although numerous studies have demonstrated AI's ability to handle different modalities, such as image captioning (Li et al., 2022), text-to-image generation (Ramesh et al., 2022), visual question answering (Chen et al., 2023), and robotic manipulation (Han et al., 2024; Duan et al., 2024; Yang et al., 2024), their designs mainly allow models to accept and process multiple modalities without specifically focusing on the mutual benefits between these modalities. Therefore, a question arises: can AI enhance its overall capabilities through an integration of different modalities like humans?

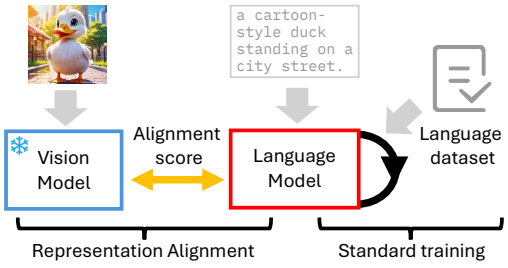


Figure 1: Idea overview. We incorporate representation alignment with other modalities, such as vision models, during the training stage of the language model to improve its performance.

Previous work to enhance model performance across modalities has primarily focused on using models from other modalities to generate data for the modality being optimized, which is then used to train the model of that specific modality. For example, Sharma et al. (2024) uses a language model to generate vision datasets for training vision models. Similarly, Baradad et al. (2022) utilizes image programs to create datasets for the same purpose. Although these methods effectively enhance the performance within individual modalities, they rely on an indirect path that merely facilitates unimodal

*Equal contribution.

improvements without fostering deep integration between modalities, which can be considered as using outputs from models as *data bridge* across modalities. However, constructing a data bridge is expensive and only conveys partial knowledge. Directly transferring knowledge between different modalities can be a more natural approach. Along that direction, Zhang et al. (2024) breaks away from building *data bridges* and learns vision-language alignment in representation space to enhance vision-language models. However, it still lacks cross-modality capability.

The Platonic Representation Hypothesis (PRH) introduced by Huh et al. (2024) suggests that, beyond information transfer through *data bridge*, representations across different modalities – such as images (X) and text (Y) – converge toward a shared underlying reality (Z). They observe a rough linear correlation between the alignment of language models with the vision model DinoV2 (Oquab et al., 2023) and their performance on Hellaswag (Zellers et al., 2019). Recently, Luo et al. (2024); Yu et al. (2024b); Subramaniam et al. (2024) have explored representation alignment along this line in various contexts. However, their focus remains within a single modality, neglecting the potential benefits of extending representation alignment to multi-modal scenarios.

To address this gap, we propose directly aligning representations between vision models and language models to enhance the performance of language models on downstream tasks. As Shukor & Cord (2024) points out, the manifolds of representations learned by different modalities are distinct. We hypothesize that aligning the representations of a language model (student model) with those of a vision model (teacher model) can enhance its ability to capture information shared across modalities. To this end, we introduce a method that adds a regularization term during the pre-training and fine-tuning stage of language models to encourage alignment with a fixed pre-trained vision model. This approach transcends traditional data bridging methods by embedding cross-modal alignment directly into the training process, fostering a shared conceptual framework that leverages the strengths of both modalities. The contributions of our method are:

- **Validating the Effectiveness of Direct Cross-Modal Feature Alignment**: We are the first to introduce a regularization term during language model training that aligns with specific layers of a vision model to enhance the performance of language models.
- **Removing Constraints on Architecture and Dimensions**: Different from previous methods for model distillation or model merging that require identical model architectures and feature dimensions to work effectively, our approach eliminates these constraints by computing the similarity of kernels in feature space. Refer to Section 2.1 for more details.
- **Proven Effective Through Various Experimental Settings**: We evaluate our method across 3 categories of tasks – Causal Reasoning, Commonsense Reasoning, and Text Generation – and demonstrate consistent improvements across 6 datasets. We also validate the effectiveness of our method in both fine-tuning and pre-training settings.

2 Related Work

2.1 Model Merging and Model Distillation.

While model merging (Matena & Raffel, 2022; Real et al., 2019; Yadav et al., 2023; Yu et al., 2024a; Akiba et al., 2024) offers a novel approach to leverage the strengths of multiple pre-trained models, it is constrained by the requirement that each model must have exactly the same architecture and identical feature dimensions in every matching layer. A similar story exists in model distillation techniques like Knowledge Distillation (Hinton, 2015) and Contrastive Representation Distillation (Tian et al., 2019), which also require the student and teacher models to have the same architecture and feature dimensions. However, in practice, we usually deal with models that have varying architectures and feature dimensions. Being able to distill or merge across such diverse models would represent a significant breakthrough. Lei et al. (2024) proposes a clever way to distill knowledge among different architectures by leveraging Mixture-of-Experts. While traditional model distillation has focused on a single modality, Cross Modal Distillation (Gupta et al., 2016) pioneers distillation across different vision-related modalities, specifically various image modalities. However, distillation across fundamentally distinct modalities like vision and language has not yet been explored. Our work addresses these limitations by eliminating the need for identical architectures and feature dimensions, covering both visual and language modalities.

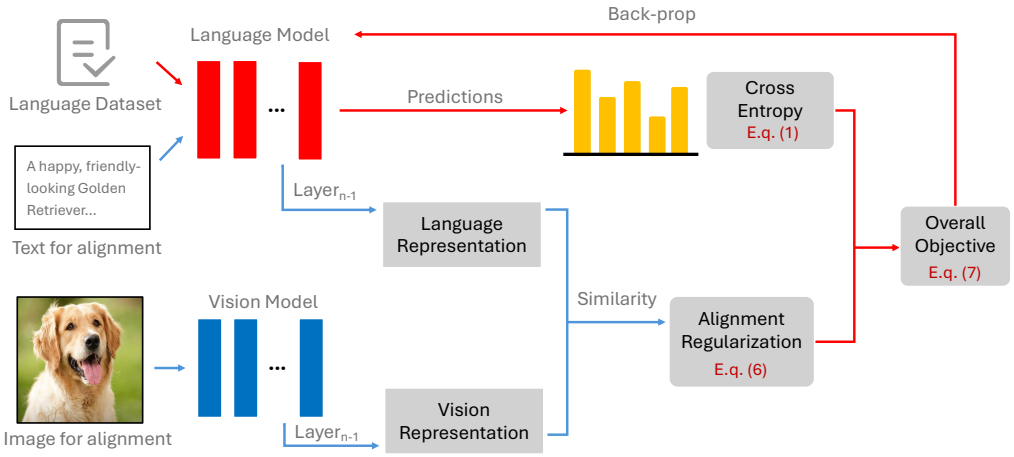


Figure 2: The workflow of the training process. Text and image inputs are processed independently to generate language and vision representations, which are used to calculate the alignment score using similarity-measure metrics. The combined loss, which includes both the original cross-entropy loss and our proposed alignment regularization term, updates the language model with back propagation to enhance its performance. The red lines indicate paths with gradient propagation, while the blue lines indicate paths with no gradient propagation.

