FROM GLOBAL TO DENSE REPRESENTATIONAL CONVERGENCE: SHARED SEMANTIC GEOMETRY SIMPLIFIES VISION-LANGUAGE CONNECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent work on representational convergence suggests that independently trained vision and language models can develop similar semantic geometries despite differing modalities and training objectives. Existing studies, however, have mainly examined global image and sentence embeddings. In this paper, we ask whether this convergence also emerges in the dense patch representations of vision transformers. Using images with detailed pixel annotations, we capture how a broad range of concepts is organized in the dense vision space and compare this structure with the corresponding text representations. Interestingly, we find substantial alignment at this level as well, with consistently stronger alignment for larger and more recent vision and language models. Furthermore, we learn linear mappings between the vision and text representation spaces to exploit the consequences of this compatibility. Better-aligned spaces are easier to connect: their mappings are more accurate, require fewer training concepts, and perform better in downstream tasks. Specifically, stronger intrinsic alignment is associated with greater generalization to held-out categories in cross-modal retrieval, higher mIoU in text-guided semantic segmentation, and higher scores in zero-shot image captioning. Altogether, these findings extend evidence of representational convergence from global embeddings to spatially grounded concepts and link intrinsic alignment to easier cross-modal mapping and better downstream performance.

1 INTRODUCTION

Modern foundation models learn rich representations from very different sources of information. Vision models learn from images, while language models and text encoders learn from linguistic data. Despite these differences, both modalities describe many of the same concepts and structures in the world. Traditionally, connecting them has required explicit cross-modal supervision. Models such as CLIP (Radford et al., 2021), for example, learn compatible visual and textual representations from large collections of paired images and captions.

Recent work suggests, however, that part of this compatibility may emerge even when vision and language models are trained independently. Maniparambil et al. (2024) showed that unimodal vision and language encoders can develop similar representational structures, while the Platonic Representation Hypothesis (Huh et al., 2024) proposed a broader principle: as models become more capable, their representations may increasingly converge toward compatible descriptions of the same underlying data. Subsequent work has explored how far this global compatibility can be exploited, including correspondence without paired data (Schnaus et al., 2025) and lightweight adaptation between pretrained unimodal models (Zhang et al., 2025).

A common feature of these studies is that they consider images primarily through *global representations*. An image is represented by a single vector, obtained for example from the *CLS* token or by globally pooling spatial features, and its geometry is compared with representations produced by language models. Yet an image is not a single semantic entity: it is a composition of objects, materials, regions, and backgrounds, each represented at different spatial locations. If representational convergence reflects models discovering common semantic structure, it is therefore natural to ask

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.

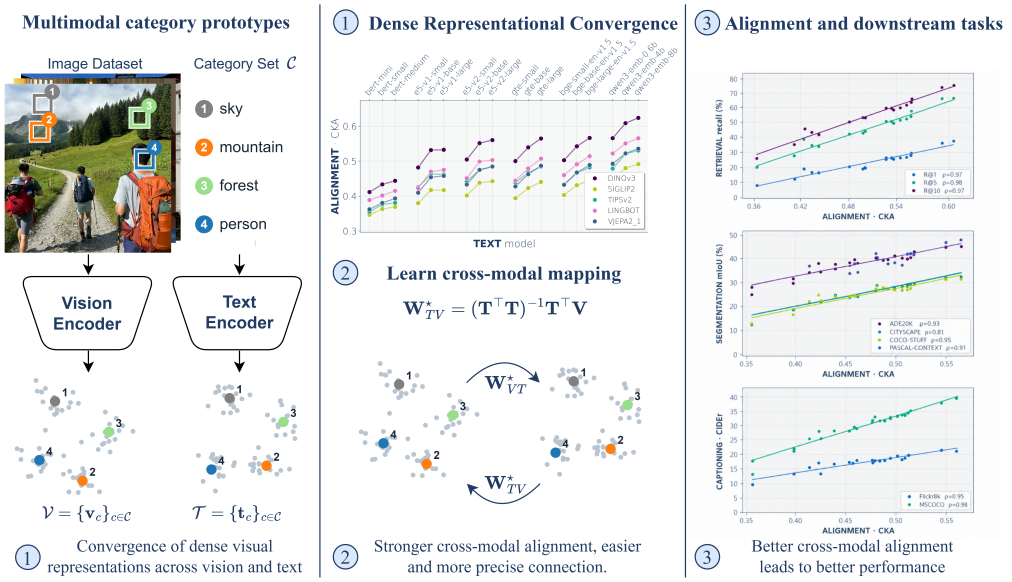

Figure 1: We study representational convergence between dense visual features and text embeddings. Alignment increases with newer and larger models and is associated with easier cross-modal mapping and better downstream vision-language performance.

whether this phenomenon extends beyond holistic image representations and also emerges *within* the image.

This question is particularly meaningful for vision transformers (Dosovitskiy et al., 2021), whose representations are naturally decomposed into spatially organized patch tokens along with global embeddings. Importantly, these local representations are not merely spatial: self-supervised vision models have shown that they can develop rich semantic structure even without language supervision. DINO (Caron et al., 2021), for example, revealed the emergence of semantic regions and object boundaries in transformer features, while DINOv2 (Oquab et al., 2024) and DINOv3 (Siméoni et al., 2026) further strengthened the quality of dense visual representations. This raises a natural question: if semantic organization already emerges locally within vision models, does its structure also resemble the one learned independently from language?

Motivated by this question, we propose to shift the semantic unit at which representational convergence is studied. Instead of comparing one representation per image with one per caption, we compare representations associated with *localized semantic categories*. This allows us to study whether independently trained vision and language models organize concepts similarly not only at the level of complete scenes, but also at the level of their spatial components.

Based on these premises, we organize our work around three main questions::

1. ***Does representational convergence extend from global to dense visual representations?***

   Independently trained vision and text models exhibit substantial alignment in how they organize spatially-grounded semantic concepts. This alignment increases consistently with model size and across newer generations of both vision and text models.

2. ***Does stronger intrinsic alignment make independently learned representation spaces easier to connect?***

   Representation spaces with stronger intrinsic alignment can be connected more accurately and with less supervision. Across model pairs, stronger initial alignment yields better cross-modal retrieval on held-out concepts and reduces the number of paired categories required to learn an effective mapping.

3. ***Does alignment predict performance on tasks that require vision and text to work together?***

   Better-aligned encoder pairs consistently perform better in both text-guided semantic segmentation and zero-shot image captioning.

2 Related Work

Emergent Alignment and Connecting Independently Trained Models. Recent work asks whether independently trained models converge toward compatible semantic representations. Maniparambil et al. (2024) found substantial similarity between unimodal vision and language encoders, while Li et al. (2024) reported partially isomorphic concept structure across model families. The Platonic Representation Hypothesis (Huh et al., 2024) generalizes these observations, proposing that stronger models increasingly converge toward a shared description of the world. Subsequent work studies where this agreement emerges across layers (He et al., 2025), whether correspondences can be recovered without paired data (Schnaus et al., 2025), and how such spaces can be connected through linear maps (Merullo et al., 2023), graph-based matching (Maniparambil et al., 2024; Schnaus et al., 2025), or lightweight projectors with limited paired supervision (Zhang et al., 2025; Maniparambil et al., 2025). Other approaches use nonlinear objectives or combine paired supervision with unpaired distributional alignment (Roschmann et al., 2026). At the same time, estimates of intrinsic alignment can depend on the metric, sample scale, and pairing assumptions (Gröger et al., 2026; Koepke et al., 2026).

Local Semantics in Self-Supervised Vision Transformers. Vision Transformers (Dosovitskiy et al., 2021) preserve spatial information through patch tokens, and self-supervised training can give these representations strong local semantics. DINO reveals object regions and boundaries without language supervision (Caron et al., 2021), while its features support object-part localization, co-segmentation, and semantic correspondence (Amir et al., 2022). DINOv2 (Oquab et al., 2024) and DINOv3 (Siméoni et al., 2026) further improve dense representations, alongside alternative objectives based on masked reconstruction (He et al., 2022) and latent prediction (Assran et al., 2023).

Language-Guided Open-Vocabulary Semantic Segmentation. Open-vocabulary segmentation transfers vision-language knowledge to dense prediction. DenseCLIP (Rao et al., 2022), MaskCLIP (Zhou et al., 2022), and GroupViT (Xu et al., 2022) adapt image-text pretraining to local features, while OVSeg (Liang et al., 2023), SAN (Xu et al., 2023b), ODISE (Xu et al., 2023a), and CAT-Seg (Cho et al., 2024) introduce specialized segmentation mechanisms. Other methods improve region-text correspondence through self-distillation (Wu et al., 2024), diffusion-based prototypes (Barsellotti et al., 2024), or self-supervised spatial priors (Lan et al., 2025). Closest to our setting, Fusioner combines frozen unimodal encoders (Ma et al., 2022), dino.txt aligns language with frozen DINOv2 at image and pixel levels (Jose et al., 2025), and Talk2DINO maps CLIP text embeddings to selected DINOv2 patches (Barsellotti et al., 2025).

Image Captioning. Image captioning increasingly relies on general-purpose vision-language models (VLMs). BLIP unifies visual understanding and generation through caption bootstrapping (Li et al., 2022), while BLIP-2 connects frozen image encoders and large language models through a trainable Querying Transformer (Li et al., 2023). Modern open-source VLMs, such as Qwen3-VL or InternVL3 (Bai et al., 2025; Zhu et al., 2025), project visual features into language-model inputs and generate text autoregressively, supported by multimodal pretraining and instruction tuning.

3 Method

This section describes how we propose to tackle the research questions introduced above. We first assess the presence of dense representational convergence. Then, we exploit the growing alignment to better link independently trained vision and text spaces. Finally, we show how such learned mappings have strong positive correlation with downstream performance.

3.1 Exploring Dense Representational Convergence

To study dense representational convergence, we first construct visual and textual prototypes for a shared semantic vocabulary $\mathcal{C}$.

Visual prototypes. For each image in a collection of semantic segmentation datasets, we extract spatial patch features $\mathbf{v}_p \in \mathbb{R}^{d_V}$ from the final block of a frozen vision transformer f_V, discarding non-spatial tokens. Each feature is normalized to unit ℓ_2 norm: $\tilde{\mathbf{v}}_p = \mathbf{v}_p / \|\mathbf{v}_p\|_2$.

