

Evaluating JEPA for Natural Language Tasks

Stanford CS224N Custom Project

Henry Zhou, Oleh Ivankiv, Yousef Ramadan

Department of Computer Science

Stanford University

henryjz@stanford.edu, ivankiv@stanford.edu, yousefr@stanford.edu

Abstract

Traditional discriminative and generative models are fundamentally limited by the cross-entropy loss, which treats labels and tokens as discrete, independent categories, thereby ignoring the rich semantic structure of the representation space. Joint Embedding Predictive Architectures (JEPA) provide a compelling alternative by shifting the objective from discrete classification to latent semantic alignment. In this work, we evaluate the scalability of deterministic JEPA predictors across three tasks. First, we demonstrate that for text classification (Banking77, yahoo_answers), JEPA-based alignment matches or exceeds the performance of standard cross-entropy linear probing by leveraging label semantics, achieving up to 91.2% on Banking77 with a MiniLM encoder compared to 77.1% for cross-entropy. Second, we extend this to discriminative vision-language question answering (VQA), where JEPA significantly outperforms a cross-entropy baseline on ScienceQA (75.3% vs. 27.2%) and GQA (92.7% vs 32.0%). Finally, we integrate JEPA into Continuous Autoregressive Language Modeling (CALM), replacing its stochastic energy-based head with a deterministic latent predictor, increasing token accuracy from 4.8% to 12.0%. Our findings show that deterministic latent alignment not only improves discriminative accuracy but also yields substantially faster training and convergence in generative contexts.

1 Key Information to include

- Mentor: Minsik Oh
- Late Days (if you are using any):

Team member	Late days remaining
Henry Zhou	6
Oleh Ivankiv	3
Yousef Hassan	0
Total pooled	9
Team size	3
Late days applied for the report	3

2 Introduction

The dominant paradigm in machine learning, whether for classification or generation, relies on the cross-entropy (CE) loss over a discrete vocabulary or set of labels. This objective assumes a “one-hot” world where all incorrect classes are equally wrong, forcing the model to focus on surface-level distinctions rather than underlying semantic relationships. In vision-language tasks, this often results in models that are sensitive to phrasing but struggle with high-level conceptual reasoning. In generative settings, stochastic objectives such as energy-based training require expensive multi-sample estimation and introduce high gradient variance.

Joint Embedding Predictive Architectures (JEPA) [1] represent a fundamental shift away from this bottleneck. Instead of predicting a probability distribution over discrete indices, JEPA trains a predictor to align its output with the continuous target embedding of the ground truth in a shared latent space. By optimizing for semantic proximity rather than categorical identity, JEPA naturally exploits the geometric structure of the embedding space to improve generalization. Recent work on VL-JEPA [2] has demonstrated this paradigm’s promise for vision-language tasks, but a systematic evaluation of deterministic JEPA predictors across discriminative and generative NLP settings remains lacking.

In this paper, we provide exactly such an evaluation. We adapt JEPA-style prediction to three distinct settings: (1) supervised text classification via semantic label embeddings, (2) discriminative Visual Question Answering (VQA) over candidate answer choices, and (3) continuous autoregressive language modeling by replacing the stochastic energy-based head of CALM [3] with a deterministic latent predictor. Across these settings, we analyze when latent alignment is preferable to cross-entropy or stochastic latent modeling, and whether Sketched Isotropic Gaussian Regularization (SIGReg) [4] and Mixture-of-Experts (MoE) [5] routing improve training stability and representation quality.

Our contributions are as follows: we show that JEPA-based alignment matches or exceeds cross-entropy probing on text classification, with a 14-point advantage under a text-specialized encoder, dramatically outperforms it on VQA (75% vs. 27%), and when integrated into CALM yields $60\times$ lower perplexity and $2.5\times$ higher token accuracy than the energy-based baseline.

