Latent Prediction Needs Alignment: A Controlled Study of Joint-Embedding Predictive Vision–Language Learning

Mohammad Kohankhaki[1,3], **Daniel Kusuma**[1], **Shirin Salehi**[1,3]
Carsten Kamp[2,3], **Sigrid Brell-Cokcan**[2,3], and **Anke Schmeink**[1,3]

[1]Chair of Information Theory and Data Analytics, RWTH Aachen University
[2]Chair of Individualized Production in Architecture, RWTH Aachen University
[3]Cluster of Excellence CARE, TU Dresden and RWTH Aachen, Germany
{mohammad.kohankhaki,shirin.salehi,anke.schmeink}@inda.rwth-aachen.de
{kamp,brell-cokcan}@ip.rwth-aachen.de

Abstract

Predictive vision–language learning aims to learn semantic representations by predicting latent targets rather than reconstructing pixels or generating text. We ask whether latent prediction alone induces the global image–text geometry required for zero-shot matching, and how the location of alignment pressure changes this outcome. In a controlled joint-embedding predictive model family with matched training and evaluation protocols, non-contrastive latent prediction retains information recoverable by supervised adapters and token-level probes but does not form a usable zero-shot image–text space under pooled cosine scoring. A target-contrastive design improves matching, but remains weaker than the directly aligned variant. Our results show that latent prediction is under-constrained at the global embedding level in this setting. Direct global alignment supplies the constraint that makes predictive representations usable for zero-shot scoring. With only a modest direct global alignment loss, latent prediction becomes competitive with matched CLIP and SigLIP baselines and improves several language-sensitive diagnostics.

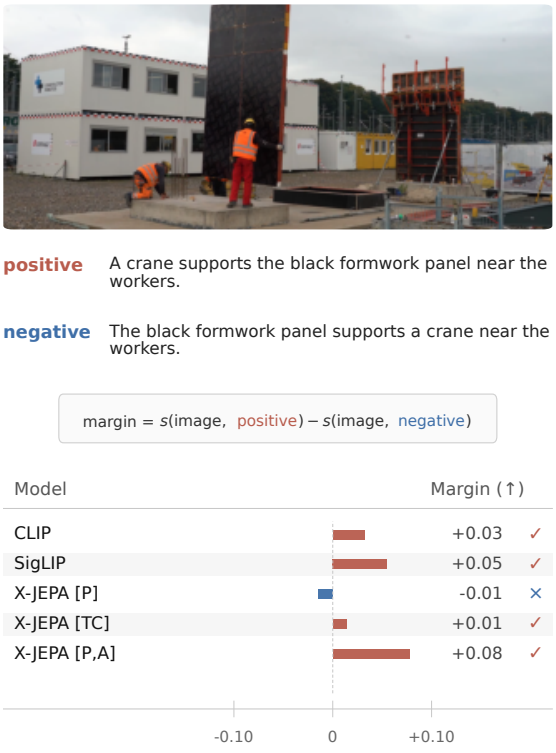


Figure 1: Illustrative hard-negative example from the evaluated model set. Positive values indicate that the correct caption receives a higher score than the hard negative.

1 Introduction

Vision–language models are often evaluated by whether matching images and texts are close in a shared embedding space. Contrastive dual encoders such as CLIP and SigLIP make this effective for retrieval and zero-shot transfer (Radford et al., 2021; Zhai et al., 2023). Yet retrieval-oriented alignment remains coarse: a model may retrieve a plausible caption while failing to bind an attribute, distinguish a spatial relation, or reject a phrase-level hard negative. Language-sensitive diagnostics such as SugarCrepe/SugarCrepe++, SVO-Probes, VSR, and NLVR2 expose such weaknesses across different settings (Thrush et al., 2022; Parcalabescu et al., 2022; Yüksekgönül et al., 2023; Hsieh et al., 2023; Dumpala et al., 2024; Hendricks and Nematzadeh, 2021; Liu et al., 2023; Suhr et al., 2019).

Figure 1 illustrates the kind of distinction hidden by coarse image–text compatibility: both captions are similar at a topical level, but only one preserves the relevant visual relation. The question is therefore not only whether a model retrieves compatible text, but whether its representation exposes the global geometry and token-level information needed to separate hard negatives.

Joint–Embedding Predictive Architectures

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 5210–5222
November 6-10, 2026 ©2026 Association for Computational Linguistics

(JEPAs) provide a complementary principle for learning semantic representations compared to contrastive learning: instead of globally aligning pooled embeddings, they learn by predicting latent representations (LeCun, 2022; Assran et al., 2023). This can encourage semantic structure without output-space generation, but it does not specify how image and text embeddings should be arranged for zero-shot matching. VL-JEPA demonstrates that embedding-space prediction can be extended to vision–language learning when paired with contrastive latent prediction and a broader large-scale setting (Chen et al., 2026). This motivates a narrower controlled question.

We ask how the location of alignment pressure affects predictive vision–language learning. We study this question through a controlled joint-embedding predictive model family called X-JEPA that compares non-contrastive latent prediction, target-contrastive prediction, and latent prediction with direct global image–text alignment under matched training and evaluation protocols. We compare these variants to matched CLIP and SigLIP baselines, which apply alignment pressure directly to pooled global image and text embeddings. We make three contributions:

- We show that prediction-only representations retain information recoverable by supervised readouts but lack a usable pooled-cosine geometry in our setting.
- We show that the target-contrastive design improves zero-shot matching, but the directly aligned variant performs substantially better in our setting.
- We show that modest direct alignment makes predictive representations competitive with matched CLIP and SigLIP baselines and improves language-sensitive diagnostics, while alignment strength is not uniformly monotonic.

Code availability. The code, configuration files, and evaluation scripts are publicly available at `https://github.com/mohkoh19/x-jepa`.

2 Related Work

Contrastive vision–language alignment. Large-scale dual encoders such as CLIP, ALIGN, and SigLIP learn shared image–text spaces by applying contrastive objectives to pooled representations (Radford et al., 2021; Jia et al., 2021; Zhai et al., 2023). Subsequent systems preserve direct alignment while strengthening data pipelines, auxiliary supervision, or alignment granularity, including align-before-fuse training, caption bootstrapping, multi-grained concept alignment, and improved SigLIP encoders (Li et al., 2021, 2022; Zeng et al., 2022; Tschannen et al., 2025). Rather than comparing against such models at scale, we fix training and evaluation protocol across our X-JEPA model family and contrastive baselines, and ask how different placements of alignment pressure affect a predictive objective.