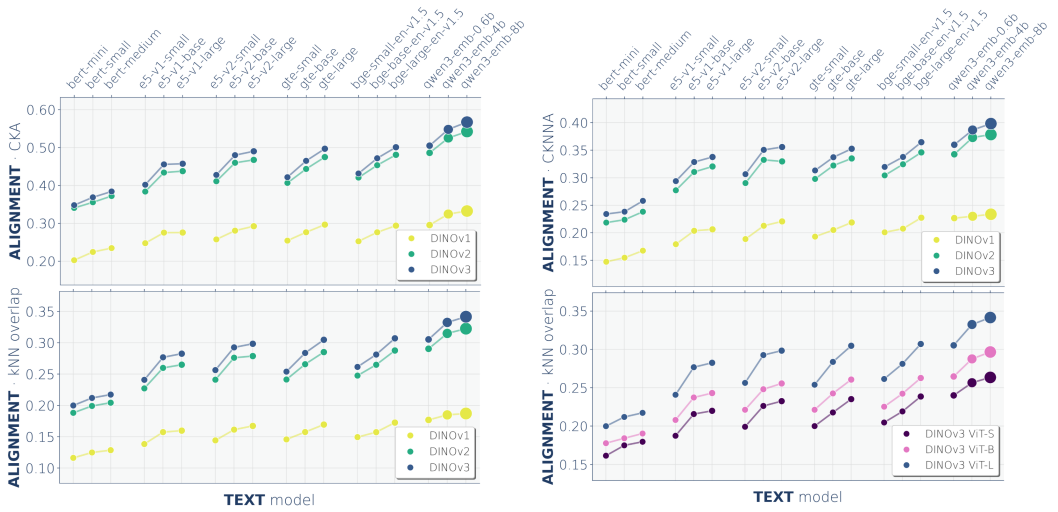

Figure 2: CKA, k-NN overlap, and CKNNA between DINOv1, DINOv2, and DINOv3 and different text encoders. The bottom-right panel shows k-NN overlap across DINOv3 scales (ViT-S, ViT-B, and ViT-L).

Using pixel-level annotations, we retain patches for which a single semantic label covers at least a fraction τ of their pixels. We group the normalized features by label across all images and datasets, yielding $\mathcal{P}_c$ for each category $c \in \mathcal{C}$. Their mean is

$$\mathbf{v}_c = \frac{1}{|\mathcal{P}_c|} \sum_{\tilde{\mathbf{v}}_p \in \mathcal{P}_c} \tilde{\mathbf{v}}_p. \tag{1}$$

Finally, we normalize each mean as $\tilde{\mathbf{v}}_c = \mathbf{v}_c / \|\mathbf{v}_c\|_2$, forming the visual prototype set $\mathcal{V} = \{\tilde{\mathbf{v}}_c\}_{c \in \mathcal{C}}$.

Text prototypes. For each category $c \in \mathcal{C}$, we construct K prompts $q_1(c), \ldots, q_K(c)$ containing its class name. A frozen text encoder f_T maps each prompt to an embedding $\mathbf{t}_{c,k} \in \mathbb{R}^{d_T}$ using its associated pooling strategy. We normalize each embedding to unit ℓ_2 norm: $\tilde{\mathbf{t}}_{c,k} = \mathbf{t}_{c,k} / \|\mathbf{t}_{c,k}\|_2$. Their mean is

$$\mathbf{t}_c = \frac{1}{K} \sum_{k=1}^{K} \tilde{\mathbf{t}}_{c,k}. \tag{2}$$

Finally, we normalize each mean as $\tilde{\mathbf{t}}_c = \mathbf{t}_c / \|\mathbf{t}_c\|_2$, forming the textual prototype set $\mathcal{T} = \{\tilde{\mathbf{t}}_c\}_{c \in \mathcal{C}}$.

The visual and textual prototypes share category identities but remain in their original representation spaces. Figure 1 illustrates their construction, from image patches and textual prompts to the corresponding semantic prototypes.

Alignment metrics. We compare latent spaces produced by different text and visual encoders using three complementary alignment metrics:

- **Centered Kernel Alignment (CKA)** measures global agreement between the pairwise similarity structures of two representation spaces (Kornblith et al., 2019). We use CKA with a linear kernel, computed directly from the centered representations.
- **k-nearest-neighbor overlap** measures local structural agreement between two spaces as the average fraction of neighboring categories shared by corresponding prototypes. Neighbors are selected independently in each space using cosine similarity, excluding the query category itself.
- **Centered Kernel Nearest-Neighbor Alignment (CKNNA)** measures local geometric agreement by combining neighborhood consistency with centered similarity values within

shared neighborhoods (Huh et al., 2024). Unlike CKA, which captures global representational similarity, CKNNA focuses on the agreement of local structure around corresponding prototypes.

Formal definitions are provided in Appendix A.1.

Figure 2 shows the dense representational convergence: larger and newer models generally exhibit stronger vision-text alignment, both in their global concept geometry (CKA) and in their local semantic neighborhoods (k-NN overlap and CKNNA). This pattern is consistent with the Platonic Representation Hypothesis (Huh et al., 2024), even though here the comparison concerns category prototypes derived from dense visual features and their textual counterparts, rather than whole-image and caption representations. Figure 3 illustrates the corresponding concept spaces for DINOv3 and Qwen3-Embedding-8B. It is worth noting how similar concepts cluster together in both spaces, highlighting how neighborhood construction is similar despite the models not being jointly trained. It is precisely these features that facilitate the learning of a cross-modal mapping.

3.2 Learning Cross-Modal Mappings

Stronger intrinsic alignment can make independently learned spaces easier to connect. In an idealized setting, consider paired, zero-mean representations $\mathbf{Q}_T, \mathbf{Q}_V \in \mathbb{R}^{n\times r}$ of equal rank, whitened so that $\mathbf{Q}_T^\top \mathbf{Q}_T = \mathbf{Q}_V^\top \mathbf{Q}_V = \mathbf{I}_r$. Their minimum normalized linear reconstruction error satisfies

$$\mathcal{E}^\star_{T\to V} = \frac{\min_{\mathbf{W}} \|\mathbf{Q}_T\mathbf{W} - \mathbf{Q}_V\|_F^2}{\|\mathbf{Q}_V\|_F^2} = 1 - \mathrm{CKA}(\mathbf{Q}_T, \mathbf{Q}_V). \tag{3}$$

Thus, under these assumptions, higher alignment implies lower optimal reconstruction error, with exact recovery at CKA equal to one. The full derivation is given in Appendix A.2, building on connections between representational similarity and linear recoverability (Kornblith et al., 2019; Williams, 2024).

Mapping semantic prototypes. Our visual and textual representation spaces may differ in dimensionality and rank, and we do not whiten their embeddings. Yet, our convergence analysis reveals increasingly higher CKA for larger and more recent models. This naturally raises the question of whether stronger intrinsic alignment also makes the two spaces easier to connect.

Let us formally define the mapping procedure. For a training vocabulary $\mathcal{C}_{\mathrm{map}} \subseteq \mathcal{C}$ of size n_{map}, let $\mathbf{T}_{\mathrm{map}} \in \mathbb{R}^{n_{\mathrm{map}}\times d_T}$ and $\mathbf{V}_{\mathrm{map}} \in \mathbb{R}^{n_{\mathrm{map}}\times d_V}$ be the input matrices to the fit, with corresponding rows indexed by concept. We fit a text-to-vision map $\mathbf{W}_{T\to V} \in \mathbb{R}^{d_T\times d_V}$ by ordinary least squares:

$$\min_{\mathbf{W}} \|\mathbf{T}_{\mathrm{map}}\mathbf{W} - \mathbf{V}_{\mathrm{map}}\|_F^2 . \tag{4}$$

The minimum-Frobenius-norm solution is

$$\mathbf{W}^\star_{T\to V} = \mathbf{T}^\dagger_{\mathrm{map}}\mathbf{V}_{\mathrm{map}}, \tag{5}$$

where † denotes the Moore-Penrose pseudoinverse. This solution exists for any rank. When T_{map} has full column rank, it reduces to

$$\mathbf{W}^\star_{T\to V} = \left(\mathbf{T}^\top_{\mathrm{map}}\mathbf{T}_{\mathrm{map}}\right)^{-1}\mathbf{T}^\top_{\mathrm{map}}\mathbf{V}_{\mathrm{map}}. \tag{6}$$

The same procedure has been applied for vision-to-text mapping $\mathbf{W}^\star_{V\to T}$, as illustrated in Fig 3.

3.3 Alignment and Downstream Tasks

After mapping the two representation spaces, we verify whether intrinsic alignment between text and vision models predicts performance on tasks requiring cross-modal cooperation: retrieval, semantic segmentation, and captioning.

Cross-modal concept retrieval. Cross-modal retrieval is a direct proxy task for testing the quality of learned mapping: successful retrieval requires preserving concept identity, locally and globally, across modalities. We fit the mappings on $\mathcal{C}_{\mathrm{map}}$ and retain a disjoint concept set $\mathcal{C}_{\mathrm{test}}$ for testing.

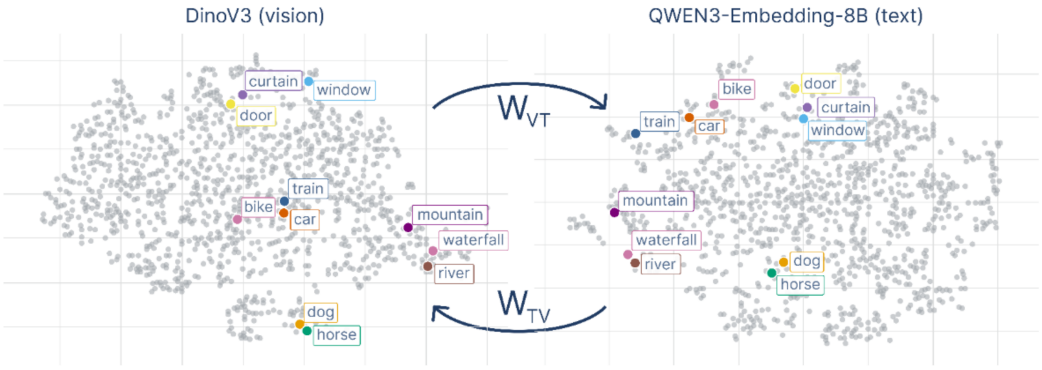

Figure 3: Visual prototypes from DINOv3 (left) and text prototypes from Qwen3-Embedding-8B (right). Similar categories cluster together in both spaces, suggesting compatible semantic structure that facilitates cross-modal mapping.

For each held-out concept $c \in \mathcal{C}_{\text{test}}$, let $\mathbf{t}_c$ and $\mathbf{v}_c$ denote the prototype inputs prepared according to Sec. 3.1. We project them into the opposite modality using the learned mappings:

$$\mathbf{q}_c^{T\to V} = \mathbf{W}_{T\to V}^{\top}\mathbf{t}_c, \qquad \mathbf{q}_c^{V\to T} = \mathbf{W}_{V\to T}^{\top}\mathbf{v}_c. \tag{7}$$

These projected representations are used to retrieve the corresponding held-out concepts in the other modality.