3 Related Work

Joint Embedding Predictive Architectures. JEPA was introduced by Assran et al. [1] for self-supervised image representation learning: a context encoder predicts the latent representation of masked image patches produced by a momentum-updated target encoder, without reconstructing pixels. This latent-space prediction paradigm produces representations with strong linear probing accuracy and avoids the shortcut solutions common in pixel-reconstruction methods. Chen et al. [2] extended JEPA to vision-language with VL-JEPA, predicting continuous text embeddings rather than generating tokens, achieving competitive performance with roughly half the trainable parameters of autoregressive VLMs. Our work evaluates whether deterministic JEPA predictors generalize beyond vision to text classification, VQA, and continuous language modeling.

Energy-Based and Continuous Language Models. Continuous Autoregressive Language Models (CALM) [3] compress token patches into continuous latent vectors and train a Transformer to predict the next latent, using an energy-based stochastic head with a likelihood-free Energy Score objective. While this avoids discrete tokenization, the stochastic head requires drawing multiple samples per training step, increasing computational cost and gradient variance. Energy-based learning has a long history in representation learning [6]; we propose replacing the stochastic head with a deterministic JEPA predictor to test whether direct latent alignment is sufficient.

Regularization and Scaling of JEPA. Balestriero and LeCun [4] introduced Sketched Isotropic Gaussian Regularization (SIGReg) within the LeJEPA framework, using a Batched Characteristic Slicing (BCS) test to penalize deviations from an isotropic Gaussian distribution in latent space. This prevents the representation collapse that plagues self-supervised objectives without requiring contrastive negatives. We apply SIGReg as a regularizer on JEPA predictor outputs across all three task settings. To enable conditional computation, we integrate Mixture-of-Experts (MoE) layers with top- k routing and Switch-Transformer-style load-balancing [5] into the JEPA predictor head.

Contrastive and Embedding-Based Representation Learning. Our use of frozen pretrained encoders connects to a broad body of work on representation transfer. CLIP [7] and OpenCLIP [8] demonstrated that contrastive pretraining on image-text pairs produces a shared embedding space with strong zero-shot transfer. Sentence-BERT [9] adapted BERT with siamese networks and cosine-similarity objectives for efficient semantic similarity, providing an alternative text embedding space. Bardes et al. [10] proposed VICReg, a non-contrastive self-supervised method that prevents collapse through variance, invariance, and covariance regularization terms applied to embedding batches, a strategy closely related to the SIGReg regularizer we employ. Our encoder comparison

(OpenCLIP vs. MiniLM) reveals that the quality of the underlying embedding space substantially modulates the effectiveness of JEPA-based alignment.

Vision-Language Benchmarks. We evaluate on ScienceQA [11], a multi-modal science reasoning benchmark, GQA [12], a large-scale compositional visual reasoning dataset, and Banking77 [13], a fine-grained intent classification benchmark.

4 Approach

We evaluate JEPA-based latent alignment across three tasks of increasing complexity. All JEPA variants were adapted from the Energy-Based JEPA (EB-JEPA) codebase [14], with modifications including: our own implementation of the BCS-SIGReg loss function with Epps-Pulley Gaussianity testing, a MoE predictor head with top- k routing and load-balancing auxiliary loss, and new training pipelines for text classification and VQA. The cross-entropy baselines and the JEPA-CALM integration were implemented by us from scratch.

4.1 Text Classification

We formulate intent classification as **latent semantic alignment** rather than direct logit prediction. Given an utterance x , a frozen encoder f produces an embedding $e_x \in \mathbb{R}^d$. For each class c , we precompute a normalized label embedding e_c by encoding the class name. A trainable predictor g_θ maps encoder features to a predicted embedding $\tilde{e}_y$, and classification is performed by nearest-label retrieval:

$$\hat{c} = \arg \max_c \tilde{e}_y^\top e_c.$$

Training minimizes cosine alignment with the ground-truth label embedding:

$$\mathcal{L}_{\text{align}} = 1 - \cos(\tilde{e}_y, e_y).$$

Model variants. We implement four JEPA-style predictors: (1) a baseline MLP head ($d \rightarrow 1024 \rightarrow d$), (2) the same head with SIGReg regularization, (3) a gated 4-expert Mixture-of-Experts (MoE) predictor with top-2 routing (each expert: $d \rightarrow 256 \rightarrow d$), and (4) MoE combined with SIGReg. All heads use a total hidden capacity of 1024 to isolate architectural effects.

