Steerable Visual Representations

Jona Ruthardt[1*], Manu Gaur[2*],
Deva Ramanan[2], Makarand Tapaswi[3†], and Yuki M. Asano[1†]

[1] University of Technology Nuremberg, Germany
[2] Carnegie Mellon University, USA
[3] International Institute of Information Technology Hyderabad, India

Abstract. Pretrained Vision Transformers (ViTs) such as DINOv2 and MAE provide generic image features that can be applied to a variety of downstream tasks such as retrieval, classification, and segmentation. However, such representations tend to focus on the most salient visual cues in the image, with no way to direct them toward less prominent concepts of interest. In contrast, Multimodal LLMs can be guided with textual prompts, but the resulting representations tend to be language-centric and lose their effectiveness for generic visual tasks. To address this, we introduce *Steerable Visual Representations*, a new class of visual representations, whose global and local features can be steered with natural language. While most vision-language models (e.g., CLIP) fuse text with visual features after encoding (*late* fusion), we inject text directly into the layers of the visual encoder (*early* fusion) via lightweight cross-attention. We introduce benchmarks for measuring *representational steerability*, and demonstrate that our steerable visual features can focus on any desired object in an image while preserving the underlying representation quality. Our method also matches or outperforms dedicated approaches on anomaly detection and personalized object discrimination, exhibiting zero-shot generalization to out-of-distribution tasks.

Project Website: jonaruthardt.github.io/project/SteerViT

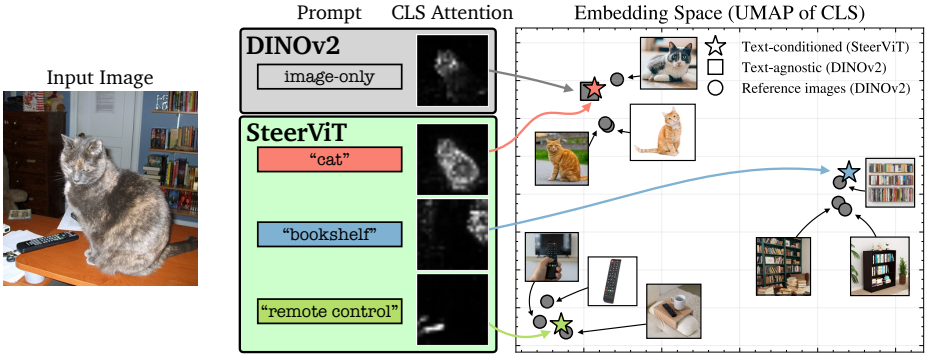


Fig. 1: Steering visual representations with language. While DINOv2 primarily encodes the salient object, producing a "cat" representation, SteerViT can be steered with text to shift its attention (middle) and global feature semantics (right) towards the queried visual concept (e.g., "bookshelf" or "remote control").

[*] Equal contribution. [†] Equal advising.

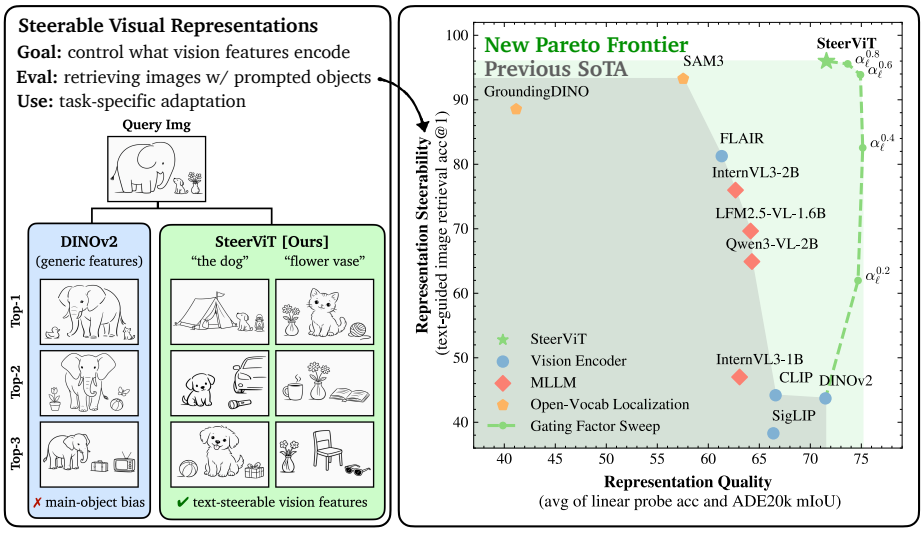


Fig. 2: SteerViT produces high-quality visual representations that can be steered by text. Left: Traditional (non-steerable) representations like DINOv2 tend to focus on the dominant object in an image and retrieve images with the same object. SteerViT can adapt to a text prompt, enabling retrieval of images even with small objects of interest. **Right:** We compare SteerViT to prior work in terms of its ability to adapt to text (measured by text-guided image retrieval (cf. Sec. 4.1)) and the quality of the visual representation (measured by the accuracy of linear probing for the CLS feature and semantic segmentation for patch features). While models typically trade off steerability for representation quality, SteerViT preserves both. By modulating a gating factor (Eq. (2)), SteerViT achieves a new Pareto frontier.