Text-guided semantic segmentation. This task tests whether alignment between category prototypes translates into accurate spatial predictions. This requires mapped textual concepts to discriminate visual content at individual image locations. We learn a text-to-vision mapping from the complete concept vocabulary. For each target category $c \in \mathcal{C}_{\text{seg}}$, we project its textual prototype into the visual feature space, obtaining $\mathbf{z}_c$ as specified in Eq. 7. We normalize it to form a classifier and score each normalized patch feature $\tilde{\mathbf{v}}_p$:

$$\tilde{\mathbf{v}}_c = \frac{\mathbf{z}_c}{\|\mathbf{z}_c\|_2}, \qquad s_{p,c} = \tilde{\mathbf{v}}_p^{\top}\tilde{\mathbf{v}}_c, \qquad c \in \mathcal{C}_{\text{seg}}. \tag{8}$$

Let $\mathbf{S}_c = (s_{p,c})_g$ be the score map for category c on the patch grid g, and let $\mathcal{B}$ denote bilinear upsampling to the original image resolution. The predicted label at pixel x is

$$\hat{y}_x = \arg\max_{c\in\mathcal{C}_{\text{seg}}} \big[\mathcal{B}(\mathbf{S}_c)\big]_x \tag{9}$$

We evaluate the resulting masks using mean Intersection-over-Union (mIoU). No segmentation decoder or additional task-specific parameters are learned.

Zero-shot image captioning. Finally, we test whether the visual semantics transferred by the mapping can support coherent natural-language descriptions through the captioning task.

We group the normalized patch features of an image $\mathbf{I}$ into k clusters using k-means algorithm. Let $\mathcal{P}_k$ contain the patch indices in cluster k-th, with $n_k = |\mathcal{P}_k|$. We compute its normalized mean feature and assign the nearest mapped textual category from the complete vocabulary:

$$\tilde{\mathbf{v}}_k = \frac{\sum_{p\in\mathcal{P}_k}\tilde{\mathbf{v}}_p}{\left\|\sum_{p\in\mathcal{P}_k}\tilde{\mathbf{v}}_p\right\|_2}, \qquad c_k^{\star} = \arg\max_{c\in\mathcal{C}} \tilde{\mathbf{v}}_k^{\top}\tilde{\mathbf{v}}_c, \tag{10}$$

where $\tilde{\mathbf{v}}_k$ is the centroid of cluster k, $\tilde{\mathbf{v}}_c$ is defined in Eq.8, and $c_k^{\star}$ is the category assigned to the k-th cluster. The matching score is $a_k = \tilde{\mathbf{v}}_k^{\top}\tilde{\mathbf{v}}_{c_k^{\star}}$. Once we know all the cluster information, our goal is to push it blindly into a language model to finally generate the dense caption of the original image. Therefore, for each predicted category, we extract a textual summary that records the number of assigned clusters, the number of patches in each cluster, and their matching scores. Then, the frozen language model can generate one caption from this summary alone, without access to the image, visual features, or spatial coordinates. Since the language model receives only a semantic summary built from dense visual features, caption quality probes the usefulness of the information provided by this connection.

4 EXPERIMENTAL RESULTS

4.1 SEMANTIC PROTOTYPE CONSTRUCTION

Visual prototypes. We construct 3,440 semantic prototypes from ADE20K (Zhou et al., 2017), MS COCO 2017 (Lin et al., 2014), COCO-Stuff 164K (Caesar et al., 2018), Cityscapes (Cordts et al., 2016), PASCAL-Context (Mottaghi et al., 2014), PASCAL VOC 2012 (Everingham et al., 2010), and SUN2012 (Patterson & Hays, 2012). Table 1 summarizes the number of images, annotated regions, and semantic categories contributed by each dataset. Images are processed at their native resolution, without resizing, in order to preserve the original spatial information and avoid merging small semantic regions. For the main analysis, we extract final-block spatial tokens from frozen DINO, DINOv2, and DINOv3 ViT-L models at native image resolution, retaining patches with at least 75% single-class coverage ($\tau = 0.75$). Unit-normalized features are averaged and normalized within each category. Across all datasets, this procedure yields approximately 332.4M retained patches aggregated in 3440 visual prototypes. Additional statistics and aggregation details appear in Appendix A.3.

Table 1: Semantic segmentation datasets used for prototype construction, with splits, image counts, category counts, and retained patches.

Dataset	Split	Images	Categories	Patch Extracted
COCO-Object	train2017	118,287	80	∼24.7M
COCO-Stuff	train2017	118,287	171	∼104.9M
Cityscapes	train	2,975	19	∼19.6M
SUN	2012-train	16,873	2,552	∼55.2M
PascalContext	train	4,998	392	∼6.2M
ADE20K	train	25,574	3,206	∼121.2M
PASCAL VOC2012	trainval	2,913	20	∼0.4M
Total	-	171,620*	3,440**	∼332.4M

* COCO-Object and COCO-Stuff share the same dataset
**3,440 after category merging and vocabulary filtering.

Text prototypes. We evaluate BERT, E5, GTE, BGE, and Qwen3-Embedding families across model scales using the same vocabulary and $K = 14$ fixed prompts per category. Each encoder uses its recommended input formatting and pooling strategy. Prompt embeddings are individually unit-normalized, averaged, and normalized again, as in Sec. 3.1. Checkpoint details and prompt templates are provided in Appendix A.4.

4.2 DENSE REPRESENTATIONAL CONVERGENCE

We compare independently learned dense visual and textual concept spaces using linear CKA, k-NN overlap, and CKNNA. All model pairs are evaluated on the same $1{,}024$ categories with the largest number of retained patches, matching the evaluation scale of Huh et al. (2024). Neighborhood-based metrics use cosine similarity with $k = 10$. Figure 2 compares DINO, DINOv2, and DINOv3 with ViT-L backbones, as well as DINOv3 with ViT-S, ViT-B, and ViT-L backbones, against text embedding models from multiple families. Alignment generally strengthens across successive DINO generations and with increasing model scale in both modalities. Interestingly, the consistent trends observed across CKA, k-NN overlap, and CKNNA indicate that this increase in alignment encompasses both global concept geometry and local semantic neighborhoods, providing evidence of dense representational convergence consistent with the hypothesis posed in our first research question in Sec. 1. In particular, DINOv3 achieves a CKA score of 0.567 when paired with Qwen-3-embedding-8B, compared with slightly above 0.20 for the pairing of DINO and BERT Mini. This corresponds to a nearly threefold increase in alignment across newer generations of vision models and newer, larger text models. Moreover, DINOv3 and Qwen-3-embedding-8B achieve a k-NN overlap of 0.341. With $k = 10$, this means that, on average across all evaluated concepts, approximately 3.4 of the 10 nearest neighbors of a concept are shared between the two latent spaces. This degree of neighborhood agreement is particularly notable given that the models are trained independently and operate on different modalities. Overall, these findings support the existence of convergence between category-level summaries of dense visual features and textual representations, extending the scope of the Platonic Representation Hypothesis beyond global image representations. Additional vision encoders and evaluation configurations are reported in Appendix A.5.

4.3 LINEAR CROSS-MODAL MAPPING

We divide the $3{,}440$ concepts into fixed, disjoint training and test sets using an $80/20$ split shared across all model pairs, ending up constructing $\mathcal{C}_{map}$ and $\mathcal{C}_{test}$ splits. Following Sec. 3.2, we independently fit text-to-vision and vision-to-text linear maps. We center prototypes using training-set means and use a small ridge penalty ($\lambda = 10^{-2}$) when estimating $\mathbf{W}$ for numerical stability.

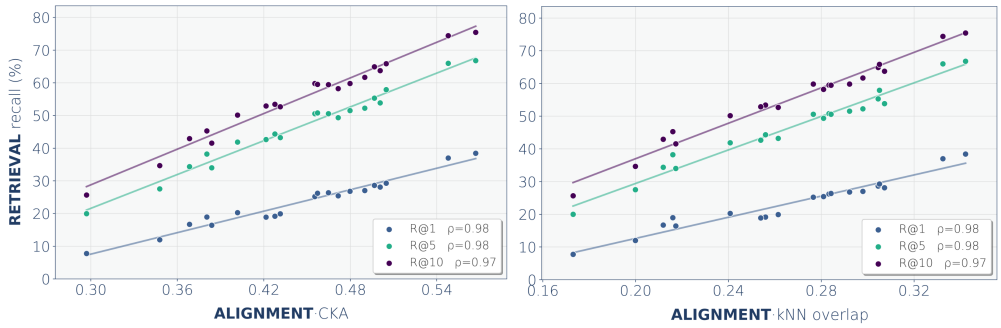

Figure 4: Held-out retrieval (R@1, R@5, and R@10, averaged across both directions) versus CKA (left) and k-NN overlap (right). Each point represents a DINOv3-text encoder pair; higher alignment is associated with better retrieval.

Retrieval of unseen concepts. Following Sec. 3.3, we project each held-out prototype into the opposite modality and rank the held-out target prototypes by cosine similarity. We report Recall@1, @5, and @10, averaged across both mapping directions. In Fig. 4, each point represents DINOv3 paired with a different text encoder. These results show that both CKA and k-NN overlap are strongly correlated with retrieval performance, providing empirical support for our second hypothesis introduced in Sec. 1. Quantitatively, CKA achieves Spearman correlation coefficients of 0.975, 0.979, and 0.971 with R@1, R@5, and R@10, respectively. These strong correlations provide evidence that greater alignment between the visual and textual representation spaces is associated with mappings that generalize more effectively to categories unseen during fitting.

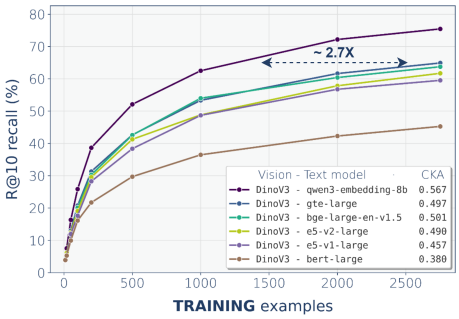

Figure 5: R@10 vs. number of training concepts for DINOv3 and different text encoders. Higher alignment (CKA, in the legend) is associated with better retrieval and greater sample efficiency.