Language-sensitive and compositional diagnostics. Strong retrieval models can still struggle with attribute binding, relational reasoning, word order, and hard negatives. Winoground, VALSE, ARO, SugarCrepe/SugarCrepe++, SVO-Probes, VSR, and NLVR2 expose such weaknesses across controlled linguistic, compositional, spatial, and two-image settings (Thrush et al., 2022; Parcalabescu et al., 2022; Yüksekgönül et al., 2023; Hsieh et al., 2023; Dumpala et al., 2024; Hendricks and Nematzadeh, 2021; Liu et al., 2023; Suhr et al., 2019). A separate line of work improves compositional matching through hard-negative fine-tuning, composition-aware objectives, or inference-time decomposition (Zhang et al., 2024; Jiang et al., 2024; Peleg et al., 2025). Simultaneously, hard-negative metrics still require care because they do not guarantee invariance to meaning-preserving hard positives (Kamath et al., 2024). We use such diagnostics to ask whether different placements of alignment pressure during predictive pretraining change language-sensitive discrimination.

Predictive joint-embedding learning and VL-JEPA. JEPA models predict latent representations rather than reconstructing in input space or aligning global representations (LeCun, 2022; Assran et al., 2023). I-JEPA and V-JEPA show that non-generative feature prediction can learn visual and video representations efficiently and improve semantic understanding (Assran et al., 2023; Bardes et al., 2024). VL-JEPA demonstrates that embedding-space prediction in vision–language can work at scale when paired with contrastive latent prediction and a broader multimodal procedure (Chen et al., 2026). VL-JEPA uses InfoNCE as its contrastive objective, which is widely used for contrastive representation learning in vision–language models (van den Oord et al., 2018). More precisely, it applies contrastive pressure at the predictor–target interface, which is a different placement than direct global alignment like in CLIP-based models.

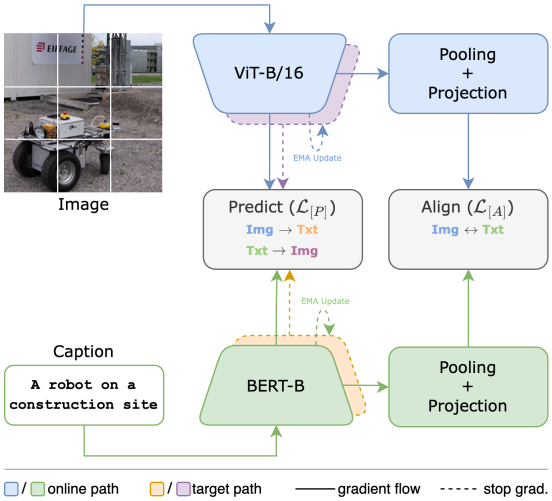


Figure 2: Shared encoder envelope for X-JEPA [P] and X-JEPA [P,A].

Our target–contrastive X-JEPA variant is a same-backbone design point that uses a predictor–target InfoNCE objective that is inspired by VL-JEPA. It is not a reproduction or benchmark comparison to VL-JEPA.

Controlled objective studies and representation bottlenecks. Controlled comparisons are most informative when they separate objective effects from architecture, data, and scale. Recent CLIP–DINO analyses follow this principle to isolate language supervision (Liu et al., 2025). Bottleneck studies further show that weak performance at a pooled interface need not imply absence of internal information: text-only recovery probes reveal limits of single-vector text embeddings, and dual-encoder analyses argue that global cosine matching can hide richer patch/token structure (Kamath et al., 2023; Kang et al., 2025; Miranda et al., 2026). Our adapter and NLVR2 token/global analyses extend this bottleneck view from inference and representation probing to pretraining-objective design by asking which alignment objective makes information usable through global image–text geometry.

3 Method

We study objective design in a controlled family of dual-encoder predictive vision–language models named X-JEPA. We differentiate three variants by where they apply alignment pressure: non-contrastive latent prediction ([P]), target-contrastive prediction ([TC]), and latent prediction with direct global image–text alignment ([P,A]). The backbone family is fixed: each model has a ViT image encoder, a BERT text encoder, pooled global features for retrieval-based scoring, and token features for predictive and probing objectives. Figure 2 schematizes this shared encoder envelope and the [P]/[P,A] prediction-and-alignment paths. The [TC] variant also changes directionality, target-encoder policy, projection heads, and retrieval features. It is a compound design point, not a single-factor alignment-placement ablation.

Shared encoders and representations. For an image–text pair (x_I, x_T), the online image encoder f_θ^I and text encoder f_θ^T produce token sequences and pooled global representations,

$$Z_I, g_I = f_\theta^I(x_I), \qquad Z_T, g_T = f_\theta^T(x_T), \quad (1)$$

where Z_I contains image-patch tokens, Z_T contains text-token representations, g_I is the ViT class-token representation, and g_T is the BERT `[CLS]` representation. All reported X-JEPA variants use a shared cross-modal predictor with modality-specific input and output adapters. Thus the same predictor block is used for image-to-text and text-to-image prediction in [P] and [P,A], while [TC] uses the image-to-text direction only.

[P]: non-contrastive latent cross-modal prediction. The prediction-only variant, X-JEPA [P], tests whether latent cross-modal prediction alone organizes a usable shared space. It uses exponential-moving-average target encoders $f_{\bar\theta}^I$ and $f_{\bar\theta}^T$ to produce stop-gradient latent targets,

$$H_I = f_{\bar\theta}^I(x_I), \qquad H_T = f_{\bar\theta}^T(x_T). \quad (2)$$

The shared predictor maps online tokens from one modality to target-token representations of the other modality. We measure prediction error by the mean squared Euclidean distance over target tokens and hidden dimensions. For a token sequence of length n and hidden dimension d, define

$$\ell_{\text{mse}}(\hat H, H) = \frac{1}{nd}\sum_{j=1}^{n} \|\hat H_j - H_j\|_2^2. \quad (3)$$

The predictive objective averages the two cross-modal directions,

$$\mathcal{L}_{[P]} = \frac{1}{2}\big(\ell_{\text{mse}}(\hat H_T, \text{sg}(H_T)) + \ell_{\text{mse}}(\hat H_I, \text{sg}(H_I))\big), \quad (4)$$

where $\text{sg}(\cdot)$ denotes stop-gradient. For text targets, padding tokens are excluded from the average. No contrastive image–text alignment is used in this variant.

