LeVLJEPA: End-to-End Vision-Language Pretraining Without Negatives

Anonymous authors
Paper under double-blind review

Abstract

Vision-language encoders are increasingly used as frozen backbones for multimodal language models and dense prediction, which consume the grid of patch tokens rather than a single pooled embedding. Image-text pretraining objectives, however, are defined on the pooled embedding, so the patch tokens are shaped only indirectly. We introduce **LeVLJEPA**, the first fully non-contrastive end-to-end vision-language pretraining method. Image and text embeddings predict each other through modality-specific predictors with stop-gradient targets, and collapse is prevented by regularizing each modality with SIGReg. Thus, no negatives, temperature, momentum encoder, or teacher-student schedule are needed. On Datacomp-L, LeVLJEPA matches CLIP and SigLIP under linear probing and is more robust to background shifts, while its zero-shot accuracy is lower. On the patch tokens, which no objective supervises directly, it improves over the strongest contrastive baseline by more than 2 mIoU on both semantic segmentation benchmarks. Used as the frozen visual backbone of a vision-language model, it obtains the highest accuracy on GQA, VQAv2, and POPE with two different language models. These results suggest that non-contrastive image-text pretraining is a viable way to learn visual features for dense prediction and multimodal language models.

1 Introduction

Self-supervised pretraining for vision has increasingly moved toward non-contrastive objectives. Methods such as DINO (Caron et al., 2021), I-JEPA (Assran et al., 2023), and LeJEPA (Balestriero & LeCun, 2025) learn strong visual representations without negative pairs, relying on predictive or distributional criteria instead. Vision-language pretraining has not followed the same path: CLIP (Radford et al., 2021), SigLIP (Zhai et al., 2023), and their extensions still use contrastive image-text alignment as the default objective for learning from paired image-caption data. In this work, we investigate whether non-contrastive prediction can learn image-text representations end-to-end, and how the resulting visual features compare with those learned by contrastive objectives.

Vision-language encoders are increasingly reused as frozen visual backbones, e.g., in visual instruction-tuned models such as LLaVA (Liu et al., 2023), in dense prediction methods such as DenseCLIP (Rao et al., 2022), and in language-conditioned robotics systems such as CLIPort (Shridhar et al., 2022) and Voltron (Karamcheti et al., 2023). In all these cases, the downstream system does not consume a single pooled image vector, but attends over the full grid of patch tokens. The relevant property is therefore the quality of the per-token features, not zero-shot classification alone. Image-text pretraining objective, however, align a single pooled embedding across modalities, which is well suited to zero-shot retrieval and classification, and supervise the patch tokens only indirectly.

In this work, we introduce **LeVLJEPA**, the first fully non-contrastive end-to-end vision-language pretraining method. Image and text embeddings are aligned through cross-modal prediction: a modality-specific predictor maps the image embedding to the text embedding, and vice versa, with the target always detached. To prevent collapse, each modality is regularized independently with SIGReg (Balestriero & LeCun, 2025), which constrains the embeddings to follow an isotropic Gaussian. The resulting recipe needs no negatives, temperature parameter, momentum encoder, or teacher-student schedule. As in contrastive pretraining, the objective operates on a single pooled embedding per modality, and the patch tokens receive no direct supervision. Nevertheless, as we

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

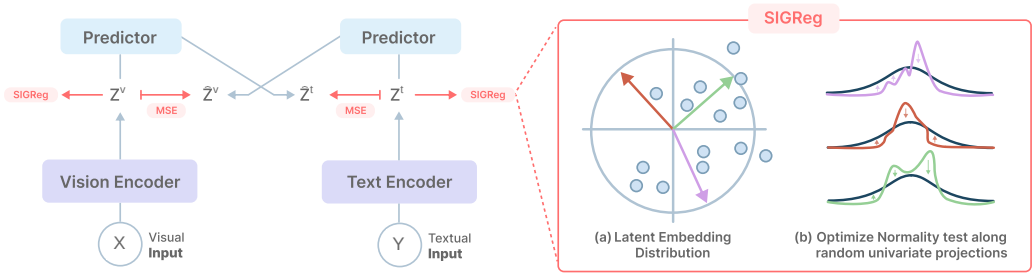


Figure 1: **LeVLJEPA overview.** A batch of images X and their captions Y are encoded into embeddings Z^v and Z^t. Each is passed through a modality-specific predictor that estimates the embeddings of the other modality, $\hat{Z}^t = h_v(Z^v)$ and $\hat{Z}^v = h_t(Z^t)$. The training objective combines (i) a cross-modal MSE loss between each prediction and the stop-gradient target from the other modality, and (ii) SIGReg (Balestriero & LeCun, 2025) applied independently to Z^v and Z^t to keep each modality's marginal embedding distribution close to isotropic Gaussian. SIGReg (right) enforces this distribution by projecting embeddings onto random univariate directions and applying a characteristic-function-based normality test to each projection, avoiding the curse of dimensionality of direct density matching.

show in Section 6, the resulting per-token features are considerably stronger than those of the contrastive baselines.

We validate our method at two scales, training on CC12M (12M samples) (Changpinyo et al., 2021) and on Datacomp-L (92M samples) (Gadre et al., 2023), and we compare it with CLIP and SigLIP trained on the same data with the same backbone. Our contributions are as follows:

- *A fully non-contrastive image-text objective.* LeVLJEPA aligns the two modalities through cross-modal prediction with stop-gradient targets and per-modality regularization. The resulting recipe is simple and stable, without the heuristics of contrastive and teacher-student objectives.
- *Stronger dense features.* Under protocols that read the patch tokens instead of a pooled embedding, LeVLJEPA outperforms CLIP and SigLIP on semantic segmentation (Table 3) and, as the frozen visual backbone of a vision-language model, obtains the highest accuracy in every benchmark–language-model combination on GQA, VQAv2, and POPE (Table 4).
- *Competitive global representations.* On CC12M, LeVLJEPA matches CLIP and SigLIP on zero-shot transfer and linear probing (Appendix F). On Datacomp-L, it remains on par under linear probing (Table 1), is the most robust of the three on the ImageNet-9 background challenge (Table 2), and trails the contrastive baselines only on zero-shot transfer (Table 1), i.e., the objective they optimize directly.