Sample efficiency. We investigate the sample efficiency of different cross-modal mappings by varying the number of training concepts, using shared subsets and a fixed held-out test set. Figure 5 compares DINOv3 paired with different text encoders. More strongly aligned vision-language representation spaces achieve higher retrieval performance with fewer paired concepts and converge to higher final performance as the amount of supervision increases. These results suggest that stronger intrinsic alignment improves sample efficiency, reducing the amount of paired-concept supervision required to connect independently trained modalities. In particular, as both encoders become more capable and their representation spaces more closely aligned, an effective cross-modal mapping can be learned from substantially fewer paired concepts.

4.4 Text-Guided Semantic Segmentation

In this task, we test whether cross-modal compatibility translates into accurate dense semantic prediction on ADE20K, Cityscapes, COCO-Stuff, and PASCAL-Context, using their respective 150, 19, 171, and 59 classes. We fix DINOv3 as the visual backbone and evaluate mappings from different text encoders, each fitted on the complete vocabulary $\mathcal{C}$. Following Sec. 3.3, we project each dataset textual class prototypes into the visual space and score normalized image patch tokens by cosine similarity. We bilinearly upsample the resulting class-score maps to image resolution and assign each pixel its highest-scoring class. We report mean Intersection-over-Union (mIoU) on the evaluation splits, without training a segmentation decoder. Similarly to the retrieval setting, Figure 6 shows that higher intrinsic alignment is associated with higher mIoU across all four datasets. More aligned text and vision representations yield mapped classifiers that better preserve semantic distinctions at individual image locations. These results extend the category-level retrieval findings to

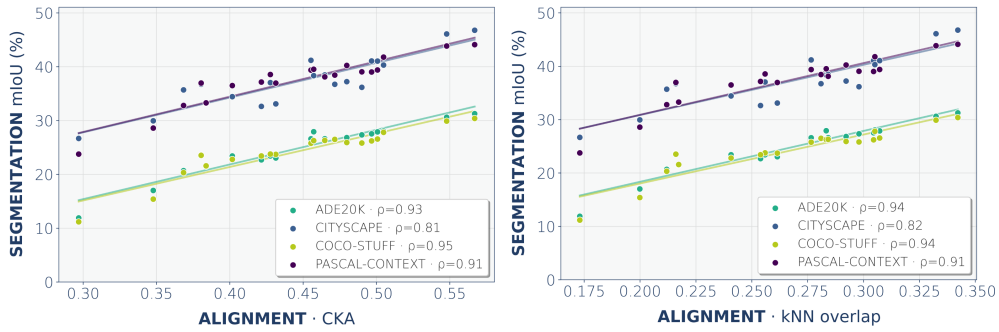

Figure 6: Intrinsic alignment and semantic segmentation performance. Each point corresponds to a DINOv3-text encoder pair, with alignment measured before mapping and mIoU evaluated using mapped textual prototypes as patch classifiers. Higher alignment is associated with higher mIoU.

a dense prediction task, further supporting the practical relevance of representational convergence. More interestingly, this setting directly tests whether semantic knowledge can be transferred from the textual to the visual modality: the inputs are textual class labels, while the outputs are spatially localized pixel predictions. Additional details and results are listed in Appendix A.6.

4.5 Zero-Shot Image Captioning

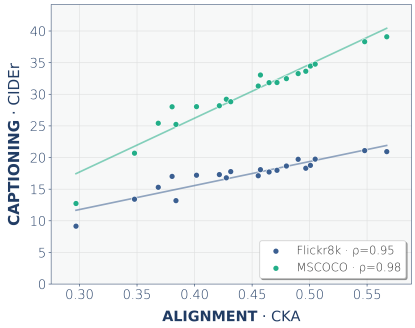

Figure 7: Intrinsic alignment and zero-shot image captioning performance on Flickr8k and MS COCO. Each point represents a DINOv3-text encoder pair.

Complementing the semantic segmentation analysis presented above, in this task we investigate the reverse direction of cross-modal transfer: how visual information at the pixel level can be mapped into the textual modality for the downstream objective of text generation. Following Sec. 3.3, we cluster DINOv3 spatial tokens using k-means and assign each centroid its nearest mapped textual prototype from the full 3,440-concept vocabulary. Repeated labels are merged into a textual summary retaining cluster multiplicities, patch counts, and matching scores. A frozen LLM generates a caption from this summary alone, without access to the image or visual features. We use QWEN3-8B as frozen caption generator LLM. We evaluate on Flickr8k and the MS COCO evaluation split using BLEU-4, METEOR, CIDEr, and SPICE, without caption-specific training. All the details appear in Appendix A.7. As for the previous tasks, also Figure 7 shows a strong positive association between intrinsic alignment and caption quality across the evaluated DINOv3-text encoder pairs. More aligned pairs provide semantic summaries that support more accurate captions, extending the alignment-performance relationship to text generation too. Together with the retrieval and segmentation experiments, these results provide strong empirical support for our final research question in Sec. 1, showing that the degree of intrinsic alignment between independently trained vision and text representation spaces is strongly associated with performance on downstream tasks that require cross-modal interaction.

5 Conclusion

We show that vision-language representational convergence extends to semantic prototypes derived from dense visual features. Across model generations and scales, stronger alignment is associated with more accurate linear mappings, better generalization to unseen categories, and greater sample efficiency. More aligned pairs also achieve better text-guided segmentation and zero-shot captioning performance. These results connect representational compatibility with cross-modal learnability and semantic transfer, supporting alignment as a criterion for connecting independently trained vision and language models.

AI USE STATEMENT

In this work, we used generative AI tools for writing assistance, i.e., polishing and proofreading the text (grammar and formatting). We have not used generative AI tools for research ideation or methodology design. These were carried out entirely by the authors. Retrieval of related work and literature review were conducted manually by the authors. All technical claims, experimental results, and artifacts were produced by the authors. We take full responsibility for the final content of this work, including all text, claims, and artifacts.

REPRODUCIBILITY STATEMENT

To support reproducibility, we describe the construction and normalization of visual and textual semantic prototypes in Section 3.1, with datasets, extraction settings, and prompt aggregation described in Section 4.1 and Table 1. The representational alignment metrics are introduced in Section 3.1 and formally defined in Appendix A.1, including linear CKA in Eq. 11. Section 3.2 describes linear cross-modal mapping, while Appendix A.2 provides the assumptions and complete proof of the relationship between alignment and optimal linear reconstruction in Eq. 3. The concept split, centering, regularization, and held-out retrieval protocol are reported in Section 4.3. Downstream prediction procedures are described in Section 3.3, with segmentation and captioning evaluations presented in Sections 4.4 and 4.5, respectively. Our anonymized code repository is available at `https://anonymous.4open.science/r/TEXT-VISION-ALIGNMENT-D059/`.

REFERENCES

Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep vit features as dense visual descriptors. *ECCVW What is Motion For?*, 2022.

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 *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15619–15629. IEEE, 2023.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025.

Luca Barsellotti, Roberto Amoroso, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Training-free open-vocabulary segmentation with offline diffusion-augmented prototype generation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3689–3698, 2024.

Luca Barsellotti, Lorenzo Bianchi, Nicola Messina, Fabio Carrara, Marcella Cornia, Lorenzo Baraldi, Fabrizio Falchi, and Rita Cucchiara. Talking to dino: Bridging self-supervised vision backbones with language for open-vocabulary segmentation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22025–22035. IEEE, 2025.

Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1209–1218. IEEE, 2018.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 9630–9640. IEEE, 2021.

Seokju Cho, Heeseong Shin, Sunghwan Hong, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4113–4123. IEEE, 2024.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3213–3223, 2016.

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. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2): 303–338, 2010.

Fabian Gröger, Shuo Wen, and Maria Brbic. Revisiting the platonic representation hypothesis: An aristotelian view. In *Forty-third International Conference on Machine Learning*, 2026.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 15979–15988. IEEE, 2022.

Zoe Wanying He, Sean Trott, and Meenakshi Khosla. Seeing through words, speaking through pixels: Deep representational alignment between vision and language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 35645–35660, 2025.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In *Forty-first International Conference on Machine Learning*, 2024.

Cijo Jose, Théo Moutakanni, Dahyun Kang, Federico Baldassarre, Timothée Darcet, Hu Xu, Daniel Li, Marc Szafraniec, Michaël Ramamonjisoa, Maxime Oquab, et al. Dinov2 meets text: A unified framework for image-and pixel-level vision-language alignment. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24905–24916. IEEE, 2025.

A. Sophia Koepke, Daniil Zverev, Shiry Ginosar, and Alexei A. Efros. Back into plato's cave: Examining cross-modal representational convergence at scale. *ArXiv*, abs/2604.18572, 2026.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pp. 3519–3529. PMlR, 2019.

Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Proxyclip: Proxy attention improves clip for open-vocabulary segmentation. In *Computer Vision – ECCV 2024*, pp. 70–88, Cham, 2025. Springer Nature Switzerland.

Jiaang Li, Yova Kementchedjhieva, Constanza Fierro, and Anders Søgaard. Do vision and language models share concepts? a vector space alignment study. *Transactions of the Association for Computational Linguistics*, 12:1232–1249, 2024.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, 2022.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 23–29 Jul 2023.

Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7061–7070. IEEE, 2023.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Chaofan Ma, Yu-Hao Yang, Yanfeng Wang, Ya Zhang, and Weidi Xie. Open-vocabulary semantic segmentation with frozen vision-language models. *ArXiv*, abs/2210.15138, 2022.

Mayug Maniparambil, Raiymbek Akshulakov, Yasser Abdelaziz Dahou Djilali, Mohamed El Amine Seddik, Sanath Narayan, Karttikeya Mangalam, and Noel E O'Connor. Do vision and language encoders represent the world similarly? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14334–14343, 2024.

Mayug Maniparambil, Raiymbek Akshulakov, Yasser Abdelaziz Dahou Djilali, Sanath Narayan, Ankit Singh, and Noel E. O'Connor. Harnessing frozen unimodal encoders for flexible multimodal alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29847–29857, June 2025.

Jack Merullo, Louis Castricato, Carsten Eickhoff, and Ellie Pavlick. Linearly mapping from image to text space. In *The Eleventh International Conference on Learning Representations*, 2023.

Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

Genevieve Patterson and James Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2751–2758, 2012. doi: 10.1109/CVPR.2012.6247998.

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 *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 18061–18070. IEEE, 2022.

Simon Roschmann, Paul Krzakala, Sonia Mazelet, Quentin Bouniot, and Zeynep Akata. SOTAlign: Semi-supervised alignment of unimodal vision and language models via optimal transport. In *Forty-third International Conference on Machine Learning*, 2026.

Dominik Schnaus, Nikita Araslanov, and Daniel Cremers. It'sa (blind) match! towards vision-language correspondence without parallel data. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24983–24992. IEEE, 2025.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seung Eun Yi, Michael Ramamonjisoa, Francisco Massa, Daniel HAZIZA, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien

Mairal, Herve Jegou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. Featured Certification.