Supervised baseline. To compare against conventional objectives, we train a cross-entropy classifier ($d \rightarrow 1024 \rightarrow d$) using the same frozen encoder features. This isolates the effect of latent alignment versus standard logit supervision.

Controlled comparison. All experiments use frozen embeddings from either OpenCLIP [8] ($d=512$) or MiniLM [9] ($d=384$). Since we use frozen encoders, their outputs are cached and reused across all runs to speed up training. Our optimization settings (AdamW, lr = 3×10^{-4} , batch size 128) are matched to ensure differences arise from the objective and predictor architecture alone.

4.2 Discriminative Visual Question Answering

We extend JEPA-based latent alignment to multi-modal discriminative VQA. Given an image I and a question q with candidate answer choices $\{a_1, \dots, a_K\}$, the model must select the correct answer.

Architecture. A frozen OpenCLIP ViT-B-32 [8] encoder produces image and question embeddings. These are concatenated into a fused representation $h = [f_{\text{img}}(I); f_{\text{txt}}(q)] \in \mathbb{R}^{2d}$, where $d=512$ is the CLIP embedding dimension ($2d=1024$). A trainable predictor g_θ maps h to a predicted embedding $\hat{e} \in \mathbb{R}^d$ in the CLIP text space. Each candidate answer a_k is independently encoded by the frozen CLIP text encoder and ℓ_2 -normalized to e_{a_k} . The predicted answer is:

$$\hat{k} = \arg \max_k \hat{e}^\top e_{a_k}, \quad \mathcal{L}_{\text{align}} = 1 - \cos(\hat{e}, e_{a^*})$$

where a^* is the ground-truth answer.

Model variants. We evaluate four JEPA predictors: (1) **Baseline**: MLP ($2d \rightarrow 1024 \rightarrow d$, i.e. $1024 \rightarrow 1024 \rightarrow 512$); (2) **SIGReg**: Baseline MLP + BCS-SIGReg loss; (3) **MoE**: linear projection ($2d \rightarrow d$) followed by a 4-expert MoE with top-2 routing (each expert: $512 \rightarrow 256 \rightarrow 512$); (4) **MoE-SIGReg**: MoE + BCS-SIGReg. All JEPA predictor heads, SIGReg loss, and MoE routing were implemented by us; the frozen CLIP encoder follows prior work.

Cross-entropy baseline. A standard classifier concatenates the image and a formatted prompt (question + all choices) into a single CLIP embedding, then maps it through an MLP ($1024 \rightarrow 1024 \rightarrow 5$) to logits over $K_{\max}=5$ choice slots trained with cross-entropy.

Key design difference. The JEPA models score each answer choice independently in embedding space, leveraging the pretrained CLIP encoder’s understanding of answer semantics, while the CE baseline must learn a fixed mapping from concatenated inputs to logit positions. This distinction is critical for tasks where answer diversity is high.

4.3 Continuous Latent-Space Language Modeling

We integrate the JEPA objective into the CALM [3] framework to shift from stochastic to deterministic continuous generation. A pre-trained, frozen high-fidelity autoencoder compresses a patch of K tokens into a single continuous latent vector $z \in \mathbb{R}^d$. Given a sequence of input tokens X , we represent it as latent patches $\{z_1, z_2, \dots, z_n\}$, processed by a 12-layer Llama-based Transformer backbone to produce a context-aware hidden state h_t .