2.2 Data Bridges and Representation Alignment in Multi-modal Models.

Multi-modal alignment aims to integrate information from different data types, such as vision and language, to build more robust and adaptable AI systems. Traditional methods typically rely on model outputs to improve another, namely *data bridge*. For example, Sharma et al. (2024) and Peng et al. (2024) employ language models to produce synthetic visual datasets for vision training, and Baradad et al. (2022) generates controlled visual inputs using image programs. In robotics, Huang et al. (2022) leverages language models as zero-shot learners, enabling robots to follow natural language instructions without task-specific training. Raman et al. (2022), Liang et al. (2022) and Qiu et al. (2023) make use of the commonsense knowledge embedded in large language models to guide and refine robot action. Although they can improve performance, their indirect integration methods hinder the development of deeper shared representations across different modalities. Instead, we eliminate this *data bridge* and directly align the representations between vision and language models.

2.3 Representation Alignment in Unimodal Models.

Representation alignment within single modality has garnered significant attention in recent deep learning research. For instance, Luo et al. (2024) explores the sharing of task vectors between various tasks. Similarly, Yu et al. (2024b) aligns the diffusion model with advanced vision models, and Bai et al. (2024) leverages vision models to infuse semantic richness for tokenization. Ren et al. (2024) finetunes language models with language model-generated data. All of them find that such alignment substantially enhances generative performance by fostering more coherent and high-quality outputs. Subramaniam et al. (2024) reveals alignment can render previously untrainable networks trainable. Qi et al. (2024) aligns the representation space of the vision encoder and action decoder. Distinct from the previous work, we aim to address a more challenging yet essential problem: aligning models across different modalities.

3 Method

We introduce Cross-Modal Alignment Regularization (CMAR) approach, which integrates a similarity-based alignment term into the training objective. Figure 2 illustrates the workflow: we first extract representations from selected layers of both the language and vision models; subsequently, we compute an alignment score based on the similarity between these representations, which is then incorporated into the loss function as a regularization term.

3.1 PRELIMINARIES AND ASSUMPTIONS

We consider a training setup with two datasets:(1) A language-only dataset $\{\mathbf{d}_j\}_{j=1}^{N}$, where each $\mathbf{d}_j$ is a sequence of tokens. This dataset is used to train a language model for next-token prediction. (2) A multi-modal dataset $\left\{\left(x_i^L, x_i^V\right)\right\}_{i=1}^{M}$ containing M samples, where x_i^L is a text sequence and x_i^V is a corresponding image. This dataset is used exclusively to measure and enforce alignment between text and vision representations.

We employ two encoders: (1) A language encoder $\psi_\theta : \mathcal{X}_L \to \mathbb{R}^{d_L}$, parameterized by θ, for text input. (2) A vision encoder $\phi_\varphi : \mathcal{X}_V \to \mathbb{R}^{d_V}$, parameterized by φ, for image input.

Each encoder has multiple layers. We allow the alignment to be computed using any one or more layers in each encoder. Concretely, let $\left\{\psi_\theta^{(p)}\right\}_{p\in\mathcal{P}}$ and $\left\{\phi_\varphi^{(q)}\right\}_{q\in\mathcal{Q}}$ denote the chosen sets of layers from the language encoder and vision encoder that we wish to align, respectively. Here, p and q denote indices corresponding to the layers chosen from the language and vision encoders, respectively. The sets $\mathcal{P}$ and $\mathcal{Q}$ define specific layers or combinations of layers that are used for alignment purposes. The choice could be a single layer, multiple layers, or even certain linear combinations of layer outputs, depending on the use case and architecture.

3.2 TRAINING OBJECTIVE

We train the language encoder ψ_θ by combining: (1) Negative Log-Likelihood Loss, $L_{\mathrm{NLL}}(\theta)$, on the text-only dataset, to optimize the language modeling objective. (2) Alignment Regularization Term, $L_{\text{align}}(\theta, \varphi)$, on the text-image dataset, to encourage alignment between the selected text and vision representations.

Negative Log-Likelihood Loss. We define $L_{\mathrm{NLL}}(\theta)$ on the language-only dataset $\{\mathbf{d}_j\}_{j=1}^{N}$. For causal language modeling, for instance:

$$L_{\mathrm{NLL}}(\theta) = -\sum_{j=1}^{N}\sum_{t=1}^{T_j} \log \psi_\theta\left(\mathbf{d}_j^{(t)} \mid \mathbf{d}_j^{(<t)}\right) \tag{1}$$

where $\mathbf{d}_j^{(t)}$ is the t-th token and $\mathbf{d}_j^{(<t)}$ the preceding tokens in the j-th text sequence. Attributed to NLL's generality, if a different supervised language task is used, this term could be replaced or supplemented by a cross-entropy or other relevant loss.

Alignment Regularization Term. For each text-image pair $\left(x_i^L, x_i^V\right)$ in the multi-modal datasets $\left\{\left(x_i^L, x_i^V\right)\right\}_{i=1}^{M}$, we extract latent vectors from the chosen layers in the language encoder:$\left\{\psi_\theta^{(p)}\left(x_i^L\right)\right\}_{p\in\mathcal{P}}$, and from the chosen layers in the vision encoder: $\left\{\phi_\varphi^{(q)}\left(x_i^V\right)\right\}_{q\in\mathcal{Q}}$.