[TC]: target-contrastive prediction. The target-contrastive variant, X-JEPA [TC], replaces the mean-squared prediction objective with a predictor–target InfoNCE loss inspired by VL-JEPA (Chen et al., 2026). [TC] does not use exponential-moving-average target encoders. The direction is only image-to-text: the image encoder produces image tokens Z_I, the shared predictor produces predicted text-token representations $\hat{Z}_T$, and the text encoder produces the paired text-token targets Z_T. For a sample index i in a minibatch of size B, the predicted and target text-space features are

$$\begin{aligned} \hat{u}_{T,i} &= \mathrm{norm}\big(W_p\,\mathrm{pool}(\hat{Z}_{T,i})\big), \\ u_{T,i} &= \mathrm{norm}\big(W_t\,\mathrm{pool}(Z_{T,i})\big), \end{aligned} \tag{5}$$

where pooling over text targets excludes padding positions. Here W_p and W_t are learned projection matrices of the predictor and target alignment heads, respectively. $\mathrm{norm}(\cdot)$ denotes ℓ_2 normalization along the feature dimension. Stacking the minibatch features gives $\hat{U}_T = [\hat{u}_{T,1}, \ldots, \hat{u}_{T,B}]^\top$ and $U_T = [u_{T,1}, \ldots, u_{T,B}]^\top$. We define the target-contrastive loss as

$$\begin{aligned} \mathcal{L}_{[TC]} = \frac{1}{2}\big(&\mathrm{CE}(\hat{U}_T U_T^\top/\tau_{TC}, y) \\ &+ \mathrm{CE}(U_T \hat{U}_T^\top/\tau_{TC}, y)\big), \end{aligned} \tag{6}$$

where y indexes matched pairs and τ_{TC} is a learned temperature. This variant disables direct global image–text alignment and the mean-squared prediction loss. [TC] therefore provides an internal design point for predictor–target contrastive training.

[P,A]: latent prediction with direct global image–text alignment. The aligned predictive variant, X-JEPA [P,A], keeps the latent prediction loss and adds direct global alignment between pooled image and text features in the same way as the matched CLIP baseline. The pooled representations g_I and g_T are mapped to normalized alignment features u_I and u_T by the alignment heads. We define the direct alignment loss as symmetric InfoNCE,

$$\begin{aligned} \mathcal{L}_{[A]} = \frac{1}{2}\big(&\mathrm{CE}(u_I u_T^\top/\tau_A, y) \\ &+ \mathrm{CE}(u_T u_I^\top/\tau_A, y)\big), \end{aligned} \tag{7}$$

where y again indexes matched image–text pairs and τ_A is a learned temperature. The [P,A] objective is

$$\mathcal{L}_{[P,A]} = \mathcal{L}_{[P]} + \lambda \mathcal{L}_{[A]}. \tag{8}$$

Here $\lambda > 0$ is the alignment weight. We vary λ to study the interaction between prediction and alignment pressure. This variant tests whether directly constraining the global image–text geometry makes the predictive representation usable for zero-shot retrieval and hard-negative scoring.

Contrastive baselines. We compare the X-JEPA model family with CLIP and SigLIP (Radford et al., 2021; Zhai et al., 2023). Throughout, CLIP and SigLIP denote matched CLIP and SigLIP baselines with the standard dual-encoder contrastive objectives, but trained and evaluated in the same controlled setup as the X-JEPA variants. CLIP uses direct global symmetric InfoNCE on pooled image and text embeddings, corresponding to the X-JEPA [P,A] variant with $\lambda = 1.0$ and without the latent prediction term. SigLIP uses a sigmoid pairwise contrastive objective over image–text pairs. These baselines provide controlled references for objectives that apply alignment pressure directly to global image–text embeddings without latent prediction.

4 Experimental Setup

Pretraining data and preprocessing. All matched models are pretrained on the same cleaned image–text mixture derived from Conceptual Captions, SBU Captions, MS-COCO, and Visual Genome (Sharma et al., 2018; Ordonez et al., 2011; Lin et al., 2014; Krishna et al., 2017). We remove unavailable images and source images that overlap with any evaluation benchmark used in the paper. The final cleaned mixture contains approximately 5M image–caption pairs. Appendix D gives source-wise counts. Pretraining images use bicubic random resized crops to 224×224, random horizontal flipping, and ImageNet normalization. Evaluation uses deterministic resize, center crop, and the same normalization. Text is tokenized with the BERT-base uncased tokenizer and truncated or padded to length 64.

Model and optimization controls. All compared models use the same backbone scale: a ViT-B/16 image encoder initialized with ImageNet-1K weights and a BERT-base uncased text encoder initialized with official pretrained weights (Dosovitskiy et al., 2021; Devlin et al., 2019). We train the models end-to-end with the same data mixture, image resolution, tokenizer, optimizer family, and training length. The reported models are trained

Model	COCO ZS↑	Flickr30k ZS↑	SugarCrepe++↑	SVO↑	VSR↑	NLVR2-T↑	NLVR2-G↑
CLIP	67.89	79.87	71.68	84.43	63.75	54.93	57.05
SigLIP	67.67	80.32	69.79	84.30	62.77	55.00	57.43
X-JEPA [P]	0.10	0.21	37.06	50.36	48.52	53.05	52.22
X-JEPA [TC]	44.60	48.30	44.17	80.73	57.26	56.42	51.08
X-JEPA [P,A] ($\lambda = 0.10$)	**69.39**	**81.53**	**73.30**	**85.13**	**63.91**	**60.11**	**58.38**

Table 1: Main evaluation results. COCO ZS and Flickr30k ZS report zero-shot mean recall, SugarCrepe++ and SVO report accuracy, and VSR reports AUROC. NLVR2-T/G denote token-interaction and global-embedding probes and report accuracy, respectively. All values are percentages.

for 20 epochs, corresponding to approximately 100.6M seen image–text pairs, with effective batch size 2048 and mixed precision. Optimization uses AdamW with a cosine learning-rate schedule and 4 epochs linear warmup. Encoder learning rates peak at 5×10^{-5}, while newly initialized prediction, projection, and temperature parameters peak at 5×10^{-4}. The comparison therefore varies where alignment pressure is applied, together with the objective-specific heads and interfaces required by each variant. More details are given in Appendix A.