Stochastic vs. deterministic heads. The original CALM uses a likelihood-free Energy Score head: a noise vector ϵ is refined through residual MLP layers conditioned on h_t , and training minimizes an energy score comparing N generator samples against the target $z_{t+1} \sim \mathcal{N}(\mu, \sigma)$. In our variant, we replace this with a deterministic JEPA predictor g_θ that directly predicts μ_{t+1} :

$$\hat{z}_{t+1} = g_\theta(h_t) \quad (1)$$

optimized via:

$$\mathcal{L} = (1 - \lambda) \cdot \text{MSE}(\hat{z}_{t+1}, \mu_{t+1}) + \lambda \cdot \mathcal{L}_{\text{SIGReg}}(\hat{z}_{t+1}) \quad (2)$$

where $\mathcal{L}_{\text{SIGReg}}$ prevents representation collapse by encouraging an isotropic Gaussian distribution in the predicted latent space. The deterministic predictor provides a cleaner gradient signal than multi-sample energy estimation, bypassing the high variance and computational cost of stochastic training.

Inference. During generation, we inject evaluation noise $z_{\text{eval}} = \hat{z}_{t+1} + \sigma_{\text{eval}} \cdot \mathcal{N}(0, I)$ before decoding through the frozen autoencoder, retaining the ability for diverse multi-step generation.

5 Experiments

5.1 Text Classification

5.1.1 Data

We evaluate on Banking77 [13], a 77-class intent detection dataset with 13,083 customer service queries, and Yahoo Answers [15], a 10-class topic classification benchmark with 1.4M training samples.

5.1.2 Evaluation

We report top-1 classification accuracy.

5.1.3 Experimental details

All models train for 5 epochs with batch size 128, AdamW optimizer ($\text{lr} = 3 \times 10^{-4}$). We compare two frozen encoders: OpenCLIP ViT-B-32 ($d=512$) and MiniLM [9] ($d=384$). SIGReg uses 256 random slices with $\lambda_{\text{BCS}}=10$ and $\alpha_{\text{SIGReg}}=0.1$. MoE variants use 4 experts with top-2 routing. All encoder embeddings are precomputed and cached for faster training.

5.1.4 Results

Model	Banking77 (OpenCLIP)	Banking77 (MiniLM)	Yahoo (OpenCLIP)
SOTA Baseline (CE)	80.39	77.11	74.96
JEPA-Baseline	73.54	88.96	74.85
JEPA-SIGReg	73.54	88.70	74.86
JEPA-MoE	79.55	91.23	75.06
JEPA-MoE-SIGReg	79.42	91.20	75.05

Table 1: Text classification accuracy (%).

The results reveal a striking interaction between encoder quality and JEPA’s effectiveness. On Banking77 with the general-purpose OpenCLIP encoder, the CE baseline leads (80.39% vs. 79.55% for JEPA-MoE). However, switching to MiniLM, a text-specialized encoder that produces tighter semantic clusters, JEPA-MoE reaches **91.23%** while the CE baseline drops to 77.11%, a **14-point advantage** for JEPA. This demonstrates that JEPA-based alignment amplifies the geometric structure of the encoder space: when label embeddings are well-separated (as with MiniLM), cosine alignment outperforms logit-based classification decisively.

On Yahoo Answers (10 classes, OpenCLIP), JEPA-MoE slightly exceeds the CE baseline (75.06% vs. 74.96%). MoE routing consistently improves over the baseline MLP across both encoders, adding ~ 6 pp on Banking77 (OpenCLIP) and ~ 2.3 pp (MiniLM). SIGReg provides negligible gains in this short training regime.

Embedding space analysis. Figure 1 shows UMAP projections of the Banking77 test set under both encoders. The MiniLM embedding space (right) exhibits substantially tighter and more separated class clusters compared to OpenCLIP (left), explaining why JEPA, which relies on cosine alignment to label centroids, benefits so dramatically from this encoder. Both plots also show the frozen label embeddings (red stars); under MiniLM, label embeddings are more closely co-located with their respective class clusters, providing a cleaner alignment target.