Alex H Williams. Equivalence between representational similarity analysis, centered kernel alignment, and canonical correlations analysis. In *UniReps: 2nd Edition of the Workshop on Unifying Representations in Neural Models*, 2024.

Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Xiangtai Li, Wentao Liu, and Chen Change Loy. CLIPSelf: Vision transformer distills itself for open-vocabulary dense prediction. In *The Twelfth International Conference on Learning Representations*, 2024.

Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18113–18123. IEEE, 2022.

Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *2023 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 2955–2966. IEEE, 2023a.

Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. San: Side adapter network for open-vocabulary semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12):15546–15561, 2023b.

Le Zhang, Qian Yang, and Aishwarya Agrawal. Assessing and learning alignment of unimodal vision and language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14604–14614. IEEE, 2025.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 5122–5130. IEEE, 2017.

Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European conference on computer vision*, pp. 696–712. Springer, 2022.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

A Appendix

A.1 Measuring Representational Convergence

Let $c_1, \ldots, c_n \in \mathcal{C}$ be the $n \geq 2$ distinct concepts used in an alignment comparison. Following Sec. 3.1, their normalized visual and textual prototypes are $\tilde{\mathbf{v}}_{c_i} \in \mathbb{R}^{d_V}$ and $\tilde{\mathbf{t}}_{c_i} \in \mathbb{R}^{d_T}$. We collect them in matrices $\mathbf{V} \in \mathbb{R}^{n \times d_V}$ and $\mathbf{T} \in \mathbb{R}^{n \times d_T}$ with corresponding rows:

$$\mathbf{V}_{i,:} = \tilde{\mathbf{v}}_{c_i}^\top, \qquad \mathbf{T}_{i,:} = \tilde{\mathbf{t}}_{c_i}^\top, \qquad i = 1, \ldots, n.$$

Using the linear kernel $\kappa(\mathbf{a}, \mathbf{b}) = \mathbf{a}^\top \mathbf{b}$ in each modality, we obtain the Gram matrices

$$\mathbf{K}_V = \mathbf{V}\mathbf{V}^\top, \qquad \mathbf{K}_T = \mathbf{T}\mathbf{T}^\top.$$

Their entries are cosine similarities because the prototype rows have unit norm. All three metrics use the same ordered concept set, without learning a cross-modal mapping.

Centered Kernel Alignment with a linear kernel (CKA). Define the centering matrix and the double-centered Gram matrices as

$$\mathbf{H} = \mathbf{I}_n - \frac{1}{n}\mathbf{1}\mathbf{1}^\top, \qquad \bar{\mathbf{K}}_V = \mathbf{H}\mathbf{K}_V\mathbf{H}, \qquad \bar{\mathbf{K}}_T = \mathbf{H}\mathbf{K}_T\mathbf{H},$$

where $\mathbf{1} \in \mathbb{R}^n$ is the all-ones vector. The empirical *Hilbert-Schmidt Independence Criterion* (HSIC) is

$$\mathrm{HSIC}(\mathbf{K}_V, \mathbf{K}_T) = \frac{1}{(n-1)^2}\,\mathrm{tr}(\bar{\mathbf{K}}_V\bar{\mathbf{K}}_T).$$

Centered Kernel Alignment is its normalized counterpart (Kornblith et al., 2019):

$$\mathrm{CKA}_{\mathrm{lin}}(\mathbf{V},\mathbf{T}) = \frac{\mathrm{HSIC}(\mathbf{K}_V,\mathbf{K}_T)}{\sqrt{\mathrm{HSIC}(\mathbf{K}_V,\mathbf{K}_V)\,\mathrm{HSIC}(\mathbf{K}_T,\mathbf{K}_T)}}. \tag{11}$$

For a nonzero denominator, CKA lies in $[0,1]$, with larger values indicating stronger agreement between the global similarity structures. It is invariant to isotropic scaling of either representation. Here, CKA always uses the linear kernels defined above.

k-nearest-neighbor overlap. For an integer $1 \le k \le n-1$, define

$$\mathcal{N}_k^V(i) = \mathrm{TopK}_{j\in\{1,\dots,n\}\setminus\{i\}}(\mathbf{K}_V)_{ij}, \qquad \mathcal{N}_k^T(i) = \mathrm{TopK}_{j\in\{1,\dots,n\}\setminus\{i\}}(\mathbf{K}_T)_{ij},$$

where TopK returns the set of indices of the k largest entries. Neighbors are selected independently in each modality using the uncentered cosine similarities; the query itself is excluded. Ties are resolved by a fixed ordering of concept indices. The overlap is

$$\mathrm{NNOverlap}_k(\mathbf{V},\mathbf{T}) = \frac{1}{nk}\sum_{i=1}^{n}\left|\mathcal{N}_k^V(i)\cap\mathcal{N}_k^T(i)\right|.$$

This is the average fraction of neighbors shared by corresponding concepts (Huh et al., 2024). It lies in $[0,1]$ and equals one exactly when the two neighbor sets coincide for every query; neighbor ordering within each set does not affect the score.

Centered Kernel Nearest-Neighbor Alignment (CKNNA). CKNNA combines shared-neighbor selection with centered similarity values (Huh et al., 2024). We specify the double-centered form used here by first constructing the directed neighborhood masks

$$\mathbf{M}_{ij}^V = \mathbb{1}[j\in\mathcal{N}_k^V(i)], \qquad \mathbf{M}_{ij}^T = \mathbb{1}[j\in\mathcal{N}_k^T(i)],$$

where $\mathbb{1}[\cdot]$ denotes the indicator function. Their diagonal entries are zero. Let $\odot$ denote elementwise multiplication and $\langle\mathbf{A},\mathbf{B}\rangle_F = \mathrm{tr}(\mathbf{A}^\top\mathbf{B})$ the Frobenius inner product. We define

$$\mathrm{CKNNA}_k(\mathbf{K}_V,\mathbf{K}_T) = \frac{\left\langle \mathbf{M}^V\odot\bar{\mathbf{K}}_V, \mathbf{M}^T\odot\bar{\mathbf{K}}_T\right\rangle_F}{\|\mathbf{M}^V\odot\bar{\mathbf{K}}_V\|_F\,\|\mathbf{M}^T\odot\bar{\mathbf{K}}_T\|_F}.$$

The numerator sums products of centered similarities only for pairs that are neighbors in both spaces; each denominator factor uses all neighbors selected in its own space. Centering is performed on the full Gram matrices *before* masking, while neighbor selection uses the original Gram matrices. The score is defined when both denominator factors are nonzero and, as a normalized Frobenius inner product, lies in $[-1,1]$. Unlike k-NN overlap, it accounts for similarity magnitudes and their signs relative to the centered geometry.

A.2 Alignment and Optimal Linear Reconstruction

This section proves Eq. 3 and states the conditions under which it holds.

A.2.1 Idealized equal-rank setting

Assumptions and whitening. Let $n \ge 2$ and $1 \le r \le n-1$. Consider paired representations $\mathbf{Q}_T, \mathbf{Q}_V \in \mathbb{R}^{n\times r}$ satisfying

$$\mathbf{1}^\top\mathbf{Q}_T = \mathbf{1}^\top\mathbf{Q}_V = \mathbf{0}^\top, \qquad \mathbf{Q}_T^\top\mathbf{Q}_T = \mathbf{Q}_V^\top\mathbf{Q}_V = \mathbf{I}_r.$$

The first condition requires zero feature means across the n paired concepts; the second requires equal rank and orthonormal columns. Such representations can be obtained from any two centered, rank-r matrices by whitening their nonzero feature subspaces. Specifically, if $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{R}^\top$ is a compact singular value decomposition, with all r singular values positive, then $\mathbf{Q} = \mathbf{X}\mathbf{R}\mathbf{\Sigma}^{-1} = \mathbf{U}$ has the required properties. This uses unit column norms; the sample covariance is therefore $\mathbf{I}_r/(n-1)$.

Statement. For an unconstrained linear map $\mathbf{W} \in \mathbb{R}^{r\times r}$, define

$$\mathcal{E}^{\star}_{T\to V} = \frac{\min_{\mathbf{W}} \|\mathbf{Q}_T\mathbf{W} - \mathbf{Q}_V\|_F^2}{\|\mathbf{Q}_V\|_F^2}.$$

Then the unique minimizer is $\mathbf{W}^{\star}_{\text{ideal}} = \mathbf{Q}_T^\top\mathbf{Q}_V$, and

$$\mathcal{E}^{\star}_{T\to V} = 1 - \text{CKA}_{\text{lin}}(\mathbf{Q}_T, \mathbf{Q}_V).$$

In particular, the error is zero if and only if the two matrices have the same column space, equivalently if and only if their linear CKA is one.

Proof. Because the feature means are zero, the linear CKA expression in Appendix A.1 applies directly. Orthonormality gives $\|\mathbf{Q}_T^\top\mathbf{Q}_T\|_F = \|\mathbf{Q}_V^\top\mathbf{Q}_V\|_F = \sqrt{r}$, so

$$\text{CKA}_{\text{lin}}(\mathbf{Q}_T, \mathbf{Q}_V) = \frac{\|\mathbf{Q}_T^\top\mathbf{Q}_V\|_F^2}{r}. \tag{12}$$

The least-squares normal equations are

$$\mathbf{Q}_T^\top\mathbf{Q}_T\mathbf{W} = \mathbf{Q}_T^\top\mathbf{Q}_V.$$

Since $\mathbf{Q}_T^\top\mathbf{Q}_T = \mathbf{I}_r$, they yield the unique minimizer $\mathbf{W}^{\star}_{\text{ideal}} = \mathbf{Q}_T^\top\mathbf{Q}_V$. No orthogonality constraint or regularization is imposed on $\mathbf{W}$.