1 Introduction

Pretrained Vision Transformers (ViTs) such as DINOv2 [25], MAE [9], and SigLIP [35] provide a generic representation of an image that can be applied to a variety of downstream tasks such as retrieval, classification, and segmentation. However, such representations tend to focus on the most prominent object in the image, likely due to the well-known photographer bias and object-centric vision datasets [30]. Consider the indoor scene from Fig. 1: DINOv2 encodes the image focusing on the dominant salient object ("cat"), neglecting smaller or less prominent objects like the "remote control" or "bookshelf".

Given the lack of other input, focusing on the dominant object is reasonable. However, tasks such as fine-grained localization may require greater consideration of less prominent visual concepts. We argue that generic visual representations that can be *steered* using task-specific priors have significant utility.

In this work, we seek to steer pretrained vision transformers with natural language. We posit two desiderata that steerable representations should satisfy:

– **Steerability:** The representation should adapt to the input text. In particular, it should be steerable in *what* it encodes (so that it captures objects

Table 1: Only SteerViT satisfies both desiderata. ✓= satisfied, ✗ = not, ◐ = partial. *Trainable MM Params* reflects multimodal training; unimodal ViTs train solely on images. *V-L Fusion* notes where modalities interact relative to vision encoder layers.

Method Family	Text Steerable	Feature Quality	V-L Fusion	Trainable MM Params
Unimodal ViT (DINOv2)	✗	✓	✗	0
Cross-modal (CLIP)	✗	✓	Late	∼200M
OV Localize (SAM3)	✓	✗	Late	∼200M–1B
MLLM (Qwen3-VL)	◐	◐	Late, in LLM	≥1B
SteerViT (Ours)	✓	✓	Early, in ViT	**21M**

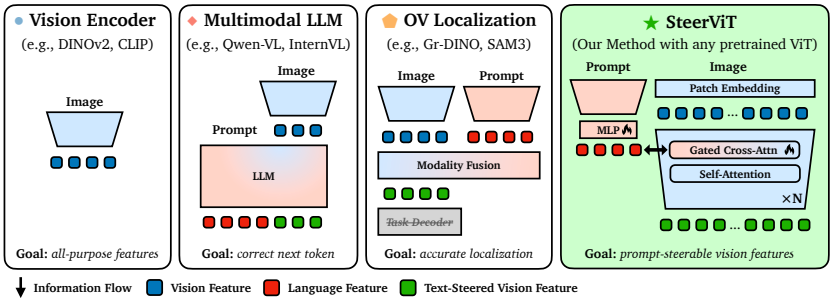


Fig. 3: Taxonomy of visual encoding. Standard vision encoders produce query-agnostic visual features. MLLMs and OV Localization models *late fuse* text after the visual encoder, modeling vision-language interactions inside the LLM or task-aligned encoder. SteerViT, instead, directly steers the internal features of a frozen ViT using text prompts (*early fusion*) via lightweight cross-attention layers.

irrespective of saliency; Sec. 4.1) and *how* it organizes the embedding space (so that clusters can be defined by specific attributes such as supercategories or object parts; Sec. 4.5).

– **Representation quality:** The steered representation should support diverse visual tasks, such as retrieval, classification, and segmentation (Sec. 4.3).

When considering prior art, we find that no existing approach simultaneously satisfies these desiderata (see Tab. 1). Most vision-language architectures first encode images independently and only later fuse the modalities (see Fig. 3). As a result, text typically cannot influence the visual encoding process at inference time and can only operate on the vision encoder's fixed outputs. This stands in contrast to human vision, where prior textual priming can change how people parse images, often through top-down, task-guided attention [5]. While Multimodal LLMs (MLLMs) come closest to steerable vision-language representations, fusion in early layers of the language model typically yields language-dominant multimodal representations with diminished visual fidelity and limited controllability, as illustrated in Fig. 2.

With Steerable Visual Representations (SteerViT), we *invert* the MLLM paradigm and condition a visual encoder on language input, producing a *vision-centric multimodal representation*. Specifically, we interleave lightweight trainable cross-attention layers [1] within frozen ViT blocks that attend to text prompts. This allows the visual encoding process to be heavily influenced or

"steered" by text. We adopt referential image segmentation as the training objective to encourage vision-language alignment.

Our method achieves a Pareto improvement over prior approaches (Fig. 2), producing visual features that are steerable with text while preserving their underlying representation quality. This is accomplished by freezing both the visual and text encoders and adding only 21M trainable parameters via cross-attention, two orders of magnitude fewer than MLLMs (see Tab. 1).