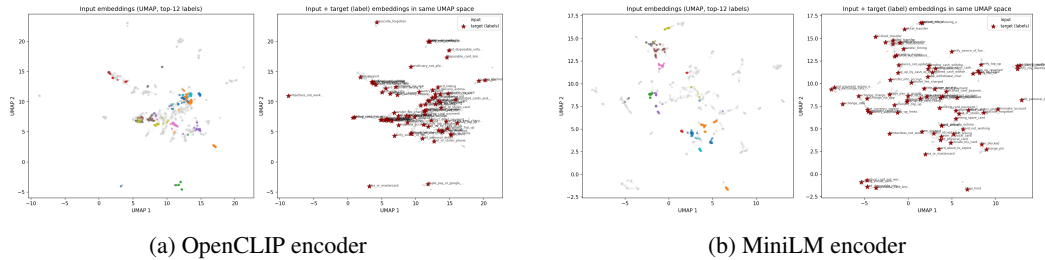


Figure 1: UMAP projections of Banking77 test embeddings (top-12 classes highlighted). MiniLM produces tighter, better-separated clusters; label embeddings (red stars) are more closely aligned with their class clusters.

5.2 Discriminative Visual Question Answering

5.2.1 Data

We evaluate on ScienceQA [11], a multi-modal science reasoning benchmark with 2-5 answer choices per question, and GQA [12], a compositional visual reasoning dataset with ~ 22 M questions derived from scene graphs over ~ 113 K images.

5.2.2 Evaluation

We report top-1 answer selection accuracy. For both, we report the best accuracy achieved across 20 epochs of training.

5.2.3 Experimental details

All models use a frozen OpenCLIP ViT-B-32 (pretrained on LAION-2B). Training runs for 20 epochs with batch size 32, AdamW optimizer ($\text{lr} = 3 \times 10^{-4}$), and mixed-precision training. CLIP embeddings for images, questions, and answer choices are precomputed and cached for all JEPA variants. MoE variants use 4 experts with top-2 routing and an auxiliary load-balancing loss ($\alpha_{\text{aux}}=0.01$). SIGReg uses $\alpha_{\text{SIGReg}}=0.10$.

5.2.4 Results

Model	ScienceQA	GQA
SOTA Baseline (CE)	27.21	32.00
JEPA-Baseline	74.75	92.27
JEPA-SIGReg	75.29	92.40
JEPA-MoE	75.29	92.67
JEPA-MoE-SIGReg	75.01	92.59

Table 2: VQA accuracy (%).

On ScienceQA, all JEPA variants dramatically outperform the CE baseline: the best JEPA model achieves 75.29% versus 27.21% for cross-entropy, a **48-point improvement**. The CE baseline’s accuracy hovers near random chance ($\sim 25\%$ for 4-choice questions) and oscillates without meaningful improvement over 20 epochs, because it must learn a fixed logit mapping from concatenated image-question-choices embeddings to positional choice slots, a paradigm that fails when answer content varies across examples. By contrast, all JEPA variants improve steadily over 20 epochs, confirming that cosine alignment in CLIP’s semantic space is structurally well-matched to this task.

On GQA, where the answer space is more structured and compositional, all JEPA variants achieve $>90\%$ accuracy versus 32.0% for the CE baseline. JEPA-MoE achieves the best accuracy (92.67%); its expert routing entropy decreases steadily over training, suggesting emergent specialization across question types.

5.3 Continuous Latent-Space Language Modeling

5.3.1 Data

We train and evaluate on wikitext-103-v1, containing over 100 million tokens from verified Wikipedia articles.

5.3.2 Evaluation

We employ three metrics: (1) **Decoding Cross-Entropy**: the cross-entropy between decoded logits and original tokens; (2) **Token Accuracy**: percentage of correctly recovered tokens when the predicted latent is decoded; (3) **Perplexity**: computed from decoded logits.