We then define a similarity-based alignment metric $\mathcal{A}$ (e.g., Canonical Correlation Analysis, Centered Kernel Alignment, or a dot-product similarity) across one or more pairs of these layer outputs. For simplicity, let

$$H_i^{\text{language}}(\theta) = \mathrm{concat}_{p\in\mathcal{P}}\left(\psi_\theta^{(p)}\left(x_i^L\right)\right), \tag{2}$$

$$H_i^{\text{vision}}(\varphi) = \mathrm{concat}_{q\in\mathcal{Q}}\left(\phi_\varphi^{(q)}\left(x_i^V\right)\right), \tag{3}$$

We can define the alignment loss as:

$$L_{\mathrm{align}}(\theta, \varphi) = \sum_{i=1}^{M}\left[1 - \mathcal{A}\left(H_i^{\text{vision}}(\varphi), H_i^{\text{language}}(\theta)\right)\right] \tag{4}$$

Minimizing L_{align} results in maximizing $\mathcal{A}$, which enforces higher correlation between image and text embeddings in the chosen layers. This alignment serves as a regularization, guiding language encoder to produce representations similar to vision ones.

Overall Objective. Putting it all together, our overall objective combines the language modeling loss and the alignment regularization:

$$\theta^*, \varphi^* = \arg\min_{\theta,\varphi} [L_{\text{NLL}}(\theta) + \lambda L_{\text{align}}(\theta, \varphi)] \tag{5}$$

Here, $L_{\text{NLL}}(\theta)$ is computed on the text-only dataset for next-token prediction, while $L_{\text{align}}(\theta, \varphi)$ is computed about the similarity alignment between feature spaces on image-text pairs. The hyperparameter λ controls the relative importance of representation alignment versus token prediction.

4 Experimental results

To investigate the effectiveness of CMAR, we evaluate our method across both pre-training and fine-tuning settings. In the first setting, a language model is aligned with a vision model during the training process, allowing us to assess whether the vision representations can effectively enhance the learning of the language model. In the second setting, a pre-trained language model is aligned with a vision model during the fine-tuning process, investigating whether the language model can also benefit from the alignment with another modality after a thorough training on the language corpus.

We introduce the setup in Section 4.1 with design choices and configurations. Next, we analyze performance and report experimental results of the two settings aforementioned in Section 4.2 and 4.3 respectively. Finally, we conduct ablation studies in Section. 4.4.

4.1 Set up

Pre-Training. For aligning the student language model with the teacher vision model, our student language model is GPT-2 (124M) (Radford et al., 2019), a language model that leverages the transformer architecture to predict the next token in the sequence. GPT-2 exhibits zero-shot capabilities across diverse language tasks, making it a robust starting point for evaluating our approach. We use OpenWebText (Gokaslan et al., 2019) for training GPT-2 from scratch. Specifically, we adapt from NanoGPT (Karpathy, 2023) to build the pipeline.

Fine-tuning. For aligning student language model with weak vision model in the fine-tuning setting, we use Llama-3.1-8B (Grattafiori et al., 2024) and Llama-3.2-1B (Grattafiori et al., 2024) as our student language models. The Llama 3 series has gained widespread adoption for its performance on par with leading proprietary models, making them reasonably strong models for testing CMAR effectiveness. Besides, our teacher vision model is DinoV2-Small (Oquab et al., 2023) whose representation is used to guide language model. We leverage the Alpaca (Taori et al., 2023) dataset and pipeline for fine-tuning.

Implementation Details. For pre-training, we use 4 NVIDIA H100 GPUs and train for 8000 iterations with learning rate $6e-3$ and batch size 32. For finetuning, we use 1 NVIDIA H100 GPU for 500 iterations with learning rate $1e-5$ and batch size 8. After a thorough search on hyperparameter space (see Figure 8 in Appendix), we choose the weight of alignment regularization λ as 10 and the interval for applying alignment regularization is 20 iterations. We use Centered Kernel Alignment (CKA) as the alignment metric for all experiments except for the Section 4.4.

Vision Models to be Aligned with. To demonstrate the robustness of our method, we align the language model with seven different types and sizes of vision models, including MAE (He et al., 2022), DinoV2 (small, base, large, giant) (Oquab et al., 2023), CLIP (Radford et al., 2021), and AugReg models (Steiner et al., 2021). Refer to Table 2 in the Appendix for more details.

Evaluation Datasets. For both experiments, we evaluate our method using 6 datasets that encompass next-token prediction, causal reasoning, and commonsense reasoning tasks: ARC (Challenge),

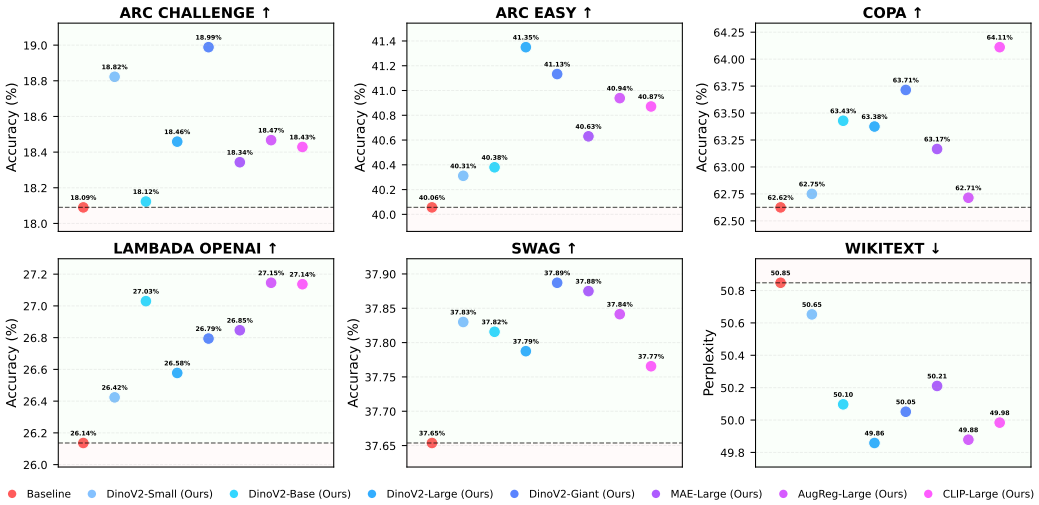


Figure 3: Results averaged on 5 experiments (details in Table 3), examining the alignment of representations in language models with those in various vision models. The points in green region show a better performance compared to the baseline, and a worse one in red region. Our results indicate that our method consistently improves performance on all datasets. Additionally, aligning with various vision models generally enhances the performance of language models.