Analogous to how prompting adapts (M)LLMs to novel tasks without retraining, language can adapt our *steerable visual representations* to novel domains without fine-tuning. Across diverse tasks spanning text-guided image retrieval (Sec. 4.1), personalized object discrimination (Sec. 4.4), and industrial anomaly detection (Sec. 4.6), SteerViT, without task-specific training, matches or outperforms strong baselines, including DINOv2, SAM3, billion-scale MLLMs and even some dedicated methods. These results suggest that conditioning vision on language – rather than language on vision – could be a new paradigm for efficient multimodal vision understanding.

In summary, we make the following contributions:

1. We introduce Steerable Visual Representations (SteerViT), a framework that equips *any* pretrained visual encoder with text-steerable representations via a simple grounding pretext task, adding only 21M parameters into the ViT.
2. We show that SteerViT achieves a Pareto improvement over prior approaches, steering visual features with text while preserving feature quality.
3. We demonstrate that text prompts enable zero-shot generalization, allowing SteerViT to transfer to new domains (e.g., personalized object discrimination or industrial anomaly detection) without task-specific training.

2 Related Work

Visual representation families. We summarize popular approaches against our desiderata in Tab. 1 and compare their architectures in Fig. 3. Unimodal self-supervised encoders (DINOv2 [25], MAE [9]) learn rich visual features but are query-agnostic. Cross-modal encoders (CLIP [26], SigLIP [35]) use text to provide training supervision; the visual encoder still cannot be steered with text. MLLMs come closest with moderate steerability and visual quality, but their features reside in language space and require billions of parameters. SteerViT inverts this paradigm: we condition vision on language, add only ∼21M trainable parameters to a frozen ViT, while yielding stronger visual features (Fig. 2).

Text-conditioned visual features. To our knowledge, *no prior work* steers a visual encoder effectively with text while preserving its representation quality. The closest attempt, FLAIR [33], applies text-conditioned attention pooling over a frozen SigLIP encoder (late fusion), resulting in suboptimal steerability and underperforming unimodal encoders on standard vision benchmarks (Fig. 2). Concurrent works condition visual features on text but target narrow pipelines. TIE [29] injects query tokens into the image encoder to reduce visual tokens in MLLMs, optimizing for document understanding. ELIP [36] prepends text in

the ViT to improve text-to-image retrieval re-ranking. TEVI [22] edits CLIP's final image features via caption-conditional SAE masking to suppress irrelevant information. In contrast, SteerViT is a general framework for producing steerable visual representations that transfer across a wide variety of tasks.

3 SteerViT: Steering Vision Transformers with Text

We describe the architectural modifications and post-training to obtain steerable representations from a pretrained ViT.

3.1 Architecture

As illustrated in Fig. 4, our SteerViT consists of four components:

A. Visual encoder. For an image $X_v \in \mathbb{R}^{H\times W\times 3}$, a ViT produces a sequence of N patch tokens $Z_v \in \mathbb{R}^{N\times d_v}$ and optionally a `[CLS]` token (d_v is the embedding dimension). All original ViT parameters remain frozen throughout training and new capacity results exclusively through the interleaved cross-attention layers described below. While most experiments adopt DINOv2 ViT-B/14 [25] as our backbone, we show that our approach also improves steerability of SigLIP and MAE.

B. Text encoder. We adopt a frozen, pretrained text encoder (RoBERTa-Large [20]) to produce token-level embeddings $Z_t \in \mathbb{R}^{L\times d_t}$ for a given input conditioning prompt X_t, where L is a variable number of text tokens and d_t is the text embedding dimension.

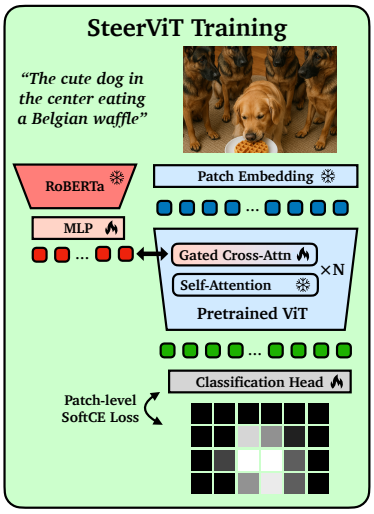


Fig. 4: Steering any ViT using text conditioning. Our method adds lightweight **vision**-to-**language** cross-attention layers within pretrained ViT blocks and applies a patch-level segmentation proxy objective to fuse prompt cues into patch tokens.

C. Multimodal adapter. Each text embedding Z_t^i is ℓ_2-normalized and fed through a trainable two-layer MLP that projects the sequence into a visual-aligned embedding space $H_t \in \mathbb{R}^{L\times d_v}$.