Let $\mathbf{P}_T = \mathbf{Q}_T\mathbf{Q}_T^\top$. This is a symmetric, idempotent matrix and hence the orthogonal projector onto $\text{col}(\mathbf{Q}_T)$. The fitted representation is $\mathbf{P}_T\mathbf{Q}_V$, while the residual is $(\mathbf{I}_n - \mathbf{P}_T)\mathbf{Q}_V$. These components are orthogonal in the Frobenius inner product because $\mathbf{P}_T(\mathbf{I}_n - \mathbf{P}_T) = \mathbf{0}$. Consequently,

$$\begin{aligned}\|\mathbf{Q}_T\mathbf{W}^{\star}_{\text{ideal}} - \mathbf{Q}_V\|_F^2 &= \|(\mathbf{I}_n - \mathbf{P}_T)\mathbf{Q}_V\|_F^2 \\ &= \|\mathbf{Q}_V\|_F^2 - \|\mathbf{P}_T\mathbf{Q}_V\|_F^2 \\ &= r - \|\mathbf{Q}_T^\top\mathbf{Q}_V\|_F^2.\end{aligned}$$

The last step uses $\|\mathbf{Q}_V\|_F^2 = r$ and $\|\mathbf{Q}_T\mathbf{A}\|_F^2 = \|\mathbf{A}\|_F^2$ for any compatible $\mathbf{A}$, a consequence of $\mathbf{Q}_T^\top\mathbf{Q}_T = \mathbf{I}_r$. Dividing by r and substituting Eq. 12 proves the identity. Zero residual is equivalent to $\text{col}(\mathbf{Q}_V) \subseteq \text{col}(\mathbf{Q}_T)$; because both spaces have dimension r, this is equality. Reversing the mapping direction gives the same normalized error in this equal-rank setting. □

A.3 Semantic Prototype Construction Details

Datasets. Table 1 reports the datasets and splits used for prototype extraction. Visual feature extraction and patch selection follow Section 4.1. When an image appears in multiple datasets, we retain the contributions from each annotation set independently, since the annotations can describe different semantic content. In particular, COCO-Object and COCO-Stuff share images but contribute their respective annotations to prototype construction.

Category cleaning and filtering. We standardize category names with respect to capitalization, whitespace, punctuation, and singular/plural forms. We exclude labels containing the word *"occluded"*, such as *"person occluded"* and *"table occluded"*, and remove the trailing word *"crop"*, so that, for example, *"fire crop"* becomes *"fire"*. Labels with identical names after cleaning are merged into a shared semantic category. We discard categories with fewer than 10 extracted patches in total across all datasets. Together, category cleaning, merging, and filtering reduce $6{,}440$ raw labels to a vocabulary of $3{,}440$ concepts. Figure 8 shows the histogram of patch counts for the 30 categories with the most extracted patches across all datasets.

A.4 Text Encoders and Prompt Templates

Checkpoints and embedding extraction. We evaluate 18 pretrained text encoders from the Qwen3-Embedding, E5, GTE, BGE, and BERT families, all obtained from Hugging Face. Table 2 lists the exact checkpoint identifiers and their names in our experiments. We use the embedding interface provided for each model and retain the checkpoint defaults for maximum input length and truncation. All encoders remain frozen during prototype construction and downstream experiments.

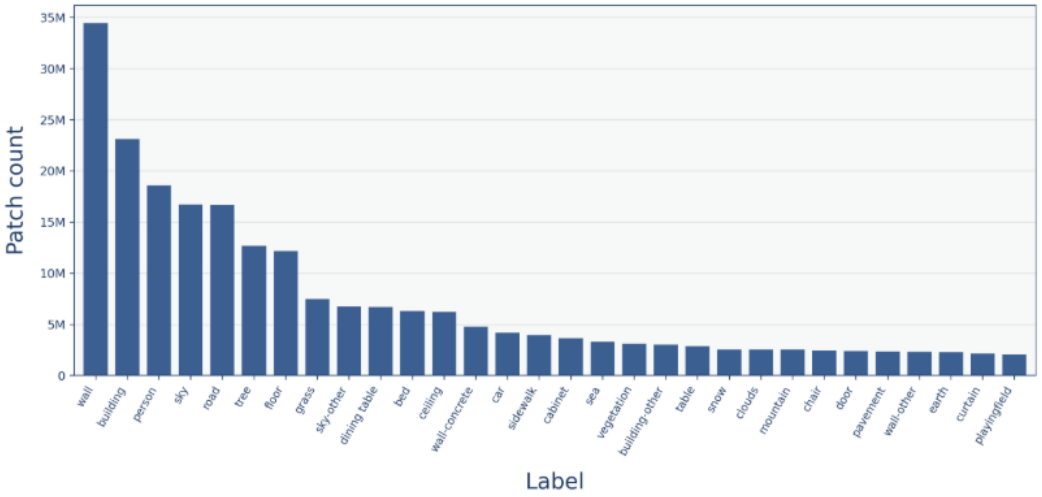

Figure 8: Distribution of extracted patches across the 30 most frequent semantic categories after label cleaning and merging.

Table 2: Text encoders and their Hugging Face checkpoint identifiers. Experiment names are aliases for the listed checkpoints; in particular,

Experiment name	Hugging Face checkpoint
`qwen3-embedding-0.6b`	`Qwen/Qwen3-Embedding-0.6B`
`qwen3-embedding-4b`	`Qwen/Qwen3-Embedding-4B`
`qwen3-embedding-8b`	`Qwen/Qwen3-Embedding-8B`
`e5-v1-small`	`intfloat/e5-small`
`e5-v1-base`	`intfloat/e5-base`
`e5-v1-large`	`intfloat/e5-large`
`e5-v2-small`	`intfloat/e5-small-v2`
`e5-v2-base`	`intfloat/e5-base-v2`
`e5-v2-large`	`intfloat/e5-large-v2`
`gte-small`	`thenlper/gte-small`
`gte-base`	`thenlper/gte-base`
`gte-large`	`thenlper/gte-large`
`bge-small-en-v1.5`	`BAAI/bge-small-en-v1.5`
`bge-base-en-v1.5`	`BAAI/bge-base-en-v1.5`
`bge-large-en-v1.5`	`BAAI/bge-large-en-v1.5`
`bert-mini`	`google/bert_uncased_L-4_H-256_A-4`
`bert-small`	`google/bert_uncased_L-4_H-512_A-8`
`bert-medium`	`google/bert_uncased_L-8_H-512_A-8`

Prompt templates. All text models use the same 14 templates and category vocabulary. Each prompt embedding is individually ℓ_2-normalized; the 14 embeddings are then averaged and normalized again. We substitute each cleaned category name for `{label}` in the following templates, preserving the capitalization and punctuation shown:

```
{label}
a {label}
A photo of a {label}
a photo of {label}
the object is a {label}
a close-up photo of {label}
a cropped photo of {label}
a photo of the large {label}
a photo of the small {label}
a low-resolution photo of {label}
a rendering of {label}
```

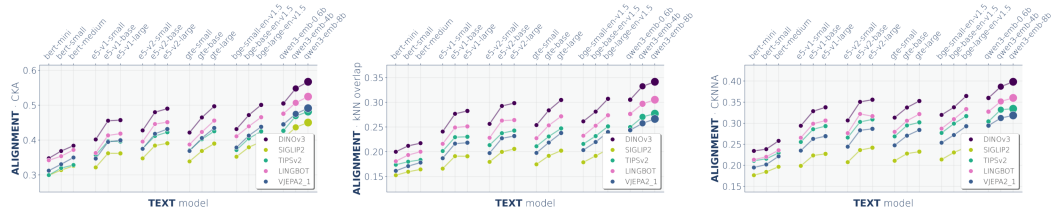

Figure 9: Dense representational alignment across vision encoder families with ViT-L backbones. Linear CKA (left), k-NN overlap (center), and CKNNA (right) are computed on the same 1,024 categories, with $k = 10$ for neighborhood metrics. Colors identify vision encoders; text encoders are grouped by family and ordered by size. Alignment generally increases with text-model size within each family, with DINOv3 attaining the highest scores across the displayed comparisons.



Figure 10: Cross-modal retrieval performance versus the number of training concepts. Recall@1 (left), Recall@5 (center), and Recall@10 (right) are averaged across text-to-vision and vision-to-text retrieval on a fixed held-out concept set. Each curve pairs DINOv3 with a different text encoder; pre-mapping CKA scores are reported in the legends. Better-aligned pairs generally achieve higher recall with fewer training concepts.

```
a drawing of {label}
art of {label}
{label} in a video game
```

A.5 Additional Dense Representational Convergence Analyses

Evaluation protocol. We use the same 1,024 categories for every encoder pair, selected from the 3,440-concept vocabulary by the largest retained-patch counts in the DINOv3 extraction. Counts are aggregated across datasets for each cleaned category as described in Appendix A.3. The selected category identities and their ordering are fixed across comparisons. We measure alignment in the original representation spaces, without learning a cross-modal mapping, using linear CKA, k-NN overlap, and CKNNA. Both neighborhood metrics use cosine similarity and $k = 10$; formal definitions appear in Appendix A.1.

Model generations and scale. The main comparisons across DINO generations are shown in Figure 2. Within DINOv3, we evaluate ViT-S/16, ViT-B/16, and ViT-L/16. The scaling panel in that figure shows increasing neighborhood overlap with larger DINOv3 models. Text encoders and their checkpoint identifiers are listed in Appendix A.4.

Additional vision encoders. We extend the comparison to SigLIP2, TIPSv2, LingBot, and V-JEPA 2.1, using ViT-L configurations for the cross-family comparison. The Hugging Face checkpoints are `google/siglip2-large-patch16-256`, `google/tipsv2-l14`, and `robbyant/lingbot-vision-vit-large`, respectively. For V-JEPA 2.1, we use `vjepa2_1_vit_large_384` from the official V-JEPA repository. Figure 9 compares their alignment with the text encoders under the shared evaluation protocol.

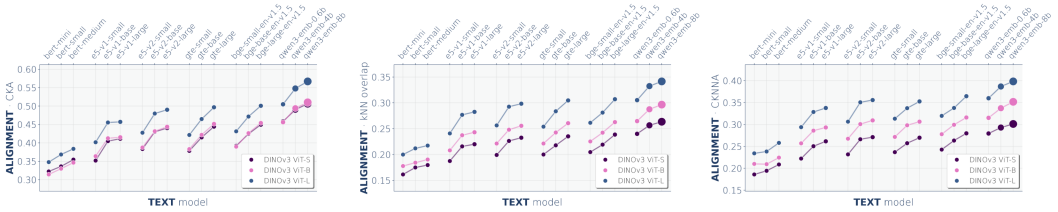

Figure 11: Dense representational alignment across DINOv3 model sizes. ViT-S, ViT-B, and ViT-L are compared with the same text encoders on a shared set of semantic categories. Alignment generally increases with vision-model size and with text-model size within each text encoder family.