ARC (Easy), COPA, LAMBADA, SWAG, and WikiText. A summary of these datasets is provided in Table 1; more details can be found in Appendix A.3.

4.2 Align Student Language Model with Teacher Vision Model in Pre-training

We test our methods on several tasks and datasets to investigate their performance. As demonstrated in Figure 3, language models that align with advanced vision models achieve superior performance on downstream tasks, underscoring the effectiveness of our CMAR method:

For the *Next-Token Prediction* task, we analyze our method using WikiText and LAMBADA, observing a performance gain of 0.99 in perplexity and an 1.01% increase in accuracy respectively.

For the *Causal Reasoning* task, our evaluation on the ARC-Easy benchmark shows model variants achieving accuracy rates exceeding 41.35%, which surpasses the baseline by approximately 1.29%. On ARC-Challenge dataset, the trend continues with model performance improving from 18.09% to 18.99%. These results suggest that integrating representations from vision models can significantly enhance the ability to tackle complex reasoning tasks as well as knowledge-based queries. Notable performance gains are also evident on the COPA with an increase of 1.49% on accuracy.

For the *Commonsense Reasoning* task, when evaluated on the SWAG dataset, the performance shows an improvement of 0.24% compared to the base model that does not incorporate representation alignment from the vision model.

Overall, these results collectively indicate that leveraging vision model representations can yield measurable improvements in language model performance across a variety of datasets and tasks, from general knowledge and commonsense reasoning to text generation and language modeling fluency. We also evaluate the effect of using different vision models in Section 4.4.4.

4.3 Align Student Language Model with Teacher Vision Model in Fine-tuning

As demonstrated in Figure 4, aligning language models with vision models during finetuning also acquires performance gains.

On the COPA dataset, the aligned *Llama3.2-1B* model achieves 82.0%, compared to 80.0% without alignment. Similarly, *Llama3.1-8B* achieves an accuracy of 90.0% with alignment, surpassing the unaligned baseline by 1.0%.

For the LAMBADA dataset, the alignment of *Llama3.1-8B* with the DinoV2-Small model results in an accuracy of 75.3%, showing a slight improvement over baseline (74.6%). Similarly, *Llama3.2-1B* achieves 61.8% when aligned, a marginal gain compared to 60.6% without alignment. These results suggest that incorporating vision model alignment positively impacts token prediction tasks as well in fine-tuning setting.

On the ARC-Easy dataset, the aligned *Llama3.1-8B* model achieves an accuracy of 82.6%, improving by 1.4% compared to the unaligned baseline (81.2%). *Llama3.2-1B* model, while achieving lower overall accuracy, also benefits from alignment, with accuracy improving from 64.2% to 65.3%. These improvements emphasize the value of visual alignment in enhancing causal reasoning capabilities.

Therefore, aligning with DinoV2-Small model provides consistent performance gains across datasets. These results collectively reinforce the effectiveness of using CMAR for aligning student language models with teacher vision models during fine-tuning, demonstrating improvements in token prediction, causal reasoning, and commonsense reasoning tasks.

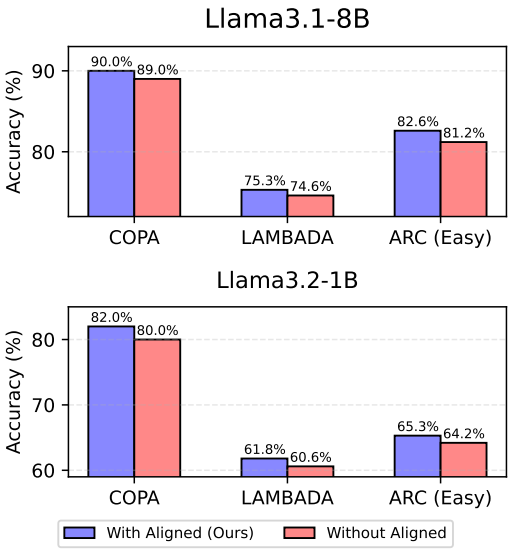


Figure 4: Alignment of student language models (Llama3.1-8B / Llama3.2-1B) with a vision teacher model (DinoV2-Small) boost the performance of language models.

4.4 Ablation Studies

To analyze the effects of alignment among different layers, alignment with various models and alignment by different metrics, we conduct ablation studies, with results and analyses below.

4.4.1 The Effect of Aligning with Different Vision Models

Results in Figure 3 need to be further investigated to clarify the underlying factors influencing the performance. We initially hypothesized that aligning the language model with larger and more capable vision models would introduce larger performance improvements. However, our results *do not* present a clear positive correlation between vision model size and downstream gains in language tasks. The ablation study includes training the student GPT-2 with alignment to a range of different vision models across four sizes of DinoV2 (small, base, large, giant), along with MAE (He et al., 2022), CLIP (Radford et al., 2021), and AugReg (Steiner et al., 2021) models.

As shown in Figure 3, on the COPA dataset, accuracy increases from 62.62% to 64.11%, with the best performance achieved when aligning with the CLIP model. This might be attributed to CLIP being pre-trained on the extensive Laion-2B dataset and subsequently fine-tuned on ImageNet-12K, which endows it with a rich knowledge base in the visual domain. This indicates that possessing substantial and relevant knowledge in the vision domain can amplify the improvements in language model training.

On the WikiText dataset, perplexity decreases from 50.85 to 49.86. Interestingly, the DinoV2-Large model outperforms the DinoV2-Giant model, indicating that model size is not the only determinant of performance when aligning language models. We hypothesize that the more powerful and senior teacher DinoV2-Giant model inherently has a larger representational gap with GPT-2, making it harder for the student language model to effectively learn from it. In contrast, DinoV2-Large model acts more like a junior teacher, allowing student model to adapt and align more effectively.