Checkpoint selection. Checkpoint selection is fixed before final testing and uses validation retrieval only. For each pretraining run, we select the checkpoint that maximizes the equally weighted mean recall on COCO Karpathy validation and Flickr30k validation. No final test split or language-sensitive benchmark is used for checkpoint selection.

Evaluation protocols. The evaluation suite is representation-centric: pretrained backbones remain frozen, and no benchmark uses full end-to-end finetuning. We evaluate three access paths.

For zero-shot retrieval, we evaluate COCO Karpathy and Flickr30k test splits (Karpathy and Fei-Fei, 2015; Young et al., 2014). Images and captions are encoded once, normalized, scored by dot product, and ranked in both image-to-text and text-to-image directions. We report mean recall, the average of R@1, R@5, and R@10 over both retrieval directions. For X-JEPA [TC], the image-side retrieval feature is the predicted text-space feature produced from image tokens, and the text-side feature is the encoded text-target feature used by the target-contrastive loss. Adapter retrieval freezes the pretrained encoders and trains only small image/text retrieval adapters and a logit scale, selecting the best checkpoint by validation mean recall.

For language-sensitive zero-shot scoring, SugarCrepe++ requires both positive captions to score above a controlled hard negative. SVO-Probes scores a sentence against a positive and a negative image. VSR evaluates binary image–caption spatial-relation compatibility.

For NLVR2, the *global embedding* probe trains an MLP over pooled left-image, right-image, and text features, while the *token interaction* probe trains a small Transformer over raw image and text token sequences. Both NLVR2 probes keep the pretrained backbone frozen and select the best checkpoint by validation accuracy. All evaluation details are given in Appendix E.

5 Results and Analysis

We organize results by representational access path. Zero-shot retrieval and hard-negative scoring test the pooled global geometry. Adapter tuning tests recoverable information under a frozen backbone. NLVR2 token interaction tests whether raw token features expose information hidden by pooled embeddings. This distinction is essential: weak zero-shot geometry does not imply absence of useful cross-modal information.

Table 1 summarizes the main results across all metrics and models. We report results for X-JEPA [P,A] with $\lambda = 0.10$ in the main results as a representative setting. This is a representative balance across metrics, not a universal optimum (Table 3).

5.1 Prediction Alone Does Not Constrain Global Geometry

Table 1 first tests whether latent prediction alone yields a usable cross-modal embedding space. Under the evaluated pooled-cosine scorer, the answer is negative. X-JEPA [P] is essentially unusable for zero-shot retrieval, while the matched contrastive baselines form strong shared spaces. The same pattern appears outside retrieval, where [P] is also weak on the language-sensitive tasks. Thus, non-contrastive latent prediction does not arrange

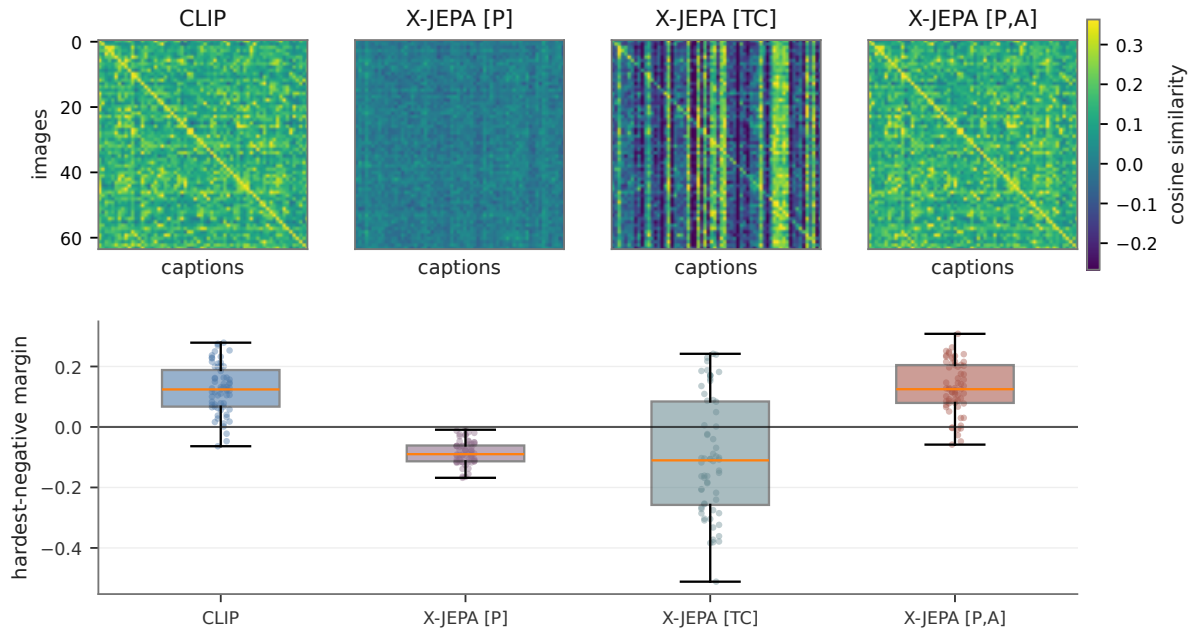


Figure 3: **First row:** image–text similarity matrices on a deterministic COCO retrieval subset. Rows correspond to images and columns to captions, ordered so matched pairs lie on the diagonal. **Second row:** margin summaries report matched-pair separation from hardest negatives.

Model	COCO		Flickr30k	
	ZS↑	Adapter↑	ZS↑	Adapter↑
CLIP	67.89	69.34	79.87	82.50
SigLIP	67.67	68.62	80.32	82.19
X-JEPA [P]	0.10	45.90	0.21	60.61
X-JEPA [TC]	44.60	63.05	48.30	75.06
X-JEPA [P,A]	**69.39**	**69.86**	**81.53**	**83.33**

Table 2: Zero-shot and adapter-tuned retrieval. [P,A] uses $\lambda = 0.10$. All values are mean recall percentages.

pooled image and text embeddings into a retrieval-compatible geometry.