2 Background

Notation. We consider a paired dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ of images $x_i \in \mathcal{X}$ and captions $y_i \in \mathcal{Y}$. An image encoder $f_\theta : \mathcal{X} \to \mathbb{R}^d$ maps images to embeddings $z^v = f_\theta(x)$, and a text encoder $g_\phi : \mathcal{Y} \to \mathbb{R}^d$ maps captions to embeddings $z^t = g_\phi(y)$. A minibatch of size B yields the embedding matrices $Z^v, Z^t \in \mathbb{R}^{B\times d}$. Contrastive baselines ℓ_2-normalize the embeddings before computing cosine similarities. In LeVLJEPA, the embeddings are not ℓ_2-normalized during training: normalization would confine them to the unit sphere, which is incompatible with the isotropic Gaussian in $\mathbb{R}^d$ that SIGReg targets. We ℓ_2-normalize them only at evaluation time, where cosine similarities are needed.

Contrastive vision-language pretraining. CLIP (Radford et al., 2021) trains the two encoders f_θ and g_ϕ jointly with a symmetric InfoNCE objective: matched image-caption pairs are pulled

together, while all other pairs in the batch act as negatives,

$$\mathcal{L}_{\text{InfoNCE}} = -\frac{1}{2B}\sum_{i=1}^{B}\left[\log\frac{\exp(z_i^v \cdot z_i^t/\tau)}{\sum_{j=1}^{B}\exp(z_i^v \cdot z_j^t/\tau)} + \log\frac{\exp(z_i^v \cdot z_i^t/\tau)}{\sum_{j=1}^{B}\exp(z_j^v \cdot z_i^t/\tau)}\right], \tag{1}$$

where τ is a learned temperature and both denominators run over the batch. The quality of this signal depends on the batch size B, since larger batches provide harder and more diverse negatives. For this reason, CLIP-style methods are typically trained with batches of tens of thousands of samples. To reduce the dependency on the batch size, SigLIP (Zhai et al., 2023) replaces the softmax with an independent sigmoid classification of every pair in the batch. Removing the global normalization improves training stability at small and large batches, although performance still benefits from larger B.

VL-JEPA (Chen et al., 2025) recasts vision-language learning as a joint-embedding predictive architecture (JEPA). Instead of aligning two independently pooled embeddings, a predictor maps the image to the continuous embedding of the target caption, learned in latent space and not by autoregressive generation. Despite this framing, the training signal remains contrastive, since VL-JEPA optimizes a bidirectional InfoNCE objective between predicted and target embeddings. Collapse is thus prevented by in-batch negatives rather than by an explicit regularizer, and the method inherits the batch-size dependence of contrastive pretraining. Notably, Chen et al. (2025) observe that the InfoNCE term could in principle be replaced by a sample-independent anti-collapse regularizer, but leave this to future work.

Non-contrastive self-supervised learning. In vision-only pretraining, non-contrastive methods have become the dominant paradigm. DINO (Caron et al., 2021) learns without negatives via self-distillation between a student and a teacher network, while I-JEPA (Assran et al., 2023) predicts the representations of masked image regions from the visible ones. BYOL (Grill et al., 2020) prevents collapse through architectural asymmetry: an online network with a prediction head regresses the output of a momentum-averaged target network, with a stop-gradient on the target. SimSiam (Chen & He, 2021) later showed that the momentum encoder is not needed. This predictor-plus-stop-gradient design is the mechanism we adapt to the cross-modal setting (Section 3).

LeJEPA (Balestriero & LeCun, 2025) replaces these mechanisms with an explicit constraint on the embedding distribution. Its central result is that isotropic Gaussian embeddings minimize the worst-case downstream probing risk, and SIGReg is the objective that enforces this distribution. Instead of matching densities directly in d dimensions, SIGReg projects the embeddings onto random one-dimensional directions and applies a characteristic-function-based normality test (Epps–Pulley) to each projection, at a cost that scales linearly with batch size and embedding dimension. Combined with a view-prediction term, this is sufficient to prevent collapse in the vision-only setting, without stop-gradients, momentum encoders, or teacher-student architectures. We use SIGReg as the non-contrastive component of our method. As we show in Section 3, however, the cross-modal setting additionally requires an asymmetry between the two modalities, which we obtain through predictors with stop-gradient targets.

3 METHODOLOGY: LEVLJEPA

Our goal is to align images and captions without negative pairs, using a single loss and no target network. Given a paired batch with image embeddings $z_i^v = f_\theta(x_i)$ and text embeddings $z_i^t = g_\phi(y_i)$, we need an objective that (i) aligns the two modalities, (ii) regularizes each embedding distribution with SIGReg, and (iii) uses no negatives. Figure 1 gives an overview of the resulting architecture.

Cross-modal prediction with stop-gradient. The most direct adaptation of LeJEPA to this setting, i.e., a symmetric latent regression between matched image-text pairs, is not sufficient in practice. Gradients from the same target flow into both encoders, so the two embeddings meet in a space determined by what image and caption share; since captions describe images at a coarse level of detail, much of the visual content is left unspecified. Adding per-modality SIGReg raises the effective rank of both embeddings but does not recover useful representations (Appendix D). We therefore introduce an asymmetry between each prediction and its target. Instead of regressing one embedding

against the other, we introduce a *cross-modal predictor* on each side: a small MLP $h_v : \mathbb{R}^d \to \mathbb{R}^d$ that predicts the text embedding from the image embedding, $\hat{z}_i^t = h_v(z_i^v)$, and a mirror predictor h_t that predicts the image embedding from the text embedding, $\hat{z}_i^v = h_t(z_i^t)$. Crucially, the prediction targets are detached, so gradients flow only through the predictor and its own encoder:

$$\mathcal{L}_{\text{cross}} = \frac{1}{B} \sum_{i=1}^{B} \Big[\|\hat{z}_i^t - \text{sg}(z_i^t)\|_2^2 + \|\hat{z}_i^v - \text{sg}(z_i^v)\|_2^2 \Big], \tag{2}$$

where $\text{sg}(\cdot)$ denotes the stop-gradient. With detached targets, each encoder receives gradients only from its own branch, and SIGReg regularizes each modality's embedding distribution while the prediction loss aligns the two. The predictors absorb the difference between image and caption content, instead of forcing it into the encoders, similar to the predictor design of BYOL and SimSiam in the vision-only setting.

Architecture. Both encoders follow the standard CLIP setup: a ViT (Dosovitskiy et al., 2020) for images and a GPT-2 (Radford et al., 2019) for captions, with the `[cls]` embedding and the final-token hidden state as the respective representations. Each representation is then passed through a one-layer MLP projection into the shared embedding space of dimension d, where the embeddings z^v and z^t are formed and SIGReg is applied. The projection is needed because the last layers of both backbones apply layer normalization (Ba et al., 2016), which prevents SIGReg from being optimized effectively on the encoder outputs (Maes et al., 2026). The cross-modal predictors h_v and h_t operate on top of these embeddings, so the stop-gradient targets are exactly the embeddings that SIGReg regularizes. Further details are given in Appendix A.