Our ablation study suggests that factors beyond raw model capacity – such as the nature of learned features, alignment calculation, or relative discrepancy between capacities of two modalities – may also contribute to the scale of improvement. Understanding the circumstances of when and

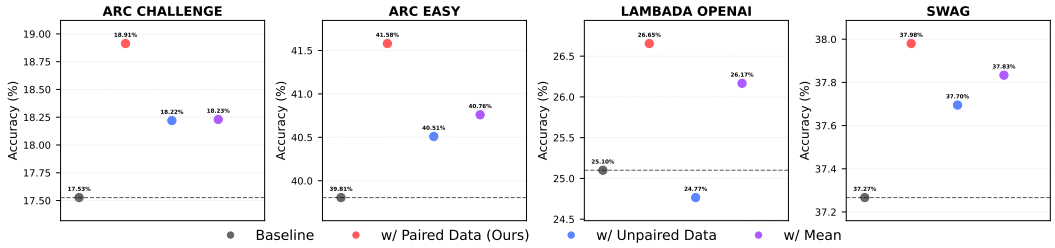


Figure 5: Ablation studies on data used for alignment. The figure shows the results of aligning paired image-text data (*w/ Paired Data*) outperform not only the baseline, but also the results of aligning with randomly-selected image-text (*w/ Unpaired Data*), and the performance of aligning with the mean of the image distribution (*w/ Mean*).

how different vision models translate into stronger cross-modal guidance is critical for deeper understanding of representation spaces in multi-modalities as well as enhancing training effectiveness, which we leave as an open question and a promising direction for future research.

4.4.2 The Effect of Aligning with Different Data Setup

In Figure 5, we conduct ablation studies on the data used for alignment, including (1) paired data, which consists of corresponding image-text pairs; (2) unpaired data, which consists of non-corresponding image-text pairs; (3) mean, which involves calculating the mean of vision representations across all samples and then aligning the language model with this mean.

The results show that aligning with mean, paired, or unpaired data all lead to clear improvements over the baseline model. This demonstrates that our method is not limited to paired data, and perhaps much of the improvement is simply about rescaling the distribution or some methods similarly straightforward. Also, we find that the performance of paired data is better than that of unpaired data, while the unpaired data also help boost the performance compared to the baseline. This actually opens up the future direction about what features are useful when deploying cross-modal alignment, and the potential effectiveness of alignment with general feature space in other modalities.

4.4.3 The Effect of Aligning at Different Layers

Figure 6 shows how model performance varies when aligning different layers with vision model representations. Since each layer captures unique representations and features, aligning vision model representations at various depths leads to varied impacts on the performance across different downstream tasks.

We also observe that while aligning multiple higher-level layers generally improves performance, the extent of improvement varies by task. Some tasks benefit more from broadly integrated features, whereas others require more specialized alignment. A detailed analysis of layer selection and alignment strategies offers a promising avenue for future research, possibly revealing richer cross-modal synergies.

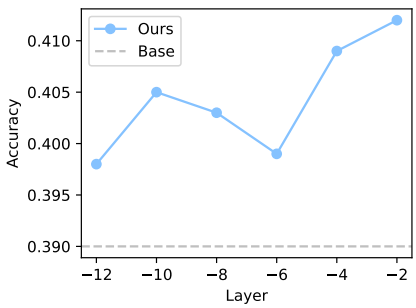


Figure 6: Ablation Studies on Aligned Layers. Results illustrate that aligning with penultimate layer works the best.

4.4.4 THE EFFECT OF ALIGNING BY DIFFERENT REGULARIZATION TERMS

To further investigate the effectiveness of the CMAR method under different regularization terms, we conduct ablation studies with additional regularization terms, such as linear CKA, SVCCA and PWCCA. As shown in Figure 7(a) and Figure 7(b), the results highlight the impact of aligning the model using CMAR over other well-known regularization methods. Among the tested methods, CMAR achieves the highest accuracy of 41.2%, substantially outperforming regularization terms L1 and L2, which only achieve accuracies of 30.5% and 28.4% respectively. Moreover, when focusing on the performance of CMAR across different alignment metrics, every metric works decently. It is also worth noticing that certain metrics are more effective in capturing useful features from vision teachers and guiding the learning. For instance, the CKA (Kornblith et al., 2019b) metric yields the highest accuracy of 41.2%, followed by SVCCA (Raghu et al., 2017b) with an accuracy of 41.0%, outperforming other methods like dot product due to its consistency in measuring similarities of representations across different architectures and datasets (Kornblith et al., 2019b). Conclusively, this ablation study demonstrates that CMAR is an effective regularization method and it works consistently for improving the performance regardless of the choice of alignment metrics. Further study could explore the effectiveness of different alignment metrics.

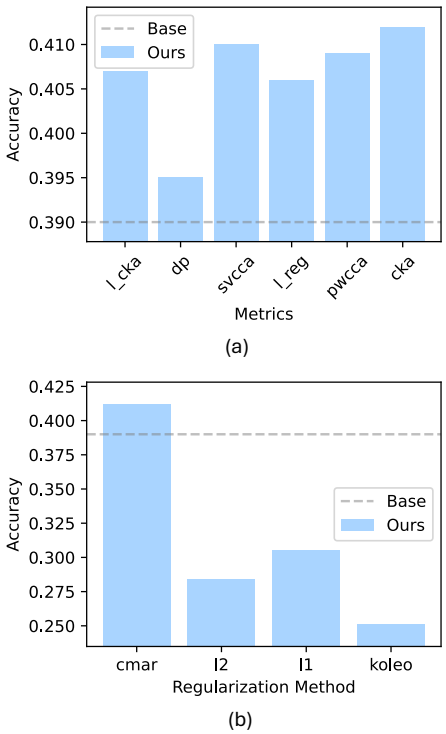


Figure 7: (a) All similarity metrics in CMAR work, among which CKA metric works the best. See Appendix A.7 for metrics details. (b) CMAR significantly outperforms other regularization terms like L1, L2 and Koleo Loss (Sablayrolles et al., 2019).