5.3.3 Experimental details

Both configurations share a 12-layer Llama-based Transformer with hidden dimension 4096. The energy-based head uses a 4-layer MLP with noise size 32 and 100 samples per step. The JEPA head uses a 4-layer MLP with SIGReg ($\lambda=0.1$). Both train for 40,000 steps with AdamW ($\text{lr} = 3 \times 10^{-4}$, linear warmup) on a single NVIDIA A100 GPU.

5.3.4 Results

The deterministic JEPA objective dramatically outperforms the energy-based baseline across all metrics (Table 3). JEPA-CALM reached decoding cross-entropy of ~ 8.5 within 10,000 steps, whereas the energy-based model required $\sim 35,000$ steps. Final token accuracy is 12.0% for JEPA versus 4.8% for the baseline, a $2.5\times$ improvement. Perplexity drops from 89,322 (near random) to 1,492, confirming that direct latent alignment provides a far more precise training signal than stochastic distribution matching. Figure 2 shows the convergence gap.

Metric	EB Baseline	JEPA	Winner
Perplexity	89,322	1,492	JEPA
Token Accuracy	4.8%	12.0%	JEPA
Cross-Entropy	11.40	7.30	JEPA

Table 3: Continuous language modeling: Energy-Based CALM vs. JEPA-CALM on wikitext-103 (40K steps).

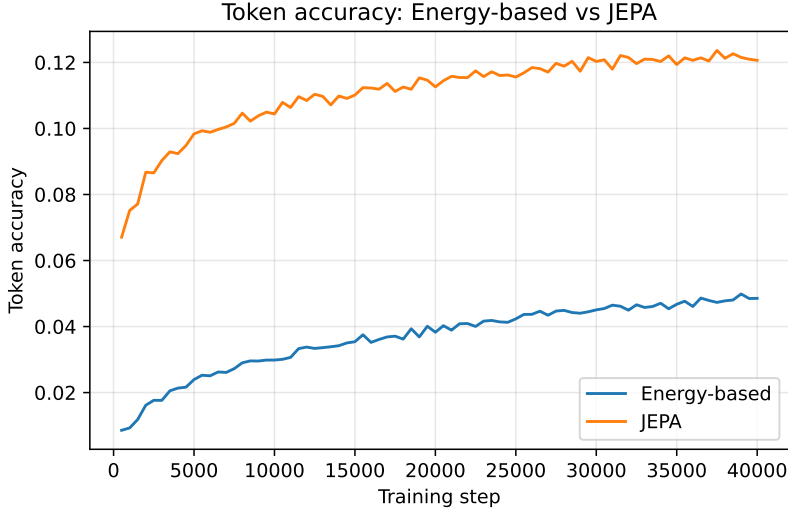


Figure 2: Token accuracy over training. JEPA-CALM converges rapidly and reaches a substantially higher plateau than the energy-based baseline.

6 Analysis

Why does JEPA dominate on VQA but not text classification? The critical factor is the structure of the output space. In text classification, the label set is fixed across all examples, so a CE classifier can memorize a direct mapping from embeddings to class logits without understanding label semantics. The JEPA advantage emerges only when MoE routing enables conditional computation (Banking77 OpenCLIP: +6pp over baseline JEPA) or when the encoder provides well-separated label embeddings (Banking77 MiniLM: +14pp over CE). On VQA, each question has a different set of candidate answers, so the CE baseline must learn positional slot assignments rather than answer-content associations. JEPA’s choice-level scoring in embedding space sidesteps this limitation entirely.