Objective. The full objective combines the cross-modal prediction loss of Equation 2 with SIGReg on each modality:

$$\mathcal{L}_{\text{LeVLJEPA}} = (1 - \lambda_v - \lambda_t)\, \mathcal{L}_{\text{cross}} + \lambda_v\, \text{SIGReg}(Z^v) + \lambda_t\, \text{SIGReg}(Z^t), \tag{3}$$

with regularization weights λ_v and λ_t. The method is trained with a standard forward-backward pass and is compatible with distributed data parallelism.

Training dynamics. In practice, this objective trains stably from initialization: the effective rank of both Z^v and Z^t increases steadily until saturation, the SIGReg terms converge to values consistent with isotropic Gaussian marginals, and the cross-modal loss decreases monotonically. Furthermore, training is largely insensitive to the predictor depth and to the weights λ_v and λ_t over a wide range of values (Appendix E). The ablations of Appendix D isolate the contribution of each component: adding SIGReg to direct image-text regression improves the effective rank, but downstream accuracy remains low, whereas prediction with stop-gradient targets improves accuracy while training remains unstable without SIGReg. Neither component is therefore sufficient on its own, and we combine cross-modal prediction with regularization of both modalities.

Batch size. Contrastive objectives draw their negatives from the batch, through the denominators of Equation 1, so their performance depends on the batch size B, and CLIP-style methods improve as B grows (Radford et al., 2021; Zhai et al., 2023). For LeVLJEPA, cross-modal prediction is computed per sample, and SIGReg acts on the empirical marginal of the embedding distribution, which converges to the same population objective regardless of B. Non-contrastive cross-modal pretraining should therefore be largely insensitive to the batch size, which would remove a constraint that ties contrastive performance to the training infrastructure. We leave an empirical verification of this property to future work.

4 Experimental Setup

Two-stage protocol. Hyperparameter selection and all objective ablations are conducted on CC12M, whose moderate scale allows us to sweep the predictor depth, the SIGReg weights λ_v and λ_t, and the learning rate over many configurations at a tractable cost (Appendix E). The selected configuration is then transferred unchanged to Datacomp-L, where all the results in the main paper are obtained, and no further tuning is performed at that scale.

Baselines. For the contrastive baselines, we use the InfoNCE and SigLIP objectives with the OpenCLIP (Ilharco et al., 2021) training hyperparameters, which are widely validated for CLIP-style training at the scale of Datacomp-L. All methods share the same ViT-B/16 vision encoder, the same data, the same number of samples seen, and the same evaluation protocols, and differ only in the pretraining objective. We refer to the CLIP objective as InfoNCE in all tables and figures.

Scale. All Datacomp-L runs are trained for a total of 819M samples seen rather than for a fixed number of epochs, so that total compute is the same for all objectives. Datacomp-L is distributed as a list of image URLs, and some of them are no longer retrievable: of the approximately 140M image-text pairs of the original release, roughly 92M were still downloadable at the time of our experiments. All methods are therefore trained on this reduced pool. Earlier results on Datacomp-L were obtained from larger downloadable fractions, so our absolute numbers are lower and not directly comparable to them.

Evaluation. All encoders are evaluated frozen, and we group the evaluations into four families. First, zero-shot classification on ImageNet (Deng et al., 2009), Places365 (Zhou et al., 2017a), FGVC-Aircraft (Maji et al., 2013), and Oxford-IIIT Pets (Parkhi et al., 2012), using a single prompt template for all methods to avoid confounding the comparison with prompt engineering. Second, linear probing of the frozen CLS features on the same four datasets, together with the ImageNet-9 background challenge (Xiao et al., 2020), which measures how much the global representation relies on background context. Third, linear semantic segmentation on the frozen patch tokens of ADE20K (Zhou et al., 2017b) and COCO-Stuff (Caesar et al., 2018), where a single linear layer predicts a class at each spatial location. Fourth, frozen-backbone evaluation of a vision-language model, in which both the encoder and the language model are kept fixed and only a lightweight MLP bridge between them is trained, on GQA (Hudson & Manning, 2019), VQAv2 (Goyal et al., 2017), and POPE (Li et al., 2023). The first two families read a single pooled embedding, while the last two operate on the patch tokens. Detailed protocols are reported in Appendix C.

5 Global Representation Quality

We first examine what removing negative pairs costs at the level of the pooled representations. The encoders are evaluated under the two standard protocols for vision-language representations, i.e., zero-shot transfer and linear probing. Both reduce an image to a single vector before scoring and therefore characterize the global representation rather than the per-token features. Furthermore, we measure how much the learned representations rely on background context.

Zero-shot alignment without negatives. The results reported in Table 1 suggest that cross-modal prediction alone is enough to induce image-text alignment. LeVLJEPA classifies far above chance on all four benchmarks, but remains below both contrastive baselines on each of them. On ImageNet it reaches 42.45%, against 47.32% for InfoNCE and 50.78% for SigLIP. The gap is expected, since zero-shot classification measures how well matched pairs are scored above unmatched ones, which is what the contrastive objectives optimize directly.

The size of this gap, however, depends on the scale. On CC12M, at a batch size of 2048, the three objectives are nearly on par, and LeVLJEPA exceeds InfoNCE on most benchmarks (Appendix F). This is consistent with the batch-size dependence of contrastive training: at a moderate batch size, the negative signal that drives zero-shot alignment is weaker and the gap narrows.

Global-feature linear probing. A linear probe on the frozen CLS token isolates the quality of the global image feature from the cross-modal alignment, since it discards the text encoder and only asks whether the pooled embedding is linearly separable by class. Under this readout the three objectives become difficult to separate, and no method dominates (Table 1, bottom): LeVLJEPA stays within 1.5 points of the strongest baseline on every benchmark and is ahead of InfoNCE on Aircraft. The same holds on CC12M (Appendix F). The lower zero-shot accuracy therefore does not correspond to a loss of class information in the visual features.

Background robustness. Linear probing measures whether global features are class-discriminative, but not whether they rely on the object or on the surrounding context. We therefore

Table 1: Zero-shot and linear-probing accuracy (%) on Datacomp-L, with an identical ViT-B/16 backbone. Zero-shot compares the pooled image embedding to a text embedding, while linear probing classifies the frozen CLS features. The three objectives differ on zero-shot transfer, but are on par under linear probing.