5.2 Adapter Recovery Shows That Information Is Present

Table 2 separates frozen-backbone information from zero-shot geometry. Although X-JEPA [P] has near-zero zero-shot retrieval, adapter tuning recovers substantial COCO and Flickr30k performance, showing recoverability under this supervised readout. Target-contrastive prediction improves zero-shot geometry but still benefits strongly from adapters, whereas X-JEPA [P,A] is already strong before adaptation and gains only modestly. Direct global alignment therefore changes what is immediately accessible through the pooled embedding space, not merely what can be recovered by adding task-specific capacity. Because the encoders start from pretrained unimodal checkpoints and the adapters and probes receive paired supervision, these results do not identify how much of the recovered signal originates in paired predictive pretraining.

5.3 Target-Contrastive Prediction Partially Improves Global Geometry

The X-JEPA [TC] variant forms the intermediate step between prediction-only learning and direct global alignment. Relative to [P], the target-contrastive design substantially improves retrieval, SVO, and VSR, suggesting that contrastive pressure at the predictor–target interface can improve the similarity structure used for matching. This observation is consistent with the larger-scale results reported by VL-JEPA (Chen et al., 2026).

Yet [TC] is not on par with direct global alignment. It remains below [P,A] on all reported metrics and below the contrastive baselines on retrieval, hard-negative scoring, and global NLVR2, with the NLVR2 token-interaction probe as the only exception. Thus, the result is not simply explained by replacing the mean-squared prediction loss with InfoNCE. Where contrastive pressure is applied matters: target-contrastive prediction helps, but does not fully impose the global image–text geometry used by retrieval and hard-negative scoring. However, [TC] changes several design choices together

and its gap to [P,A] cannot therefore be attributed solely to alignment placement or to evaluating an interface different from the one trained.

5.4 Geometry Diagnostics Visualize Global-Space Misalignment

Figure 3 shows the same mechanism in the similarity matrix. CLIP and X-JEPA [P,A] produce a visible matched-pair diagonal, whereas X-JEPA [P] lacks this structure and has negative hardest-negative margins. X-JEPA [TC] partially increases matched-pair structure, but remains uneven. Together with the adapter results, this diagnostic links weak zero-shot retrieval to the pooled global space rather than absence of internal information.

5.5 Direct Global Alignment Makes the Predictive Space Usable

Adding a modest direct global alignment loss produces a strong aligned predictive variant. X-JEPA [P,A] exceeds the matched baselines and the other X-JEPA variants on all metrics reported in Table 1. The largest gains over [TC] occur in metrics that directly use the global image–text space. Thus, under the evaluated pooled scorer, latent prediction is not a replacement for alignment: direct global alignment supplies the topology needed for zero-shot scoring, after which the predictive model also improves the language-sensitive diagnostics central to this study.

5.6 Alignment Strength Is Not Uniformly Monotonic

Table 3 studies alignment strength. The dominant discontinuity is between no global alignment and any positive alignment weight: once $\lambda > 0$, retrieval scores are near or above the matched contrastive baselines. Within the positive range, however, larger alignment pressure is not uniformly better. The smallest positive weight $\lambda = 0.03$ is strongest on SVO and VSR, $\lambda = 0.10$ is strongest on SugarCrepe++ and NLVR2, and $\lambda = 0.30$ is slightly strongest on retrieval. Direct alignment is decisive in this setting, but increasing its weight shifts tradeoffs rather than improving all metrics.

5.7 Token Interaction Separates Information from Global Geometry

The two NLVR2 protocols provide another view of representational access under frozen-probe evaluation. Table 4 compares token-interaction and global-embedding probes. The matched contrastive baselines perform better with global features than with token interaction, while the predictive variants show the opposite pattern. This pattern is clearest for [TC], whose token-interaction score exceeds both contrastive baselines despite a weak global score. [P,A] improves both access paths and achieves the best score under both protocols. This suggests that global geometry, adapter-recoverable information, and token-level information are distinct but related aspects of representation quality under these probes.

5.8 Language Slices Show Where Aligned Prediction Helps

Figure 4 tests whether the aggregate gains are concentrated in one metric or appear across language-sensitive categories. X-JEPA [P] is weak across SugarCrepe++ categories and VSR relation families. [TC] closes much of the gap on SVO and several VSR families, but remains weak on SugarCrepe++ hard negatives. [P,A] yields the most consistent behavior across object, relation, role, and spatial-relation slices in this evaluation, indicating that the aggregate gains are not only retrieval effects.

6 Discussion

The results separate representational content from representational access. Adapter recovery and NLVR2 token interaction show that predictive variants can retain cross-modal information even when pooled image–text matching is weak. This connects our objective study to bottleneck analyses of dual encoders, where weaknesses at a global scorer can reflect how information is exposed rather than whether it is present (Kang et al., 2025; Miranda et al., 2026). In our setting, this bottleneck is not only an inference artifact. It depends on the training objective and on where alignment pressure enters.

The [P,A] objective directly constrains the pooled scoring interface, whereas [P] does not. [TC] also trains its own predictor–target scoring interface and improves matching, but its compound design prevents attributing its gap to [P,A] solely to alignment location.

A simple invariance illustrates the underconstraint. For a freely parameterized directional predictor p, latent vectors z_I, z_T, and orthogonal matrices Q_I, Q_T, let $p_Q(z) = Q_T p(Q_I^\top z)$. Then

$$\|p_Q(Q_I z_I) - Q_T z_T\|_2^2 = \|p(z_I) - z_T\|_2^2,$$

Model	λ	COCO ZS↑	Flickr30k ZS↑	SugarCrepe++↑	SVO↑	VSR↑	NLVR2-T↑	NLVR2-G↑
X-JEPA [P]	0	0.10	0.21	37.06	50.36	48.52	53.05	52.22
X-JEPA [P,A]	0.03	69.11	81.22	71.28	**85.32**	**64.99**	58.56	57.57
	0.10	69.39	81.53	**73.30**	85.13	63.91	**60.11**	**58.38**
	0.30	**69.47**	**81.66**	72.59	85.19	64.58	59.67	58.32
	1.00	69.05	81.35	72.21	84.75	63.86	57.84	57.64

Table 3: Alignment-weight ablation. [P] has no direct global alignment. [P,A] rows share the same MSE prediction loss and vary only the direct global alignment weight λ. VSR reports AUROC and all values are percentages.