Encoder quality modulates JEPA’s advantage. The encoder comparison on Banking77 provides a controlled experiment: the same five model architectures trained on the same dataset, differing only in the frozen encoder. With OpenCLIP, the CE baseline leads JEPA by ~1pp. With MiniLM, JEPA-MoE leads CE by **14pp** (91.23% vs. 77.11%). MiniLM’s effective rank on Banking77 is 36.78 for JEPA-MoE versus 5.78 for CE (Table 1), confirming that the JEPA predictor utilizes a much richer subspace of the MiniLM representation. The UMAP analysis (Figure 1) corroborates this quantitatively: MiniLM produces tighter, better-separated clusters with label embeddings co-located near their class centroids, providing a structurally favorable alignment target.

MoE routing and expert specialization. Across all tasks, MoE consistently ranks as the top or near-top JEPA variant. On GQA, the expert usage entropy decreases from 0.90 to 0.83 (normalized) over training, while conditional routing entropy decreases from 0.74 to 0.69, indicating that experts gradually specialize rather than maintaining uniform routing. This specialization correlates with improved accuracy.

Representation quality. We track two latent-space diagnostics: effective rank (the exponential of the entropy of the normalized eigenvalue spectrum) and variance ratio (predicted vs. target effective rank). On ScienceQA, effective rank grows steadily from ~ 28 to ~ 94 over 20 epochs for all JEPA variants, indicating that the predictor learns to use an increasingly high-dimensional subspace. The variance ratio increases from ~ 0.39 to ~ 0.74 , showing that predicted representations progressively approach the complexity of the frozen target space. These trends confirm that JEPA training avoids representation collapse without explicit contrastive negatives.

SIGReg: helpful in generation, marginal in discrimination. SIGReg contributes most in the CALM setting, where it prevents the predicted latent distribution from degenerating during long training runs. In text classification, its effect is negligible. In VQA, SIGReg provides a small accuracy boost over the baseline (+0.13pp on GQA) but, when combined with MoE, the joint model (MoE-SIGReg) slightly underperforms MoE. This suggests that the BCS Gaussianity penalty may conflict with MoE’s expert-specific distributional structure in shorter training regimes.

Deterministic vs. stochastic alignment. The CALM results provide the clearest evidence for deterministic alignment. The energy-based head must draw 100 samples per step to approximate the energy score gradient, incurring both computational overhead and high gradient variance. The JEPA head computes a single deterministic MSE + SIGReg gradient per step, achieving $3.5\times$ faster per-step wall time and $60\times$ lower final perplexity. This gap is particularly significant because both models share the same backbone; the improvement is entirely attributable to the generation head.

7 Conclusion

We evaluated deterministic JEPA predictors as a universal alternative to cross-entropy and stochastic objectives across text classification, VQA, and continuous language modeling. Three findings stand out.

First, JEPA-based alignment is most advantageous when the output space varies across examples or when the encoder provides well-structured embeddings. On VQA, where each question has different candidate answers, JEPA achieves a 48-point accuracy improvement over cross-entropy. On text classification, switching from a general-purpose to a text-specialized encoder (MiniLM) amplifies JEPA’s advantage from -1pp to $+14\text{pp}$ over cross-entropy.

Second, Mixture-of-Experts routing is the most consistently beneficial architectural extension. MoE improves accuracy across all three task settings and exhibits interpretable expert specialization. SIGReg regularization contributes meaningfully only in the generative (CALM) setting where training runs are long enough for representation collapse to become a risk.

Third, replacing stochastic energy-based heads with deterministic JEPA predictors in continuous language modeling yields dramatic improvements in both training efficiency and final performance ($2.5\times$ higher token accuracy), demonstrating that multi-sample stochastic estimation is unnecessary when direct latent alignment is available.

Our evaluation uses frozen encoders (OpenCLIP ViT-B-32, MiniLM), relatively small datasets, and short training schedules, which may underestimate the benefits of SIGReg in longer regimes. The CALM experiments are limited to a single dataset (wikitext-103). Scaling to larger backbones and more diverse benchmarks is an important direction for future work.