Method	ImageNet	Places365	Aircraft	Pets
Zero-shot				
InfoNCE	47.32	**34.46**	8.10	68.98
SigLIP	**50.78**	33.76	**10.62**	**77.27**
LeVLJEPA	42.45	29.97	7.65	59.63
Linear probing				
InfoNCE	65.75	**37.11**	44.10	**82.86**
SigLIP	**66.34**	36.81	**47.46**	82.64
LeVLJEPA	65.42	36.07	46.38	81.28

evaluate the encoders on the ImageNet-9 background challenge (Xiao et al., 2020), where a linear classifier trained on the *Original* split is evaluated, without retraining, on images whose background is replaced by one of the same class (*Mixed-Same*) or of a random class (*Mixed-Rand*). As reported in Table 2, LeVLJEPA is the most accurate in all three settings and the least affected by the background substitution: it loses 5.95 points on Mixed-Same and 17.21 on Mixed-Rand, against 6.57 and 18.67 for InfoNCE and 7.03 and 18.09 for SigLIP. Although the three methods are on par under standard linear probing, the features learned by LeVLJEPA are less sensitive to the background, which suggests that the non-contrastive objective encourages a more object-centric representation.

Table 2: ImageNet-9 background robustness. A linear classifier is trained on frozen CLS features from *Original* images and evaluated without retraining on background-shifted variants. All encoders use the same ViT-B/16 backbone and are pretrained on Datacomp-L.

Method	Original	Mixed-Same	Mixed-Rand
InfoNCE	95.98	89.41	77.31
SigLIP	96.44	89.41	78.35
LeVLJEPA	**96.96**	**91.01**	**79.75**

Across the global readouts, LeVLJEPA therefore matches or exceeds the contrastive baselines everywhere except on zero-shot transfer. The next section evaluates the same encoders on the patch tokens, which downstream systems consume in place of the pooled embedding.

6 Dense Feature Quality

A ViT-B/16 produces a grid of 14×14 patch tokens per image. We evaluate these tokens in two settings, semantic segmentation and visual question answering, keeping the encoder frozen and training only a linear segmentation head or an MLP bridge to a frozen language model. Any difference is therefore attributable to the pretrained features rather than to downstream adaptation.

Semantic segmentation. We train a single linear layer on the frozen patch tokens to predict a class at each spatial location, and report mean intersection-over-union (mIoU) on ADE20K and COCO-Stuff. LeVLJEPA improves over the strongest contrastive baseline by 2.25 mIoU on ADE20K and 2.08 on COCO-Stuff, with the ordering preserved on both benchmarks (Table 3). Since the same single linear layer is trained on every method's tokens, a higher mIoU means that the semantic class is more linearly separable in the token at each location. The difference therefore comes from pretraining rather than from the capacity of the head, and it does not appear under the global readouts of Section 5, where the three objectives were closely matched.

Visual question answering with a frozen backbone. Frozen encoders are increasingly used as the visual front-end of a vision-language model, and we evaluate each method in that setting. We

Table 3: Linear semantic segmentation (mIoU, %) on frozen patch tokens, Datacomp-L with an identical ViT-B/16 backbone. A single linear head is trained per method.

Method	ADE20K	COCO-Stuff
InfoNCE	20.90	29.02
SigLIP	19.24	28.88
LeVLJEPA	**23.15**	**31.10**

freeze both the encoder and the language model and train only an MLP bridge that maps the patch tokens into the language-model input space. Accuracy then depends on the pretrained features and the bridge alone. To separate properties of the visual features from those of a particular language model, we repeat the evaluation with Llama-1B and Qwen-1.5B, and report accuracy on GQA, VQAv2, and POPE. A randomly initialized encoder provides a floor for what the bridge achieves on its own.

LeVLJEPA is the strongest backbone in every benchmark–language-model combination, and all the pretrained encoders exceed the random floor by a wide margin (Table 4). On GQA, which requires compositional reasoning over object attributes and spatial relations, LeVLJEPA exceeds the best contrastive baseline by 1.9 points with Llama-1B and 1.5 with Qwen-1.5B; on VQAv2, a general visual question answering benchmark, the margins are 2.6 and 4.8 points. The two contrastive baselines separate as well, with InfoNCE ahead of SigLIP under both language models, although they are on par on the global readouts of Section 5.

POPE measures object hallucination with balanced yes/no questions about object presence. Because the questions are balanced, a backbone that pushes the model toward answering "yes" scores well on present objects while hallucinating absent ones, and we therefore report the absolute deviation of the yes-rate from 50% as Bias in Table 5. LeVLJEPA is the most accurate under both language models and has the lowest bias, whereas both contrastive backbones are skewed toward "yes". Under Llama-1B, LeVLJEPA's F1 is below that of the baselines, which follows from the same lower yes-rate: removing false positives costs recall on present objects. On POPE, therefore, the advantage of LeVLJEPA is in how rarely it reports absent objects, and not only in overall accuracy.

Table 4: Frozen VLM-backbone accuracy (%) on GQA, VQAv2, and POPE, with two frozen language models. Both the ViT-B/16 vision encoder and the LLM are frozen; only an MLP bridge is trained. All encoders are pretrained on Datacomp-L. Subscripts are standard deviations over three bridge-training seeds.

	Llama-1B			Qwen-1.5B		
Backbone	GQA	VQAv2	POPE	GQA	VQAv2	POPE
Random	$36.4_{\pm 0.6}$	$43.2_{\pm 0.1}$	$49.6_{\pm 1.3}$	$37.0_{\pm 0.7}$	$43.6_{\pm 0.4}$	$52.4_{\pm 0.7}$
InfoNCE	$42.7_{\pm 0.4}$	$51.6_{\pm 0.4}$	$65.8_{\pm 2.4}$	$42.2_{\pm 0.9}$	$49.4_{\pm 0.4}$	$71.5_{\pm 0.4}$
SigLIP	$42.4_{\pm 0.3}$	$49.2_{\pm 0.2}$	$62.0_{\pm 1.5}$	$41.6_{\pm 0.3}$	$47.7_{\pm 0.6}$	$70.4_{\pm 1.7}$
LeVLJEPA	$\mathbf{44.6}_{\pm 0.2}$	$\mathbf{54.2}_{\pm 1.4}$	$\mathbf{66.9}_{\pm 2.6}$	$\mathbf{43.7}_{\pm 0.2}$	$\mathbf{54.1}_{\pm 0.7}$	$\mathbf{75.0}_{\pm 1.2}$

In summary, the ranking of the encoders is unchanged irrespective of the language model, and LeVLJEPA is ahead on every evaluation that reads the patch tokens: linear segmentation on two datasets, and three vision-language benchmarks under two language models. These readouts differ in the task, in the head trained on top, and in whether a language model is involved at all, so the difference is more plausibly a property of the pretrained tokens than of a particular downstream setting. Furthermore, the ordering of the three objectives under zero-shot transfer does not carry over to these evaluations, and is in fact reversed: SigLIP is the strongest zero-shot model and the weakest backbone, whereas LeVLJEPA is the weakest zero-shot model and the strongest backbone.