5 CONCLUSION AND DISCUSSION

In conclusion, we demonstrate that our CMAR improves language models performance by aligning their representations with those from vision models. Our results underscore the effectiveness of the method, with empirical evidence showing improvements in accuracy or perplexity across six datasets in three categories, exemplified by notable increases of 1.01% and 1.20% on the LAMBADA dataset, and 1.49% and 2.00% on the COPA dataset, for pre-training and fine-tuning, respectively. The success of CMAR demonstrates that cross-modality distillation through direct alignment of their representations is doable for both pre-training and fine-tuning settings. Our findings advance research beyond previous distillation work, which has been confined to models sharing the same architecture and feature dimensions. Although demonstrating potential in cross-modal alignment, we acknowledge that some limitations remain for future work. For example, we have evaluated our method on a limited number of model configurations. Future research can investigate whether the benefits of CMAR extend to more complex domains and explore how interactions among different model configurations affect cross-modal integration. Future research can also explore CMAR across diverse settings, model configurations, and task domains. This may include applying the method to additional modalities and examining the scaling law of cross-modal alignment, namely performance vs. alignment score, potentially transforming current practices in training multi-modal models.

REFERENCES

Takuya Akiba, Makoto Shing, Yujin Tang, Qi Sun, and David Ha. Evolutionary optimization of model merging recipes, 2024. URL https://arxiv.org/abs/2403.13187.

Zechen Bai, Jianxiong Gao, Ziteng Gao, Pichao Wang, Zheng Zhang, Tong He, and Mike Zheng Shou. Factorized visual tokenization and generation. 2024. URL https://api.semanticscholar.org/CorpusID:274281440.

Manel Baradad, Richard Chen, Jonas Wulff, Tongzhou Wang, Rogerio Feris, Antonio Torralba, and Phillip Isola. Procedural image programs for representation learning. Advances in Neural Information Processing Systems, 35:6450–6462, 2022.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme, Andreas Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. Pali: A jointly-scaled multilingual language-image model, 2023. URL https://arxiv.org/abs/2209.06794.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. ArXiv, abs/1803.05457, 2018.

Jiafei Duan, Wentao Yuan, Wilbert Pumacay, Yi Ru Wang, Kiana Ehsani, Dieter Fox, and Ranjay Krishna. Manipulate-anything: Automating real-world robots using vision-language models, 2024. URL https://arxiv.org/abs/2406.18915.

Aaron Gokaslan, Vanya Cohen, Ellie Pavlick, and Stefanie Tellex. Openwebtext corpus, 2019.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and Amy Yang et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Saurabh Gupta, Judy Hoffman, and Jitendra Malik. Cross modal distillation for supervision transfer. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2827–2836, 2016.

Dongge Han, Trevor McInroe, Adam Jelley, Stefano V. Albrecht, Peter Bell, and Amos Storkey. Llm-personalize: Aligning llm planners with human preferences via reinforced self-training for housekeeping robots, 2024. URL https://arxiv.org/abs/2404.14285.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 16000–16009, June 2022.

Geoffrey Hinton. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531, 2015.

Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language Models as Zero-Shot Planners: Extracting Actionable Knowledge for Embodied Agents, March 2022.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. arXiv preprint arXiv:2405.07987, 2024.

Andrej Karpathy. https://github.com/karpathy/nanoGPT, 2023.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In International Conference on Machine Learning (ICML), pp. 3519–3529, 2019a.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited, 2019b. URL https://arxiv.org/abs/1905.00414.

Hongyang Lei, Xiaolong Cheng, Dan Wang, Qi Qin, Huazhen Huang, Yetao Wu, Qingqing Gu, Zhonglin Jiang, Yong Chen, and Luo Ji. Alt-moe: Multimodal alignment via alternating optimization of multi-directional moe with unimodal models. ArXiv, abs/2409.05929, 2024. URL https://api.semanticscholar.org/CorpusID:272550700.

Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, Hehong Chen, Guohai Xu, Zheng Cao, Ji Zhang, Songfang Huang, Fei Huang, Jingren Zhou, and Luo Si. mplug: Effective and efficient vision-language learning by cross-modal skip-connections, 2022. URL https://arxiv.org/abs/2205.12005.

Jacky Liang, Wenlong Huang, F. Xia, Peng Xu, Karol Hausman, Brian Ichter, Peter R. Florence, and Andy Zeng. Code as policies: Language model programs for embodied control. 2023 IEEE International Conference on Robotics and Automation (ICRA), pp. 9493–9500, 2022. URL https://api.semanticscholar.org/CorpusID:252355542.

Grace Luo, Trevor Darrell, and Amir Bar. Task vectors are cross-modal. arXiv preprint arXiv:2410.22330, 2024.

Michael Matena and Colin Raffel. Merging models with fisher-weighted averaging, 2022. URL https://arxiv.org/abs/2111.09832.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer Sentinel Mixture Models, 2016a. URL http://arxiv.org/abs/1609.07843.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models, 2016b.

Ari S Morcos, Maithra Raghu, and Samy Bengio. Insights on representational similarity in neural networks with canonical correlation. In Advances in Neural Information Processing Systems (NeurIPS), 2018.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193, 2023.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Quan Ngoc Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. The LAMBADA dataset, Aug 2016.

Yi-Hao Peng, Faria Huq, Yue Jiang, Jason Wu, Amanda Li, Jeffrey P. Bigham, and Amy Pavel. Dreamstruct: Understanding slides and user interfaces via synthetic data generation. In European Conference on Computer Vision, 2024. URL https://api.semanticscholar.org/CorpusID:273023108.

Han Qi, Haocheng Yin, and Heng Yang. Control-oriented clustering of visual latent representation. ArXiv, abs/2410.05063, 2024. URL https://api.semanticscholar.org/CorpusID:273186586.

Jielin Qiu, Mengdi Xu, William Jongwon Han, Seungwhan Moon, and Ding Zhao. Embodied executable policy learning with language-based scene summarization. ArXiv, abs/2306.05696, 2023. URL https://api.semanticscholar.org/CorpusID:259129746.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. OpenAI blog, 1(8):9, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In Advances in Neural Information Processing Systems (NeurIPS), 2017a.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability, 2017b. URL https://arxiv.org/abs/1706.05806.

S. Sundar Raman, Vanya Cohen, Eric Rosen, Ifrah Idrees, David Paulius, and Stefanie Tellex. Cape: Corrective actions from precondition errors using large language models. 2024 IEEE International Conference on Robotics and Automation (ICRA), pp. 14070–14077, 2022. URL https://api.semanticscholar.org/CorpusID:253707906.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. URL https://arxiv.org/abs/2204.06125.