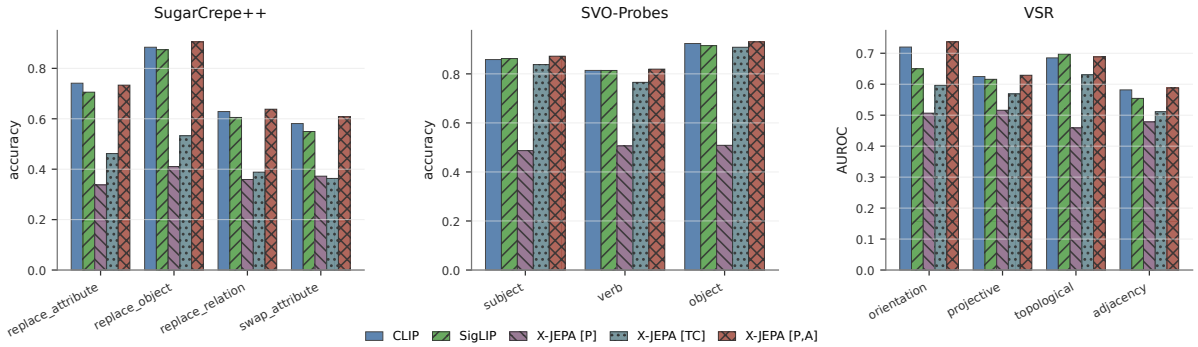


Figure 4: Language-sensitive slice metrics for the evaluated model set. SugarCrepe++ and SVO report accuracy by category or role, while VSR reports AUROC by relation family. Bars use both color and hatch patterns to distinguish models.

Model	Token↑	Global↑	Token–Global
CLIP	54.93	57.05	-2.12
SigLIP	55.00	57.43	-2.43
X-JEPA [P]	53.05	52.22	+0.83
X-JEPA [TC]	56.42	51.08	+5.34
X-JEPA [P,A]	**60.11**	**58.38**	+1.74

Table 4: NLVR2 frozen probes. Token is token-interaction accuracy, Global is global-embedding accuracy, and Token–Global is their difference. [P,A] uses $\lambda = 0.10$. Values are percentages.

while $\cos(Q_I z_I, Q_T z_T)$ need not equal $\cos(z_I, z_T)$. Thus identical prediction errors can coexist with different raw cross-modal scores.

The aligned predictive model changes both evaluated access paths. With a modest direct alignment weight, X-JEPA [P,A] keeps the token-level benefits suggested by NLVR2 while making the global representation usable for zero-shot scoring. Its scores on SugarCrepe++, SVO, VSR, and NLVR2 suggest that once global geometry is constrained, latent prediction can contribute to language-sensitive vision–language representations rather than only improving retrieval.

More broadly, the results argue for reporting which representation interface an evaluation queries. A single pooled score is appropriate for retrieval and hard-negative ranking, but it can hide information available to adapters or token-level modules. Conversely, strong token-level access does not guarantee a retrieval-compatible global space. Predictive vision–language objectives should therefore be assessed across both pooled and token-level access paths.

Future work should test whether richer alignment locations, such as dense or multi-grained predictor–target objectives, can combine the token-level benefits of predictive learning with the global accessibility of direct alignment. Scaling this comparison across larger data mixtures and models would further clarify which effects are stable design principles.

7 Conclusion

Latent prediction needs alignment. In our setting, prediction-only representations retain information recoverable by supervised readouts but lack the pooled-cosine geometry needed for zero-shot matching. The target-contrastive design improves matching, and the directly aligned variant performs substantially better. Once alignment constrains the shared geometry, predictive objectives can match or slightly exceed matched contrastive baselines in this controlled setting. These results suggest that

the value of predictive learning in vision–language models depends not only on what information is retained, but also on whether the objective makes that information accessible through the representation used at inference time. This makes alignment location a central design choice for predictive vision–language learning. Predictive vision–language objectives should therefore be evaluated by both the information they preserve and the interface through which that information becomes usable.

Limitations

This study is a controlled comparison at moderate scale. We fix the data mixture, backbone family, and evaluation protocol to reduce confounds among objective variants, but the results should not be read as state-of-the-art claims against larger VLMs, instruction-tuned systems, or models trained on substantially larger and more diverse corpora.

Each pretraining setting is represented by a single run. We do not establish run-to-run stability or statistical significance for small differences between aligned models or positive alignment weights. The central conclusion rests on the much larger prediction-only versus directly aligned contrast under the specified scorer.

The target-contrastive variant of X-JEPA is an internal design point rather than a pure single-factor ablation: relative to [P] and [P,A], it changes the loss, directionality, target-encoder policy, and retrieval interface. This is why we use it to probe a plausible predictor–target contrastive placement, not to make broad claims about VL-JEPA-like training in general.

The [P] and [P,A] variants use EMA target encoders to stabilize latent prediction. This choice avoids adding auxiliary regularization losses that would confound the alignment-location comparison, but it requires additional target-encoder state, two forward-only target-encoder passes, and bidirectional prediction. [TC] uses one prediction direction without EMA. Matching backbone scale therefore does not establish equal runtime or FLOPs.

Ethical Considerations

This work studies objective design for image–text representation learning and does not introduce a deployed system or collect new human-subject data. The pretraining mixture and evaluation benchmarks are derived from public image–text datasets, which may contain social, geographic, and linguistic biases inherited from their original sources. Improvements in image–text matching and hard-negative discrimination could be useful for retrieval and evaluation, but could also be incorporated into downstream ranking, filtering, or surveillance-like systems without adequate safeguards. We therefore limit our claims to controlled representation behavior under the evaluated protocols and recommend task-specific validation, bias assessment, and human oversight before deployment in consequential settings. Dataset construction, evaluation protocols, and implementation details are documented in the appendix to support reproducibility.

Acknowledgments

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 3115 – 533767731. Computational resources were provided by the German AI Service Center WestAI.

AI tools were used for literature searches, code debugging, and manuscript editing. The authors take full responsibility for the content.

References

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. 2023. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15619–15629.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. 2024. Revisiting feature prediction for learning visual representations from video. *Transactions on Machine Learning Research*. Featured Certification; ICLR 2025 Journal Track.