Table 5: POPE object-hallucination results with frozen encoders and frozen language models. Accuracy and F1 are higher-is-better. Bias is the absolute deviation of the yes-rate from the balanced 50%, $|\text{yes-rate} - 50|$, for which lower is better. Subscripts are standard deviations over three bridge-training seeds.

Backbone	Acc	F1	Bias ↓
Llama-1B			
Random	$49.6_{\pm 1.3}$	$59.3_{\pm 1.2}$	$23.8_{\pm 6.2}$
InfoNCE	$65.8_{\pm 2.4}$	$\mathbf{69.4}_{\pm 2.4}$	$11.6_{\pm 3.1}$
SigLIP	$62.0_{\pm 1.5}$	$68.3_{\pm 1.6}$	$19.7_{\pm 1.8}$
LeVLJEPA	$\mathbf{66.9}_{\pm 2.6}$	$66.6_{\pm 1.7}$	$\mathbf{1.2}_{\pm 1.6}$
Qwen-1.5B			
Random	$52.4_{\pm 0.7}$	$63.8_{\pm 2.4}$	$31.8_{\pm 6.8}$
InfoNCE	$71.5_{\pm 0.4}$	$73.1_{\pm 1.5}$	$5.8_{\pm 4.6}$
SigLIP	$70.4_{\pm 1.7}$	$72.7_{\pm 1.6}$	$8.4_{\pm 6.5}$
LeVLJEPA	$\mathbf{75.0}_{\pm 1.2}$	$\mathbf{75.8}_{\pm 0.9}$	$\mathbf{3.0}_{\pm 1.0}$

7 Discussion and Conclusion

This work indicates that end-to-end vision-language pretraining does not require contrastive negatives. Cross-modal prediction between matched pairs, with stop-gradient targets and a distributional regularizer applied to each modality, trains stably at the scale of Datacomp-L and yields representations that match contrastive pretraining on global-feature readouts. Negative pairs are therefore one way to obtain useful vision-language representations, but not a requirement.

The main advantage of LeVLJEPA appears in the quality of its patch representations. While linear-probing accuracy is similar across methods and contrastive objectives retain an advantage in zero-shot classification, LeVLJEPA performs best on semantic segmentation and frozen-backbone vision-language evaluation (Tables 3 and 4). All three methods apply their pretraining loss to a pooled embedding, so the improvement on dense tasks emerges without an explicit patch-level objective.

These results have a practical implication for how such encoders are compared. Visual instruction tuning, dense prediction, and language-conditioned control read the patch tokens, while all the objectives considered here, and the benchmarks conventionally used to compare them, are defined on the pooled embedding alone. Why the non-contrastive objective yields better patch tokens remains open. One possibility is that the predictors absorb the difference between image and caption content, leaving each encoder free to retain visual detail that the caption does not specify. The background-robustness results of Section 5 are consistent with this, since LeVLJEPA is less sensitive to background substitution than the contrastive baselines even under a global readout (Table 2).

Limitations and future work. Several directions remain open. The contrastive objectives retain a stronger mechanism for zero-shot alignment and we do not close that gap; whether competitive alignment and the dense-feature advantage can be obtained within a single objective is the most direct extension of this work. Our controlled comparisons use a ViT-B/16 backbone on Datacomp-L, and while the objective trains stably at that scale, its behavior at the model and data scales of current vision-language foundation models remains to be characterized. The training signal reaches the patch tokens only indirectly, and whether an explicit dense term improves on the structure that emerges without one has not been tested; the same question applies to the hallucination behavior measured by POPE, where the answer bias we observe is a property of the frozen features that the objective does not address explicitly.

Conclusion. LeVLJEPA shows that end-to-end vision-language pretraining without negative pairs can produce competitive visual representations. Compared with contrastive baselines, it achieves similar linear-probing accuracy, lower zero-shot accuracy, and stronger performance on the dense-feature evaluations considered here. As vision-language encoders increasingly serve as frozen visual backbones, our findings suggest that their patch representations should be evaluated alongside global image-text alignment.

AI USE STATEMENT

We used a large language model as a writing assistant during the preparation of this manuscript. Its use was limited to text we had already drafted: rewriting passages for concision and clarity, checking the consistency of terminology, notation, and cross-references, and identifying typographical errors. The research itself, including the method, the experimental design, the implementation, and the analysis of the results, was carried out by the authors without the assistance of a language model. All text was reviewed by the authors, who take full responsibility for the content of the paper.

REPRODUCIBILITY STATEMENT

The full training protocol, optimizer settings and hardware configuration are given in Appendix A; dataset composition and preprocessing in Appendix B; and the evaluation protocols for linear probing and zero-shot classification in Appendix C. Ablations comparing the components of the objective are reported in Appendix D, and hyperparameter sensitivity in Appendix E. The code to reproduce all experiments is included as a zipped folder in the supplementary material, and will be publicly released upon acceptance.

REFERENCES

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. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15619–15629, 2023.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization, 2016. URL `https://arxiv.org/abs/1607.06450`.

Randall Balestriero and Yann LeCun. Lejepa: Provable and scalable self-supervised learning without the heuristics. *arXiv preprint arXiv:2511.08544*, 2025.

Randall Balestriero, Hugues Van Assel, Sami BuGhanem, and Lucas Maes. stable-pretraining-v1: Foundation model research made simple. *arXiv preprint arXiv:2511.19484*, 2025.

Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1209–1218, 2018.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3558–3568, 2021.

Delong Chen, Mustafa Shukor, Theo Moutakanni, Willy Chung, Jade Yu, Tejaswi Kasarla, Yejin Bang, Allen Bolourchi, Yann LeCun, and Pascale Fung. Vl-jepa: Joint embedding predictive architecture for vision-language. *arXiv preprint arXiv:2512.10942*, 2025.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15750–15758, 2021.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36:27092–27112, 2023.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6904–6913, 2017.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6700–6709, 2019.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021. URL `https://doi.org/10.5281/zenodo.5143773`. If you use this software, please cite it as below.