Esteban Real, Alok Aggarwal, Yanping Huang, and Quoc V Le. Regularized evolution for image classifier architecture search, 2019. URL https://arxiv.org/abs/1802.01548.

Xuan Ren, Biao Wu, and Lingqiao Liu. I learn better if you speak my language: Understanding the superior performance of fine-tuning large language models with llm-generated responses. In Conference on Empirical Methods in Natural Language Processing, 2024. URL https://api.semanticscholar.org/CorpusID:270211000.

Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, and Hervé Jégou. Spreading vectors for similarity search, 2019. URL https://arxiv.org/abs/1806.03198.

Pratyusha Sharma, Tamar Rott Shaham, Manel Baradad, Stephanie Fu, Adrian Rodriguez-Munoz, Shivam Duggal, Phillip Isola, and Antonio Torralba. A vision check-up for language models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14410–14419, 2024.

Mustafa Shukor and Matthieu Cord. Implicit multimodal alignment: On the generalization of frozen llms to multimodal inputs, 2024. URL https://arxiv.org/abs/2405.16700.

Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge, 2018. URL https://arxiv.org/abs/1612.03975.

Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your vit? data, augmentation, and regularization in vision transformers, 2021.

Vighnesh Subramaniam, David Mayo, Colin Conwell, Tomaso Poggio, Boris Katz, Brian Cheung, and Andrei Barbu. Training the untrainable: Introducing inductive bias via representational alignment. arXiv preprint arXiv:2410.20035, 2024.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. 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. 4149–4158, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1421. URL https://aclanthology.org/N19-1421.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Alpaca: A strong, replicable instruction-following model. Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html, 3(6):7, 2023.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. arXiv preprint arXiv:1910.10699, 2019.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems, 2020. URL https://arxiv.org/abs/1905.00537.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. Ties-merging: Resolving interference when merging models, 2023. URL https://arxiv.org/abs/2306.01708.

Jingkang Yang, Yuhao Dong, Shuai Liu, Bo Li, Ziyue Wang, Chencheng Jiang, Haoran Tan, Jiamu Kang, Yuanhan Zhang, Kaiyang Zhou, and Ziwei Liu. Octopus: Embodied vision-language programmer from environmental feedback, 2024. URL https://arxiv.org/abs/2310.08588.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. Language models are super mario: Absorbing abilities from homologous models as a free lunch, 2024a. URL https://arxiv.org/abs/2311.03099.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. arXiv preprint arXiv:2410.06940, 2024b.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. Swag: A large-scale adversarial dataset for grounded commonsense inference. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2018.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? arXiv preprint arXiv:1905.07830, 2019.

Le Zhang, Qian Yang, and Aishwarya Agrawal. Assessing and learning alignment of unimodal vision and language models, 2024. URL https://arxiv.org/abs/2412.04616.

A Appendix

A.1 Implementation Details

All experiments are conducted using PyTorch. Each model is trained with a batch size of 64 and a learning rate of 0.006. We employed four 80GB Nvidia H100 GPUs, allocating approximately 10 hours to train each model. Computations are distributed across these GPUs to enhance training efficiency and scalability. The interval for updating the alignment loss is set to every 20 iterations, which facilitated the periodic integration of alignment loss into the training process.

A.2 Additional Ablation Studies

We conduct ablation studies on the λ and the interval. For λ, increasing it from 5 to 10 significantly improves accuracy, but further increases show minimal gains. For interval, a smaller value maintains higher accuracy, with a noticeable decline as the interval increases, indicating that the optimal interval should be around 20.

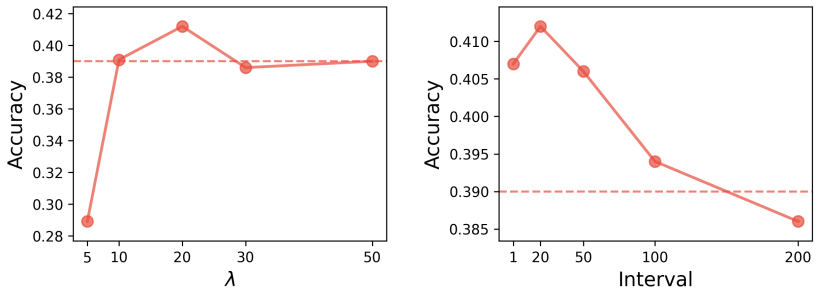


Figure 8: Additional studies on hyper parameter search.

A.3 Datasets Details

Dataset Name	Task Type	Size	Source
Wikitext	Next-Token Prediction	100M Words	Merity et al. (2016b)
LAMBADA	Next-Token Prediction	465M Words	Paperno et al. (2016)
COPA	Causal Reasoning	1.6k Q&A	Merity et al. (2016a)
ARC	Causal Reasoning	7k Q&A	Clark et al. (2018)
Commonsense QA	Commonsense Reasoning	12k Q&A	Talmor et al. (2019)
SWAG	Commonsense Reasoning	113k Q&A	Zellers et al. (2018)

Table 1: Dataset details. This table provides an overview of the datasets used in our experiments, categorized by their task type, size, and source. These datasets span various tasks such as next-token prediction, causal reasoning, and commonsense reasoning.

Below are introductions to six datasets we used for evaluation:

Wikitext (Merity et al., 2016b) dataset comprises over 100 million tokens from Wikipedia's verified Good and Featured articles. Tailored for language modeling, it maintains original article formatting, links, and structure, presenting a realistic challenge for language models.

LAMBADA (Paperno et al., 2016) evaluates computational models' language understanding by requiring the prediction of the final word in narrative passages that seem straightforward to a human reader. It emphasizes the use of broader textual contexts for deeper comprehension.

The *Choice of Plausible Alternatives (COPA)* (Merity et al., 2016a) dataset, part of the SuperGLUE Wang et al. (2020) benchmark, tests causal reasoning by asking participants to choose

between two plausible outcomes based on a given premise, enhancing understanding of cause-effect relationships in text.

AI2 Reasoning Challenge (ARC) (Clark et al., 2018) offers multiple-choice questions that demand advanced reasoning across various topics, challenging models to apply deep knowledge and reasoning capabilities beyond mere pattern recognition.