Delong Chen, Mustafa Shukor, Théo Moutakanni, Willy Chung, Jade Yu, Tejaswi Kasarla, Allen Bolourchi, Yann LeCun, and Pascale Fung. 2026. VL-JEPA: Joint embedding predictive architecture for vision-language. In *International Conference on Learning Representations*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Sri Harsha Dumpala, Aman Jaiswal, Chandramouli Shama Sastry, Evangelos Milios, Sageev Oore, and Hassan Sajjad. 2024. SUGARCREPE++ dataset: Vision-language model sensitivity to semantic and lexical alterations. In *Advances in Neural Information Processing Systems*, volume 37, pages 17972–18018.

Lisa Anne Hendricks and Aida Nematzadeh. 2021. Probing image-language transformers for verb understanding. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3635–3644, Online. Association for Computational Linguistics.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. 2023. SugarCrepe: Fixing hackable benchmarks for vision-language compositionality. In *Advances in Neural Information Processing Systems*, volume 36, pages 31096–31116.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4904–4916. PMLR.

Kenan Jiang, Xuehai He, Ruize Xu, and Xin Wang. 2024. ComCLIP: Training-free compositional image and text matching. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6639–6659, Mexico City, Mexico. Association for Computational Linguistics.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. 2023. Text encoders bottleneck compositionality in contrastive vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4933–4944, Singapore. Association for Computational Linguistics.

Amita Kamath, Cheng-Yu Hsieh, Kai-Wei Chang, and Ranjay Krishna. 2024. The hard positive truth about vision-language compositionality. In *Computer Vision – ECCV 2024*, pages 37–54. Springer.

Raphi Kang, Yue Song, Georgia Gkioxari, and Pietro Perona. 2025. Is CLIP ideal? no. can we fix it? yes! In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22436–22446.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3128–3137.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73.

Yann LeCun. 2022. A path towards autonomous machine intelligence. Version 0.9.2, 2022-06-27.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. 2022. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR.

Junnan Li, Ramprasaath R. Selvaraju, Akhilesh D. Gotmare, Shafiq Joty, Caiming Xiong, and Steven C. H. Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. In *Advances in Neural Information Processing Systems*, volume 34, pages 9694–9705.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755. Springer.

Fangyu Liu, Guy Emerson, and Nigel Collier. 2023. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651.

Yiming Liu, Yuhui Zhang, Dhruba Ghosh, Ludwig Schmidt, and Serena Yeung-Levy. 2025. Data or language supervision: What makes CLIP better than DINO? In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 1868–1874, Suzhou, China. Association for Computational Linguistics.

Imanol Miranda, Ander Salaberria, Eneko Agirre, and Gorka Azkune. 2026. Revisiting compositionality in dual-encoder vision-language models: The role of inference. *Preprint*, arXiv:2604.11496.

Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. 2011. Im2Text: Describing images using 1 million captioned photographs. In *Advances in Neural Information Processing Systems*, volume 24, pages 1143–1151.

Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. 2022. VALSE: A task-independent benchmark for vision and language models centered on linguistic

phenomena. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8253–8280, Dublin, Ireland. Association for Computational Linguistics.

Amit Peleg, Naman Deep Singh, and Matthias Hein. 2025. Advancing compositional awareness in CLIP with efficient fine-tuning. In *Advances in Neural Information Processing Systems*, volume 38.

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. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image Alt-Text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2019. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6418–6428, Florence, Italy. Association for Computational Linguistics.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5238–5248.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. 2025. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *Preprint*, arXiv:2502.14786.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *Preprint*, arXiv:1807.03748.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Mert Yüksekgönül, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. 2023. When and why vision-language models behave like bags-of-words, and what to do about it? In *International Conference on Learning Representations*.

Yan Zeng, Xinsong Zhang, and Hang Li. 2022. Multi-grained vision language pre-training: Aligning texts with visual concepts. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 25994–26009. PMLR.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11975–11986.

Le Zhang, Rabiul Awal, and Aishwarya Agrawal. 2024. Contrasting intra-modal and ranking cross-modal hard negatives to enhance visio-linguistic compositional understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13774–13784.

A Architecture and Optimization Details

The image encoder is a ViT-B/16 initialized with ImageNet-1K weights, with 12 Transformer layers, hidden size 768, 12 attention heads, 16×16 patches, and approximately 85.8M trainable parameters (Dosovitskiy et al., 2021). The text encoder is BERT-base uncased initialized with official pretrained weights, with 12 layers, hidden size 768, 12 attention heads, maximum length 64, and approximately 109.5M trainable parameters (Devlin et al., 2019). All models use the BERT-base uncased tokenizer. All runs used 8 NVIDIA H100 GPUs. Shared training controls are reported in Table 5.

Setting	Value
Training length	20 epochs
Seen image–text pairs	$\approx$ 100.6M
Effective batch size	2048
Precision	bf16 mixed
Optimizer	AdamW
Betas / epsilon	$(0.9, 0.98)$ / 10^{-6}
Weight decay	0.01
LR schedule	cosine with 4-epoch linear warmup
Final LR	10^{-6}
Encoder peak LR	5×10^{-5}
Head/temp. peak LR	5×10^{-4}

Table 5: Shared pretraining controls for the matched comparison.

The X-JEPA variants use the shared predictor described in Section 3. This predictor has width 384, depth 4, and 2 attention heads, contributing approximately 8.28M trainable parameters. Fixed sinusoidal predictor position embeddings add 0.10M non-trainable parameters. The non-contrastive predictive variants use EMA target encoders with a linear momentum schedule from 0.996 to 1.0. The target-contrastive variant uses the same encoder envelope but applies image-to-text target-contrastive prediction with online text targets.

The objective-specific additions are lightweight. X-JEPA [TC] adds two bias-free 768-dimensional projection heads for predicted and target text-space features and one learned temperature, contributing approximately 1.18M trainable parameters. X-JEPA [P,A] uses analogous global image/text projection heads and one learned temperature for direct alignment. These additions are small relative to the ViT-B/16 and BERT-base encoders, so they do not introduce a large capacity difference. [TC] nevertheless changes several design choices jointly.

B Full Retrieval Results

Tables 6 and 7 provide the full zero-shot retrieval recalls underlying the mean-recall values in the main paper.