Siddharth Karamcheti, Suraj Nair, Annie S Chen, Thomas Kollar, Chelsea Finn, Dorsa Sadigh, and Percy Liang. Language-driven representation learning for robotics. *arXiv preprint arXiv:2302.12766*, 2023.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 292–305, 2023.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Lucas Maes, Quentin Le Lidec, Damien Scieur, Yann LeCun, and Randall Balestriero. Leworld-model: Stable end-to-end joint-embedding predictive architecture from pixels. *arXiv preprint arXiv:2603.19312*, 2026.

S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. Technical report, 2013.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 3498–3505. IEEE, 2012.

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, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18082–18091, 2022.

Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Cliport: What and where pathways for robotic manipulation. In *Conference on robot learning*, pp. 894–906. PMLR, 2022.

Kai Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. *arXiv preprint arXiv:2006.09994*, 2020.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11975–11986, 2023.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017a.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 633–641, 2017b.

A Training Protocol

All methods are trained under a matched compute budget and a shared optimization schedule, so that the reported differences reflect the pretraining objective rather than the training configuration. Each model uses the same ViT-B/16 backbone, the same global batch size, the same warmup, and the same cosine learning-rate decay, and is evaluated under an identical protocol. Hyperparameters are selected on CC12M and the resulting configuration is transferred unchanged to the Datacomp-L runs on which the main results are reported. All experiments use stable-pretraining (Balestriero et al., 2025).

Optimization and hardware. Each model is trained on 16 GPUs with a per-rank batch size of 256, for a global batch size of 4096. Training uses a 1,000-step warmup followed by cosine decay. Each run requires approximately 256 H100 GPU-hours.

Learning-rate selection. The learning rate is tuned independently for each method over a small grid on CC12M, and the best configuration under the shared schedule is retained, so that each objective operates at an appropriate step size while the data, backbone, number of updates, warmup, and evaluation protocol are held fixed. At the Datacomp-L scale the contrastive baselines instead use the OpenCLIP (Ilharco et al., 2021) hyperparameters, which are well validated for CLIP-style training at this scale. They are therefore configured as a practitioner would deploy them, and the reported gaps are not an artifact of under-tuning.

LeVLJEPA configuration. LeVLJEPA uses a learning rate of 4×10^{-2}, which we found stable across the batch sizes considered. The vision and text SIGReg weights are set to $\lambda_v = \lambda_t = 0.01$. The cross-modal predictor is a depth-4 MLP of width 2048 with BatchNorm and GELU, and applies 10% dropout to limit overfitting at scale. SIGReg statistics are gathered across devices, as for the contrastive baselines.

B Datasets

Pretraining. We pretrain on two image-text datasets. CC12M (Changpinyo et al., 2021) contains approximately 12M image-caption pairs and is used for hyperparameter selection, ablations, and small-scale validation. Datacomp-L (Gadre et al., 2023) is used for the main results; of its $\sim$140M originally released pairs, $\sim$92M remained downloadable at the time of our experiments due to link rot, and all models are trained on this surviving pool.

Classification (zero-shot and linear probing). We evaluate on ImageNet (Deng et al., 2009) (1000 classes), Places365 (Zhou et al., 2017a) (scene recognition, 365 classes), FGVC-Aircraft (Maji et al., 2013) (fine-grained recognition, 100 classes), and Oxford-IIIT Pets (Parkhi et al., 2012) (37 classes).

Semantic segmentation. ADE20K (Zhou et al., 2017b) contains 20,210 training and 2,000 validation images across 150 semantic classes. COCO-Stuff (Caesar et al., 2018) provides dense annotations over 171 classes. Both are evaluated with a linear head on frozen patch tokens at the 14×14 grid resolution.

Background robustness. ImageNet-9 (Xiao et al., 2020) is a 9-superclass subset of ImageNet with controlled foreground/background recompositions. We use the *Original*, *Mixed-Same*, and *Mixed-Rand* variants, training a linear classifier on *Original* and evaluating on all three.

Vision-language model evaluation. We evaluate the frozen-backbone VLMs on GQA (Hudson & Manning, 2019) (compositional question answering), VQAv2 (Goyal et al., 2017) (general visual question answering), and POPE (Li et al., 2023) (object hallucination, balanced yes/no presence questions).

Table 6: Objective ablation for LeVLJEPA after $20{,}000$ training steps. We report whether SIGReg is used, the effective rank of the text and vision embeddings, and ImageNet zero-shot Top-5 and linear probing Top-1.

Alignment Objective	SIGReg	Eff. rank Text	Eff. rank Vision	ZS Top-5	LP Top-1
Direct MSE	No	25	3	0.52	0.18
Direct MSE	Yes	63	170	1.82	2.56
Predictor + stop-gradient	No	27	161	9.40	14.36
Predictor + stop-gradient	Yes	**358**	**477**	**25.24**	**27.51**

C Evaluation Protocol

Feature extraction. Frozen ViT-B/16 vision features are evaluated without the projection head. For each image we extract the CLS token, yielding a 768-dimensional vector, and apply ℓ_2 normalization. Images are resized so that the shortest side is 256, center-cropped to 224×224, and normalized with the ImageNet mean and standard deviation.

Linear probing. The linear probe is a single linear layer from the frozen feature dimension to the number of classes, without bias. Prior to fitting, features are standardized per dimension using the training-set mean and standard deviation, with standard deviations clamped at 10^{-8}, and the same statistics are applied to the test set. The probe is trained for 50 epochs with AdamW at learning rate 0.1, weight decay 10^{-4}, and batch size 4096, with the training set reshuffled each epoch.

Zero-shot evaluation. For zero-shot classification we use the frozen image and text encoders without fitting any classifier. Unless stated otherwise, class names are embedded with the single prompt template `a photo of a {class}`. Image embeddings are compared against text embeddings by cosine similarity, and the nearest text class is taken as the prediction. We apply this single-template protocol to all methods, which avoids confounding objective-level differences with prompt engineering, prompt ensembling, or dataset-specific calibration. Absolute accuracy is therefore lower than in fully tuned CLIP-style evaluations, and the zero-shot results of Table 1 should be read as a controlled comparison of objectives rather than as an attempt to maximize accuracy.