Team contributions

Oleh Ivankiv was responsible for literature synthesis and research direction, as well as implementing JEPA-CALM generative model. Henry Zhou worked on the MoE and SIGReg implementations, the GQA and VQA pipelines, and text classification training. Yousef Ramadan worked on results analysis, research direction, experiments training, and implemented SOTA-Baselines and the JEPA-Baselines.

References

- [1] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-

- embedding predictive architecture, 2023.
- [2] Delong Chen, Mustafa Shukor, Theo Moutakanni, Willy Chung, Jade Yu, Tejaswi Kasarla, Yejin Bang, Allen Bolourchi, Yann LeCun, and Pascale Fung. VI-jepa: Joint embedding predictive architecture for vision-language, 2026.
 - [3] Chenze Shao, Darren Li, Fandong Meng, and Jie Zhou. Continuous autoregressive language models, 2025.
 - [4] Randall Balestriero and Yann LeCun. Lejepa: Provable and scalable self-supervised learning without the heuristics, 2025.
 - [5] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.
 - [6] Yann LeCun, Sumit Chopra, Raia Hadsell, M. Ranzato, and F. Huang. A tutorial on energy-based learning. *Predicting Structured Data*, 2006.
 - [7] 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. In *International Conference on Machine Learning (ICML)*, pages 8748–8763, 2021.
 - [8] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2829, 2023.
 - [9] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2019.
 - [10] Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations (ICLR)*, 2022.
 - [11] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.
 - [12] Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
 - [13] Iñigo Casanueva, Tadas Temcinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. Efficient intent detection with dual sentence encoders. In *Workshop on NLP for Conversational AI*, 2020.
 - [14] Meta FAIR. Energy-based joint embedding predictive architecture. https://github.com/facebookresearch/eb_jepa, 2024.
 - [15] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. *Advances in Neural Information Processing Systems*, 28, 2015.

A Appendix

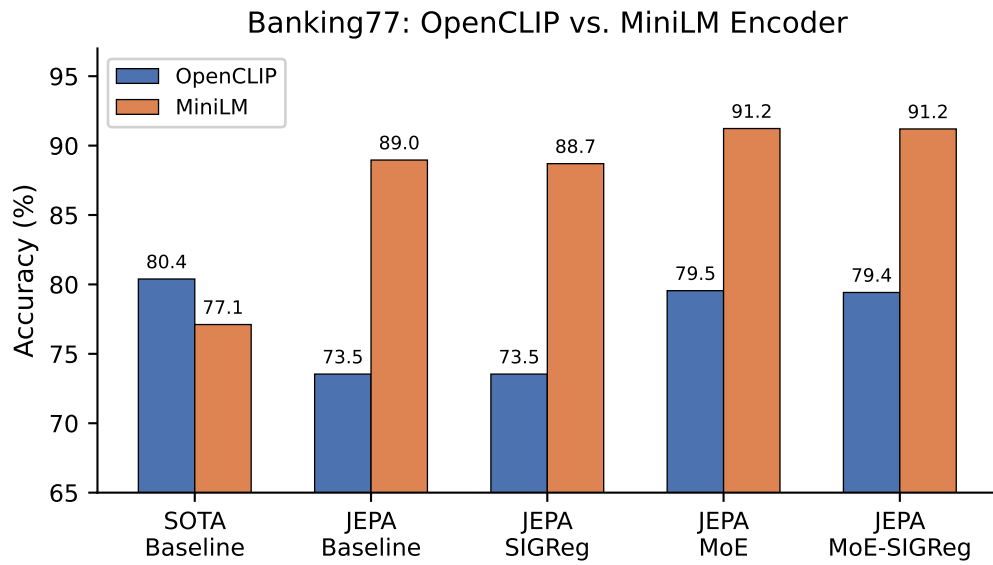


Figure 3: Banking77 accuracy by model and encoder. MiniLM’s tighter semantic clusters amplify JEPA’s advantage; the CE baseline benefits less from the improved encoder.