Metric	CLIP	SigLIP	[P]	[TC]	[P,A]
I2T R@1	52.78	52.26	0.02	21.74	**56.90**
I2T R@5	80.42	80.16	0.12	42.44	**81.68**
I2T R@10	88.24	88.68	0.12	50.46	**88.98**
T2I R@1	38.72	38.20	0.01	27.29	**39.90**
T2I R@5	68.09	68.04	0.11	56.35	**69.29**
T2I R@10	79.06	78.66	0.22	69.31	**79.57**

Table 6: COCO Karpathy zero-shot retrieval recalls for the evaluated model set. Values are percentages. I2T is image-to-text retrieval and T2I is text-to-image retrieval.

Metric	CLIP	SigLIP	[P]	[TC]	[P,A]
I2T R@1	69.70	69.70	0.00	17.30	**71.80**
I2T R@5	90.20	91.20	0.00	37.70	**92.00**
I2T R@10	**95.90**	**95.90**	0.30	48.70	95.80
T2I R@1	54.32	54.88	0.08	38.98	**57.04**
T2I R@5	80.86	81.68	0.30	68.56	**83.02**
T2I R@10	88.22	88.54	0.56	78.56	**89.52**

Table 7: Flickr30k zero-shot retrieval recalls for the evaluated model set. Values are percentages. I2T is image-to-text retrieval and T2I is text-to-image retrieval.

C Language-Slice Tables

The main paper visualizes language-sensitive slices in Figure 4. Tables 8–10 report more detailed slice values.

Slice	CLIP	SigLIP	[P]	[TC]	[P,A]
Repl. attr.	**74.11**	70.56	33.76	46.19	73.35
Repl. obj.	88.38	87.41	41.04	53.27	**90.50**
Repl. rel.	62.87	60.53	35.92	38.83	**63.80**
Swap attr.	58.11	54.95	37.24	36.34	**60.81**
Swap obj.	38.78	42.04	26.94	28.16	**45.71**

Table 8: SugarCrepe++ slice accuracies for the evaluated model set. Values are percentages.

Slice	CLIP	SigLIP	[P]	[TC]	[P,A]
Object	92.38	91.52	50.87	90.86	**93.04**
Subject	85.84	86.25	48.69	83.78	**87.21**
Verb	81.40	81.39	50.61	76.55	**81.94**

Table 9: SVO-Probes role accuracies for the evaluated model set. Values are percentages.

Slice	CLIP	SigLIP	[P]	[TC]	[P,A]
Adjacency	58.17	55.43	47.82	51.12	**58.78**
Directional	**51.38**	50.20	48.02	45.45	50.40
Orientation	72.00	65.00	50.67	59.67	**73.67**
Projective	62.48	61.59	51.60	56.88	**62.90**
Proximity	**47.27**	43.36	37.89	40.23	46.48
Topological	68.50	**69.75**	45.88	63.04	68.83

Table 10: VSR relation-family AUROC for the evaluated model set. Values are percentages.

D Construction of the Pretraining Dataset

The pretraining dataset combines Conceptual Captions, SBU Captions, MS-COCO, and Visual Genome. We apply two filtering steps before pretraining. First, we remove records whose image files are unavailable. Second, we remove records whose source images overlap with any evaluation benchmark used in this paper. Table 11 reports the resulting source-wise counts.

Count	CC	SBU	COCO	VG	Total
Raw	2,919,397	859,739	566,747	768,536	5,114,419
Path-valid	2,919,397	837,642	566,747	768,536	5,092,322
Removed	0	113	48,561	14,180	62,854
Final	**2,919,397**	**837,529**	**518,186**	**754,356**	**5,029,468**

Table 11: Construction of the cleaned pretraining mixture. CC denotes Conceptual Captions and VG denotes Visual Genome. Removed denotes evaluation-overlap removals after path-valid filtering.

E Detailed Evaluation Protocols

Zero-shot scoring. All zero-shot evaluations freeze the pretrained model and use it in evaluation mode. For retrieval, we encode each image and caption once, normalize features, compute all image–text dot products, and rank candidates in both directions. COCO and Flickr30k report R@1, R@5, R@10, and their mean over image-to-text and text-to-image retrieval. For X-JEPA [TC], image features are obtained from the image-to-text prediction path: image tokens are mapped to a predicted text-space feature and scored against encoded text-target features.

Retrieval adapters. Adapter retrieval trains only the image adapter, text adapter, and scalar logit scale. The pretrained feature extractor remains frozen and in evaluation mode. Each adapter is a 768-to-768 residual projection head consisting of LayerNorm followed by a linear projection. Because input and output dimensions match, the head adds a learnable residual branch whose scale is initialized to zero, so the initial adapter is a no-op up to normalization. Adapter outputs are normalized before scoring. We train the adapters for 10 epochs with AdamW, learning rate 10^{-3}, weight decay 0.01, batch size 512, and cosine annealing. The best adapter state is selected by validation mean recall and then evaluated on the test split.

Hard-negative zero-shot diagnostics. SugarCrepe++, SVO-Probes, and VSR train no task-specific parameters. They use the same normalized image–text compatibility score as zero-shot retrieval. In SugarCrepe++, each example contains one image, two positive captions, and one controlled hard negative. An example is correct only if both positives score above the negative. In SVO-Probes, a sentence is scored against a positive and a negative image, and accuracy measures whether the positive image receives the higher score. In VSR, we score image–caption pairs for spatial-relation compatibility and report AUROC over the resulting scores. We use category, role, and relation-family summaries for the language-slice analysis.

NLVR2 frozen probes. Both NLVR2 diagnostics freeze the pretrained backbone and train only the probe head for 10 epochs. We use AdamW with learning rate 10^{-4}, weight decay 0.01, dropout 0.1, batch size 256, and cosine annealing. The best probe state is selected by validation accuracy and then evaluated on the test split.

The global-embedding probe encodes the left image, right image, and sentence as 768-dimensional pooled features. Its classifier receives the concatenation of the three features, the two image–text products, the two image–text absolute differences, and the left–right absolute difference. This yields an 8×768 input followed by LayerNorm, a 1536-dimensional hidden layer, GELU, dropout, and a binary output layer.

The token-interaction probe operates on raw image and text token sequences. Left-image, right-image, and text tokens are normalized and projected to dimension 256, augmented with a learned CLS token, type embeddings, and position embeddings, and passed through a 2-layer Transformer encoder with 4 heads and MLP ratio 2.0. Image tokens are spatially pooled by a factor of 2 before the probe to keep the interaction sequence compact. The CLS output is classified by a 512-dimensional MLP. Padding positions in the text sequence are masked.