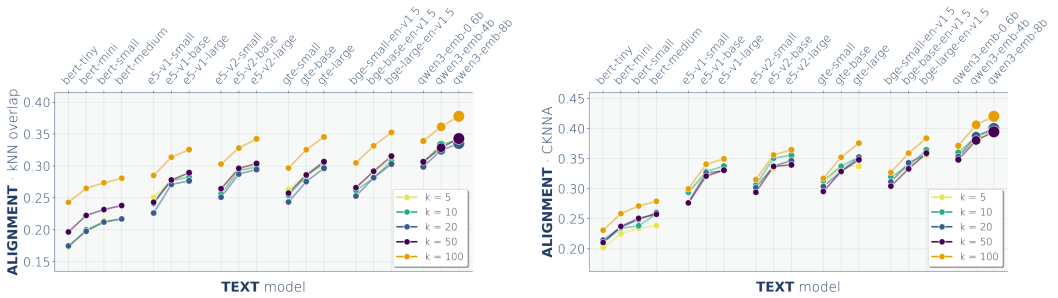

Figure 12: Sensitivity of neighborhood-based alignment to neighborhood size. k-NN overlap (left) and CKNNA (right) are reported for DINOv3 paired with different text encoders using $k \in 5, 10, 20, 50, 100$. Text encoders are grouped by family and ordered by size. Larger neighborhoods generally yield higher scores, while the relative ordering of text encoders and within-family scaling trends remain largely consistent.

Observed alignment trends. Across the vision encoders in Figure 9, alignment generally increases with text-model size within each text encoder family, following the trend observed for the DINO models. DINOv3 achieves the highest alignment scores across the displayed text-encoder comparisons for all three metrics. SigLIP2, TIPSv2, LingBot, and V-JEPA 2.1 exhibit similar within-family text-model scaling trends, although their absolute alignment scores differ.

Sensitivity to evaluation settings. We examine how alignment scores vary with neighborhood size and the number of evaluated concepts. Figure 12 reports k-NN overlap and CKNNA for $k \in \{5, 10, 20, 50, 100\}$, using DINOv3 paired with different text encoders. Larger neighborhoods generally yield higher alignment scores, while the relative ordering of text encoders and the within-family scaling trends remain largely consistent.

Figure 13 examines linear CKA, k-NN overlap, and CKNNA as the number of evaluated concepts increases. All three metrics decrease with larger concept sets, but the relative ordering of the evaluated text encoders remains broadly stable, with Qwen3-Embedding-8B consistently achieving the highest alignment. Together, these analyses show that absolute alignment scores depend on the evaluation settings, whereas the observed comparisons between text encoders are largely preserved across the tested configurations.

Null-calibrated alignment metrics. As an additional robustness analysis, we report null-calibrated versions of all three alignment metrics following the permutation-based framework of Gröger et al. (2026). Calibration provides each similarity measure with an empirical chance-level reference and controls for positive null baselines that can arise as a function of representation dimensionality and finite sample size. For each encoder pair and metric $s(\cdot, \cdot)$, we first compute the observed alignment

$$s_{\text{obs}} = s(V, T).$$

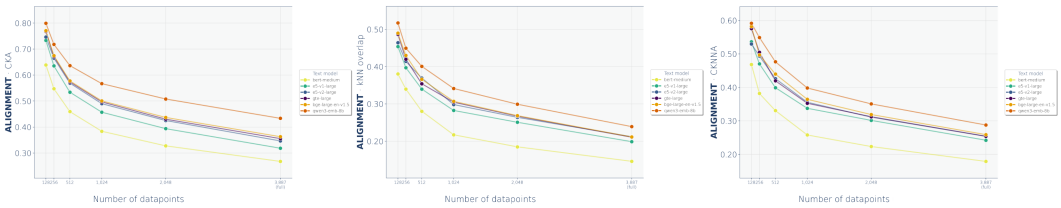

Figure 13: Sensitivity of representational alignment to the number of evaluated concepts. Linear CKA (left), k-NN overlap (center), and CKNNA (right) are computed between DINOv3 and different text encoders as the concept set grows. Alignment scores decrease with larger concept sets, while the relative ordering of text encoders remains broadly consistent across evaluation sizes.

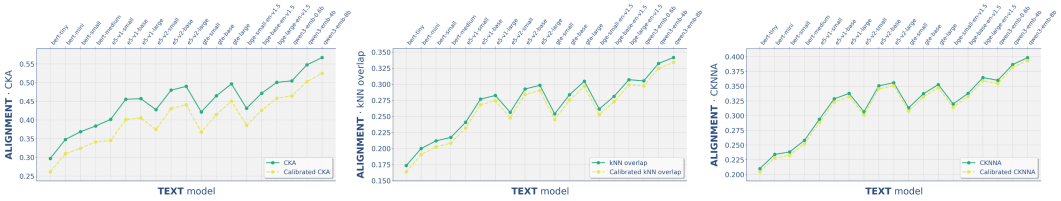

Figure 14: Comparison between the original and null-calibrated alignment metrics for DINOv3 and the selected text encoders. From left to right, we report CKA, k-NN overlap, and CKNNA; calibrated scores are obtained using the permutation-based procedure of Gröger et al. Across all three metrics, calibration shifts the absolute values while preserving the same qualitative trends across text-model families and scales, indicating that the observed alignment patterns are not driven by metric-specific null baselines.

We then generate $K = 200$ null scores by randomly permuting the correspondence between visual and textual concepts while leaving the internal geometry of each representation unchanged:

$$s^{(r)} = s(V, \pi_r(T)), \qquad r = 1, \dots, K.$$

Following Gröger et al. (2026), we combine the observed score with the K permutation scores and define the right-tail critical value at $\alpha = 0.05$ as

$$\tau_\alpha = s_{(\lceil (1-\alpha)(K+1) \rceil)},$$

where $s_{(j)}$ denotes the j-th order statistic of $\{s_{\text{obs}}, s^{(1)}, \dots, s^{(K)}\}$. Since the alignment metrics considered here have maximum value $s_{\max} = 1$, the calibrated score is

$$s_{\text{cal}} = \max\left(\frac{s_{\text{obs}} - \tau_\alpha}{1 - \tau_\alpha}, 0\right).$$

We additionally compute the corresponding permutation p-value as

$$p = \frac{1 + \sum_{r=1}^{K} \mathbb{1}\left[s^{(r)} \geq s_{\text{obs}}\right]}{K + 1}.$$

Because our analysis compares one fixed representation from each encoder rather than selecting the maximum over multiple layer pairs, we use the scalar calibration procedure rather than aggregation-aware calibration. Figure 14 compares the raw and calibrated alignment scores. Both lead to the same qualitative outcome: the cross-modal alignment trends observed with the original metrics are preserved after null calibration, indicating that the trends reported in our setting are not explained by the metric-specific permutation baseline.

A.6 Additional Semantic Segmentation Results

A.6.1 Alignment and Segmentation Performance

Figure 15 completes the analysis in Section 4.4 and Figure 6 by reporting segmentation performance against all three alignment metrics: linear CKA, k-NN overlap, and CKNNA, on PASCAL-Context, COCO-Stuff, Cityscapes, and ADE20K. Each point corresponds to DINOv3 paired with

a different text encoder. Alignment is measured before mapping, while mIoU is obtained using the Full-Mapping setting described in the main paper, with the map fitted on the complete concept vocabulary. These plots provide the per-dataset comparisons for each alignment metric under this shared segmentation protocol.

A.6.2 Open-Vocabulary Semantic Segmentation

We compare our approach with existing semantic segmentation methods on PASCAL-Context, COCO-Stuff, Cityscapes, and ADE20K, using the respective 59, 171, 19, and 150 evaluation categories. Table 3 reports mIoU for ViT-L configurations without mask refinement. Baseline scores are taken from the comparison reported by Barsellotti et al. (2025); we do not re-evaluate these methods. Both variants of our approach pair frozen DINOv3 with frozen Qwen3-Embedding-8B.

Compared methods. MaskCLIP (Zhou et al., 2022) adapts CLIP features for dense prediction. SCLIP modifies CLIP's final-layer self-attention to improve localization; ClearCLIP modifies the final-layer feature computation, including removal of the residual connection; and NACLIP incorporates spatial neighborhood information into attention. The remaining methods connect self-supervised visual features with language: `dino.txt` (Jose et al., 2025) trains a text encoder to align with frozen DINOv2, FreeDA (Barsellotti et al., 2024) constructs visual prototypes using diffusion-generated images, ProxyCLIP (Lan et al., 2025) uses self-supervised visual features to guide CLIP's dense predictions, and Talk2DINO (Barsellotti et al., 2025) learns a mapping from CLIP text embeddings to DINOv2 features.

Open-vocabulary mapping. For each evaluation dataset D, we learn a separate text-to-vision map using all available concepts except those whose category names exactly match its evaluation categories. Writing $\mathcal{C}_D$ for these categories, the mapping vocabulary is

$$\mathcal{C}_{\text{map}}^{(D)} = \mathcal{C} \setminus \mathcal{C}_D. \tag{13}$$

For ADE20K, this leaves $3{,}440 - 150 = 3{,}290$ training concepts. We fit a single linear map for each dataset using ridge regression, following the centering and regularization settings in Section 4.3. At inference, we project the textual prototypes of the excluded evaluation categories into the visual space and apply the patch scoring and bilinear upsampling procedure in Section 3.3. This setting tests transfer to category prototypes excluded from mapping training. Exclusion is based on exact category-name matches; semantically related categories may remain in the training vocabulary.

Full-Mapping with target-category supervision. This is the setting used in Section 4.4, here reported for DINOv3 paired with Qwen3-Embedding-8B. We fit the map using the complete $3{,}440$-concept vocabulary, including the evaluation categories. Supervision enters through the annotation-derived visual prototypes used to learn the mapping. Both encoders remain frozen, and inference uses the same procedure as in the open-vocabulary setting, without a learned segmentation decoder or mask refinement. We report this setting separately because its mapping has access to target-category prototypes during training.

Results. Our open-vocabulary variant achieves 28.8 average mIoU, compared with 26.1 for `dino.txt`, 27.8 for the DINOv2-only FreeDA variant, and 30.7 for Talk2DINO. Its 24.5 mIoU on ADE20K is the highest among the open-vocabulary entries in Table 3, while several methods achieve higher scores on the other datasets or in the overall average. These results show competitive transfer to held-out categories using a single ridge-regression map and direct patch classification, without a segmentation-specific architecture. Full-Mapping achieves 44.1, 30.4, 46.8, and 31.3 mIoU on PASCAL-Context, COCO-Stuff, Cityscapes, and ADE20K, respectively, for an average of 38.1. Its improvement over the open-vocabulary variant reflects the setting in which target-category prototypes are also available for fitting the map.