D Objective Ablations

We ablate the components of LeVLJEPA to test which parts are necessary for stable non-contrastive vision-language pretraining. We compare four variants: direct latent MSE between matched image and text embeddings, direct MSE with per-modality SIGReg, cross-modal predictors with stop-gradient targets but without SIGReg, and the full LeVLJEPA objective. All ablations are run for $20{,}000$ training steps in an earlier, reduced experimental setting. These runs predate the final configuration of Appendix A and used different training hyperparameters, so absolute numbers are not comparable to the main results or to the other appendices. The ablation is intended to compare the variants with each other, all of which share the same setting.

Table 6 shows that symmetric cross-modal regression is not sufficient. Direct MSE collapses almost completely, especially on the vision side, yielding effective rank 3 and near-random ImageNet performance. Adding SIGReg increases the effective rank but does not recover useful image-text alignment or transferable vision features, indicating that marginal regularization alone cannot overcome the degeneracy induced by symmetric regression. Introducing predictors and stop-gradient targets improves zero-shot alignment substantially, but remains unstable and produces weak linear probing without SIGReg. The full objective combines both ingredients: the predictor and stop-gradient asymmetry stabilizes cross-modal alignment, while SIGReg preserves high-rank modality-specific representations, yielding the highest effective ranks and the strongest zero-shot and linear probing performance among the ablated variants.

E HYPERPARAMETER SENSITIVITY

We study LeVLJEPA's sensitivity to two choices, the depth of the cross-modal predictor and the SIGReg trade-off weights λ_v and λ_t. The sweeps were run during method development, before the final training configuration of Appendix A was fixed, and use training hyperparameters that differ from it. Within each sweep, all runs share the same setting and differ only in the ablated dimension, so the sweeps characterize relative sensitivity; absolute accuracies are not directly comparable to the main results. Numbers are ImageNet zero-shot and linear probing Top-1 at batch size 2048 on CC12M.

Predictor depth. Figure 2 sweeps the cross-modal predictor's depth from 1 to 4 hidden layers. A single hidden layer is insufficient: zero-shot drops to 7.4% and linear probing to 26.1%, both well below the configurations with more capacity. Beyond two layers, performance is stable and improves monotonically with depth, gaining 0.7 points on zero-shot and 1.9 points on linear probing between depths 2 and 4. The default configuration of 4 hidden layers used throughout the paper sits within the stable region.

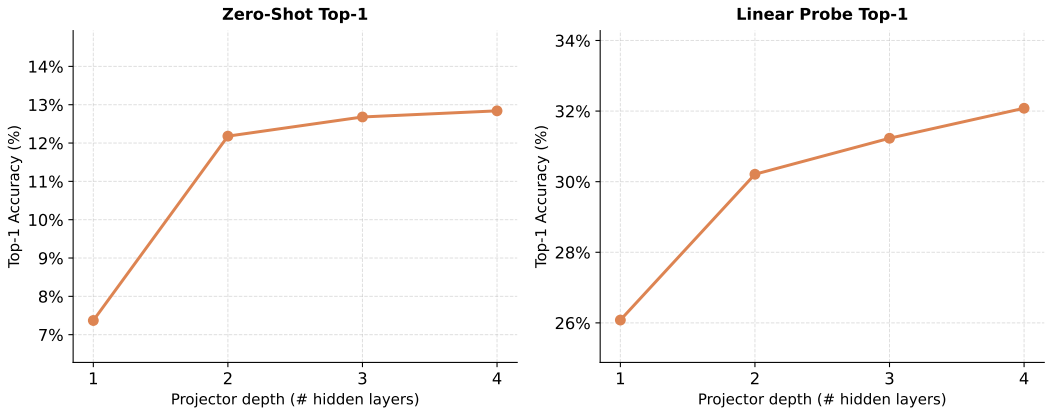


Figure 2: **Sensitivity to predictor depth.** Zero-shot (left) and linear probing (right) Top-1 accuracy as a function of the number of hidden layers in the cross-modal predictor.

SIGReg weight. Figure 3 sweeps λ_v and λ_t symmetrically, denoted λ, across two orders of magnitude. At $\lambda = 0.001$, SIGReg is too weak to enforce isotropy and training degenerates, with zero-shot below 1% and linear probing at 3%. For $\lambda \in [0.005, 0.04]$, performance is stable, with probing accuracy varying within a 4-point range and zero-shot within a 2-point range. LeVLJEPA therefore does not require precise tuning of λ, provided SIGReg is given enough weight to actively shape the embedding distribution.

Together, these sweeps indicate that LeVLJEPA is robust to architectural and optimization choices once two minimum thresholds, two predictor layers and a sufficiently strong SIGReg term, are satisfied.

F CC12M GLOBAL REPRESENTATION RESULTS

This appendix reports the full CC12M results summarized in Section 5. All models use an identical ViT-B/16 backbone and are trained on CC12M under the protocol of Appendix A, except for a global batch size of 2048.

Zero-shot transfer. Table 7 reports zero-shot accuracy across five classification benchmarks. At this scale the three objectives are nearly on par: LeVLJEPA exceeds InfoNCE on STL-10, ImageNet Top-1, CIFAR-100 Top-1, Places365, and SUN397, and matches SigLIP on SUN397 Top-1, while SigLIP retains an advantage on ImageNet and CIFAR-100. LeVLJEPA's largest margins occur on scene classification, where it improves over InfoNCE by 4.6 Top-1 and 6.3 Top-5 on Places365 and

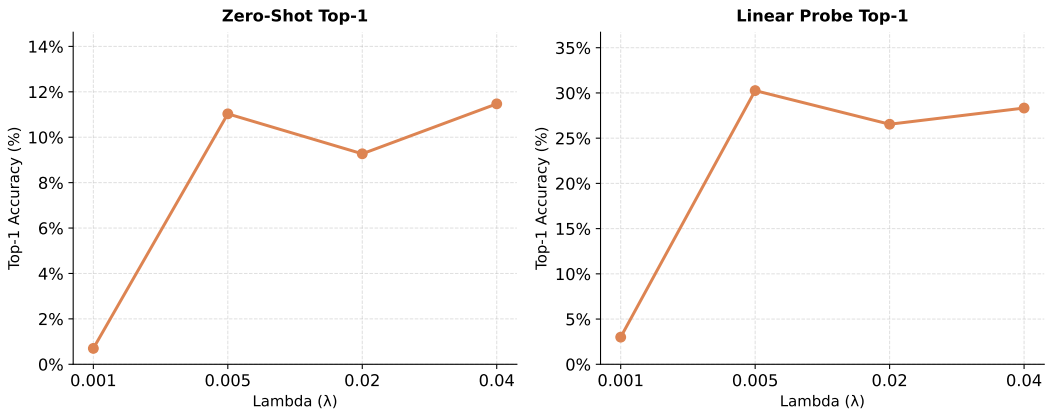


Figure 3: **Sensitivity to the SIGReg trade-off weight λ.** Zero-shot (left) and linear probing (right) Top-1 accuracy as a function of λ on a log scale.

by 2.8 Top-1 and 4.6 Top-5 on SUN397. The contrastive advantage observed on zero-shot transfer at the scale of Datacomp-L (Table 1) is thus substantially reduced in this regime, consistent with the batch-size dependence of contrastive training.