Commonsense QA (Talmor et al., 2019), derived from ConceptNet Speer et al. (2018), challenges models to apply everyday commonsense knowledge to answer intuitive questions, pushing the limits of what AI can understand and respond to coherently.

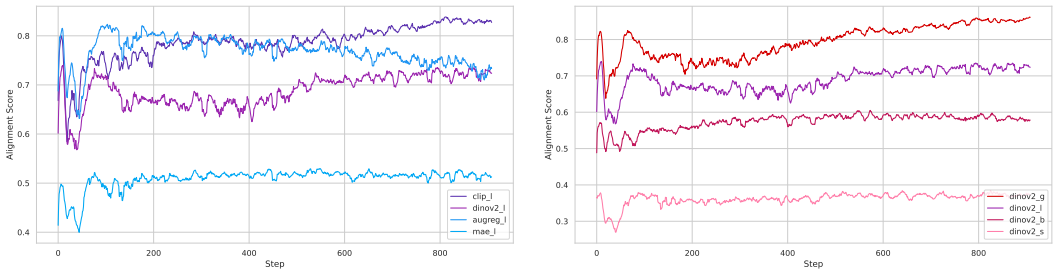
Situations With Adversarial Generations (SWAG) (Zellers et al., 2018) assesses models' ability to predict logical sentence endings in diverse scenarios. It features nearly 113,000 multiple-choice questions that test understanding and prediction of plausible outcomes.

A.4 Details of Vision Models used for Alignment

	Model Size	Arch.	Pre-trained Data	Fine-tune Data	Patch Size
MAE (He et al., 2022)	0.30B	ViT	ImageNet-21K	-	16
DinoV2-Giant (Oquab et al., 2023)	1.14B		LVD-142M	-	14
DinoV2-Large	0.30B		LVD-142M	-	14
DinoV2-Base	0.09B		LVD-142M	-	14
DinoV2-Small	0.02B		LVD-142M	-	14
CLIP (Radford et al., 2021)	0.30B		Laion2B	ImageNet-12K	16
AugReg (Steiner et al., 2021)	0.33B		ImageNet-21K	-	16

Table 2: Details of vision models used for alignment. This table summarizes the key properties of various vision models used for alignment, including their architecture, model size, pre-training datasets, fine-tuning datasets, and patch sizes. The models, such as DinoV2, MAE, and CLIP, represent a range of state-of-the-art ViT architectures trained on diverse datasets.

A.5 The Temporal Dynamics of Alignment Scores



(a) Alignment Score across different vision models (b) Alignment Score across different size of DinoV2

Figure 9: (a) Evolution of the alignment score over training steps when incorporating different vision models. (b) Evolution of the alignment score over training steps under different sizes of DinoV2. The plots highlight how alignment quality changes throughout training and illustrate that vision model choice and scale can influence the stability and strength of cross-modal integration.

In Figure 9a, we observe that aligning with different vision models exhibits different trajectories in their alignment scores. While some configurations decrease, others display a more gradual increase or fluctuations before reaching a relatively stable plateau. Besides, we can see how adjusting the size of a single vision model family (DinoV2) affects the trajectories in Figure 9b. Larger models have a higher alignment score, but they do not uniformly guarantee a more rapid increase.

Also, it is worth noticing that all configurations exhibit a drop around 50 steps, indicating a potential early-stage adjustment period due to the competence of the maximum process of alignment score

and the minimum process of cross-entropy loss, whose underlying mechanisms and factors remain unclear. Future research that explores the cases and implications of this early fluctuation can provide deeper insights into the dynamics of cross-modal alignment.

A.6 Experiment Results with Standard Error

	ARC-Challenge ↑	ARC-Easy ↑	COPA ↑	LAMBADA ↑	SWAG ↑	WikiText ↓
Baseline	18.09 ± 0.29	40.06 ± 0.53	62.62 ± 0.92	26.14 ± 0.39	37.65 ± 0.14	50.85 ± 0.36
DinoV2-Small	18.82 ± 0.23	40.31 ± 0.21	62.75 ± 1.32	26.42 ± 0.73	37.83 ± 0.08	50.65 ± 0.19
DinoV2-Base	18.12 ± 0.13	40.38 ± 0.29	63.43 ± 1.25	27.03 ± 0.59	37.82 ± 0.05	50.10 ± 0.19
DinoV2-Large	18.46 ± 0.16	**41.35 ± 0.18**	63.38 ± 0.42	26.58 ± 0.47	37.79 ± 0.07	**49.86 ± 0.17**
DinoV2-Giant	**18.99 ± 0.31**	41.13 ± 0.25	63.71 ± 0.78	26.79 ± 0.63	**37.89 ± 0.07**	50.05 ± 0.10
MAE	18.34 ± 0.20	40.63 ± 0.09	63.17 ± 0.48	26.85 ± 1.02	37.88 ± 0.09	50.21 ± 0.28
AugReg	18.47 ± 0.18	40.94 ± 0.22	62.71 ± 1.15	**27.15 ± 0.39**	37.84 ± 0.04	49.88 ± 0.18
CLIP	18.43 ± 0.21	40.87 ± 0.20	**64.11 ± 0.45**	27.14 ± 0.42	37.77 ± 0.06	49.98 ± 0.17

Table 3: Model Performance Results with Standard Error.

A.7 Ablation Study: Different Kernel Metrics Comparison

Below is the information of each metric shown on x-axis on Figure 7(a):

- **CKA (RBF)** (Kornblith et al., 2019a): A non-linear variant using an RBF kernel, which captures higher-order similarities between features.
- **Linear CKA** (Kornblith et al., 2019a): Measures similarity via the Hilbert-Schmidt Independence Criterion (HSIC) with a linear kernel, capturing invariant relational structures.
- **SVCCA** (Raghu et al., 2017a): Combines SVD and Canonical Correlation Analysis to identify shared subspaces in reduced representations.
- **Dot Product:** Computes the average elementwise dot product between aligned features.
- **Linear Regression** Fits a multi-output linear regression and evaluates using the average R^2 score.
- **PWCCA** (Morcos et al., 2018): Enhances CCA by weighting each canonical correlation according to its explained variance, reflecting component-wise importance.