A.7 Zero-Shot Image Captioning Details

Semantic summary construction. We extract unit-normalized spatial tokens from the frozen DINOv3 encoder and partition them into $K = 32$ clusters using Euclidean k-means with 10 initializations and a fixed random seed across all runs. Each cluster is represented by its normalized mean

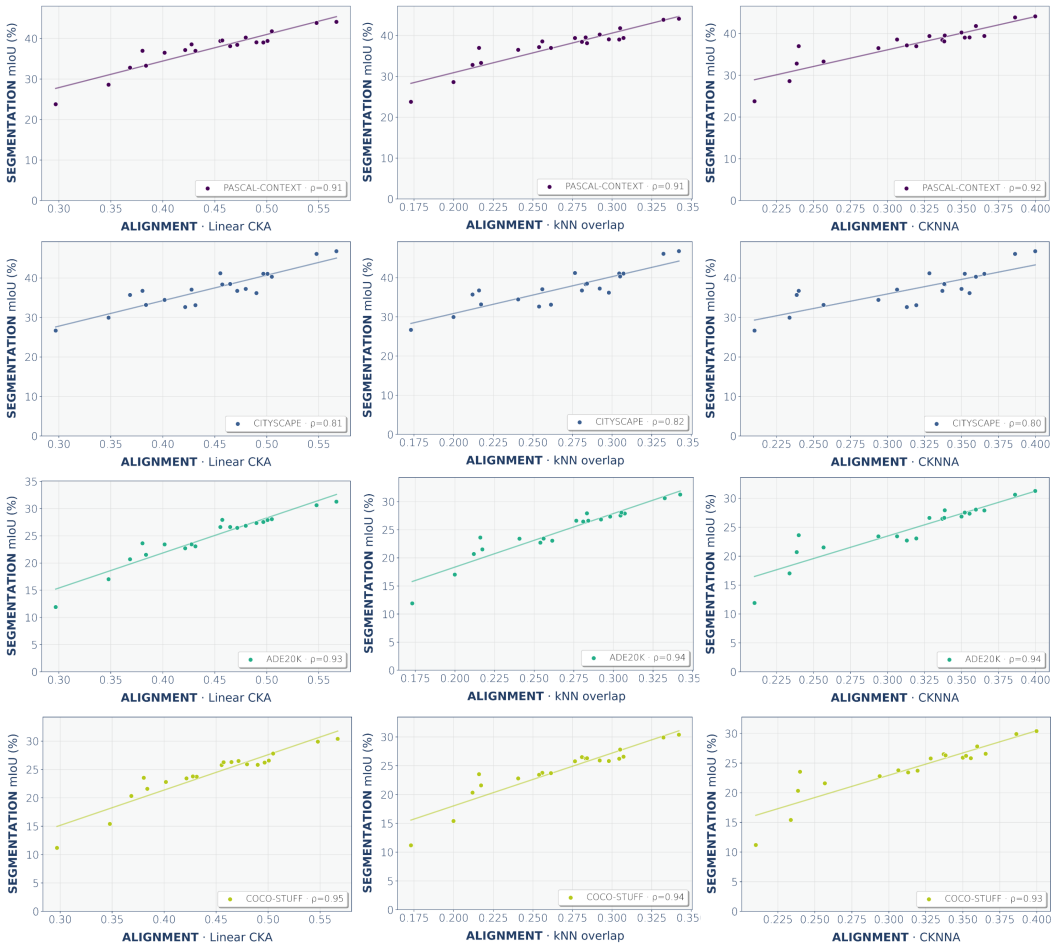

Figure 15: Intrinsic alignment and semantic segmentation performance on PASCAL-Context, COCO-Stuff, Cityscapes, and ADE20K. Each point represents DINOv3 paired with a different text encoder. Pre-mapping linear CKA, k-NN overlap, and CKNNA are compared with mIoU obtained in the Full-Mapping setting, using the complete concept vocabulary to fit the text-to-vision map.

feature and assigned the category with the highest cosine similarity among the $3{,}440$ mapped textual prototypes, following Section 3.3. Clusters assigned the same category are merged into a single summary entry, retaining the category label, the number of assigned clusters, each cluster's patch count, and its matching score. Cluster multiplicity describes the feature grouping rather than the number of object instances.

Caption generation. We supply the textual semantic summary to the frozen `Qwen/Qwen3-8B` language model, which generates a caption without access to the image, visual embeddings, or spatial coordinates. We use greedy decoding with a maximum of 64 generated tokens and disable thinking mode. No caption-specific parameters are trained.

Evaluation and additional results. We evaluate DINOv3 paired with every text encoder listed in Appendix A.4, using all images in the evaluation splits of Flickr8k and MS COCO. Caption quality is measured using BLEU-4, METEOR, CIDEr, and SPICE. Figure 16 completes the analysis in Section 4.5, reporting captioning performance against pre-mapping linear CKA, k-NN overlap, and CKNNA on both datasets. Each point corresponds to a DINOv3-text encoder pair, with the caption generator and decoding settings held fixed across comparisons.

Table 3: Semantic segmentation performance (mIoU) with ViT-L backbones and no mask refinement. Baseline scores are reproduced from Barsellotti et al. (2025). Context, Stuff, City, and ADE denote PASCAL-Context, COCO-Stuff, Cityscapes, and ADE20K; Avg is their unweighted mean. Our Open-Vocabulary variant excludes exact matches to each dataset's evaluation categories from mapping training. Full-Mapping includes these categories and is reported separately as a setting with target-category supervision. Both variants use DINOv3 and Qwen3-Embedding-8B. Bold values identify the Full-Mapping results.

		ViT-Large (mIoU)				
Model	Vision Encoder	Context	Stuff	City	ADE	Avg
Without mask refinement						
MaskCLIP	CLIP	12.4	8.8	11.5	7.2	10.0
SCLIP	CLIP	25.2	17.6	21.3	10.9	18.7
ClearCLIP	CLIP	29.6	19.9	27.9	15.0	23.1
NACLIP	CLIP	32.1	21.4	31.4	17.3	25.5
`dino.txt`	DINOv2(reg)	30.9	20.9	32.1	20.6	26.1
FreeDA	DINOv2	35.4	24.2	32.3	19.4	27.8
FreeDA	CLIP+DINOv2	39.7	26.3	33.6	21.4	30.2
ProxyCLIP	CLIP+DINOv2(reg)	36.2	24.6	35.2	21.6	29.4
ProxyCLIP	CLIP+DINO	37.7	25.6	40.1	22.6	31.5
Talk2DINO	DINOv2(reg)	39.1	27.0	35.8	21.1	30.7
Our Open-Vocabulary	DINOv3 - QWEN3-Embeddings-8B	35.3	19.7	35.7	24.5	28.8
Our Full-Mapping	DINOv3 - QWEN3-Embeddings-8B	**44.1**	**30.4**	**46.8**	**31.3**	**38.1**

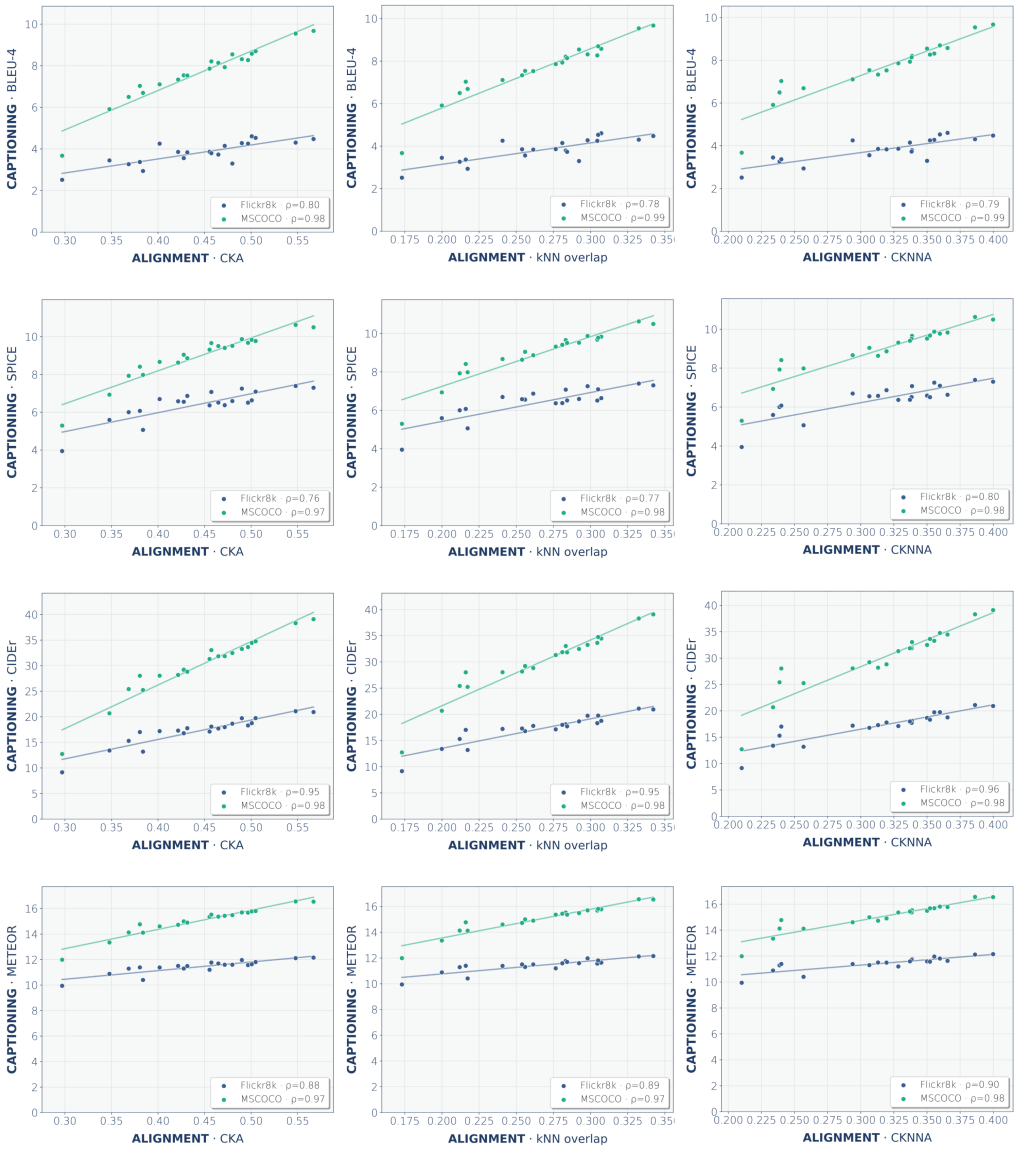

Figure 16: Intrinsic alignment and zero-shot image captioning performance on Flickr8k and MS COCO. Columns show pre-mapping linear CKA, k-NN overlap, and CKNNA; rows report BLEU-4, SPICE, CIDEr, and METEOR. Each point represents DINOv3 paired with a different text encoder, with the caption generator and decoding settings held fixed. Colors distinguish datasets, and lines show linear fits. Higher alignment is consistently associated with better captioning performance across both datasets and all four evaluation metrics.