Table 7: Zero-shot classification accuracy (%) on CC12M at batch size 2048 with an identical ViT-B/16 backbone.

Method	**STL-10**	**ImageNet**		**CIFAR-100**		**Places365**		**SUN397**	
	Top-1	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
InfoNCE	68.05	12.39	28.55	9.77	30.74	15.68	38.78	22.43	49.36
SigLIP	68.33	**14.05**	**31.28**	**13.77**	**36.06**	18.40	43.45	**25.27**	**54.78**
LeVLJEPA	**74.16**	12.84	30.23	11.03	29.70	**20.24**	**45.04**	25.27	53.98

Linear probing. Under linear probing of the frozen CLS feature on ImageNet the three objectives are indistinguishable, with LeVLJEPA at 31.84% Top-1, InfoNCE at 31.91%, and SigLIP at 31.83%. This matches the parity observed at the scale of Datacomp-L (Table 1), indicating that the equivalence of the global visual features across objectives is not specific to scale.

G Scaling

The scaling experiments were run during method development and use training hyperparameters that differ from the final configuration, so absolute numbers can differ from those in Appendix F.

G.1 Scaling Model Size

Figure 4 reports ImageNet performance across three ViT backbone sizes (Tiny, Small, Base) under the same update budget. Both methods benefit from increased model capacity. On zero-shot classification, LeVLJEPA outperforms InfoNCE at Tiny (4.90% against 2.26%), is within 0.34 points at Small (9.95% against 10.29%), and slightly exceeds InfoNCE at Base (13.49% against 12.39%). Linear probing shows the same pattern: the two methods are tied at Tiny (22.13% against 22.10%), InfoNCE is ahead at Small (29.37% against 27.43%), and LeVLJEPA slightly exceeds it at Base (32.20% against 31.89%). The non-contrastive formulation therefore scales with model capacity and is not restricted to a single backbone size.

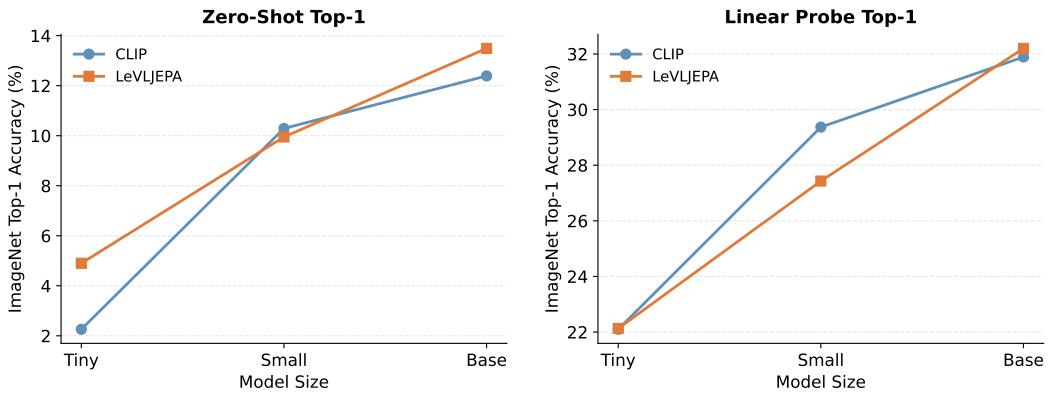


Figure 4: **Model size scaling on CC12M.** LeVLJEPA against InfoNCE on ImageNet zero-shot (left) and linear probing (right) across three ViT backbone sizes: Tiny, Small, and Base.

G.2 Data Efficiency

Figure 5 reports ImageNet zero-shot and linear probing accuracy when both methods are trained on subsets of CC12M under the same update budget. Both methods scale smoothly with dataset size. At smaller data fractions InfoNCE holds a modest advantage: at 25% of CC12M it reaches 5.15% zero-shot and 18.21% linear probing Top-1, against 4.78% and 17.33% for LeVLJEPA, and at 50% it reaches 9.86% and 25.62%, against 8.41% and 23.57%. The gap narrows as data increases. At 75%, LeVLJEPA slightly exceeds InfoNCE on zero-shot accuracy (10.91% against 10.46%), while InfoNCE remains ahead on linear probing (29.54% against 28.53%). At the full dataset the methods are tied on linear probing and LeVLJEPA is slightly ahead on zero-shot accuracy. LeVLJEPA therefore scales competitively with data under the matched training regime.

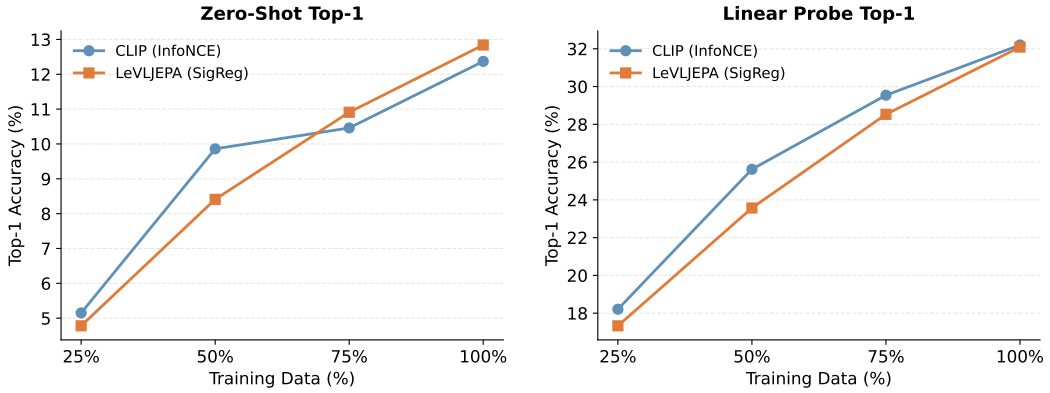


Figure 5: **Data efficiency on CC12M.** LeVLJEPA against InfoNCE on ImageNet zero-shot (left) and linear probing (right) when trained on 25%, 50%, 75%, and 100% of CC12M under the same update budget.