D. Gated cross-attention layers. To fuse textual conditioning into the ViT's residual stream, we invert the gated cross-attention (CA) formulation from Flamingo [1] by allowing hidden vision states to attend to language tokens (CA is language → vision in [1]). We insert CA layers into every other Transformer encoder block (e.g., 6 CA layers for 12 ViT-B blocks). The visual patch tokens $Z_v^{(\ell)}$ at layer ℓ are queries and the adapted text tokens H_t are keys and values:

$$\hat{Z}_v^{(\ell)} = \mathrm{CA}(Z_v^{(\ell)}, H_t) = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V, \quad Q = Z_v^{(\ell)}W_Q,\ K = H_tW_K,\ V = H_tW_V. \tag{1}$$

The output is integrated into the residual stream through a tanh gate with a layer-specific learnable scalar α_ℓ, initialized to zero:

$$Z_v^{(\ell+1)} = Z_v^{(\ell)} + \tanh(\alpha_\ell)\cdot \hat{Z}_v^{(\ell)}. \tag{2}$$

Since $\tanh(0) = 0$, the model is identical to the frozen ViT at initialization. Despite the zero-initialization, the gate still receives a learning signal since $\frac{\partial Z_v^{(\ell+1)}}{\partial \alpha_\ell} = \operatorname{sech}^2(\alpha_\ell) \cdot \hat{Z}_v^{(\ell)}$, and $\operatorname{sech}^2(0) = 1$. Thus, α_ℓ can move away from zero during optimization, gradually activating the conditioning pathway and allowing the model to incorporate language-based clues.

3.2 Training Objective

In order to encourage the vision encoder to leverage and incorporate language clues, we design a pretext task requiring consideration of the prompt to be solved. We select referential segmentation for this purpose, where, given an image X_v and a prompt X_t referring to a target object or entity, the model predicts which patches correspond to the referred region.

The ground-truth y_i is the fraction of foreground pixels of a pixel-wise binary segmentation mask that is patchified to match the ViT's $n \times n$ grid. A linear classifier head maps each patch representation $Z_v^i \in \mathbb{R}^d$ to a mask probability p_i through softmax. We adopt the soft cross-entropy loss to train our model:

$$\mathcal{L} = -\sum_{i=1}^{n \times n} y_i \log p_i \,, \tag{3}$$

as it encourages the CA layers to route textual information into the corresponding visual patch tokens, producing steered representations. Performing segmentation on a patch- rather than pixel-level reduces the training complexity and forgoes the need for pixel-level decoders.

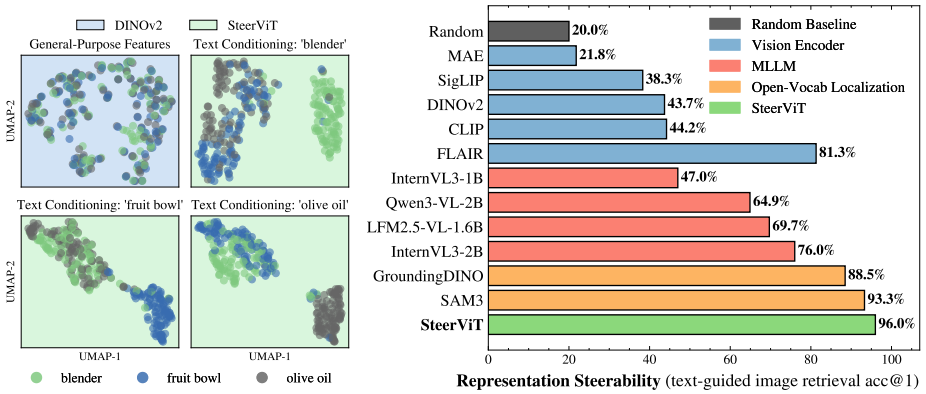
3.3 Training Data

We train on a mixture of referential segmentation and grounding datasets with diverse visual domains and textual expressions, comprising 162k unique images and 2.28M image-text pairs. Specifically, we use RefCOCO/+/g [14, 34], Visual Genome [16], LVIS [8], and Mapillary Vistas [24]. Details in the Supplement.

4 Experiments

In this section, we empirically validate properties of our learned steerable representations and show applications to diverse downstream tasks.

Baselines. We compare SteerViT (w/ DINOv2 ViT-B/14) against multiple model families that differ in how, and whether, they incorporate text into visual features (see Supplement for feature-extraction details): (1) **Unimodal vision encoders**: DINOv2 [25] and MAE [9] produce query-agnostic features with no text conditioning. (2) **Cross-modal encoders**: CLIP [26] and SigLIP [35]; we fuse visual and text embeddings via post-hoc element-wise addition (late fusion). (3) **MLLMs**: InternVL3 [39], Qwen3-VL [3] and LFM-2.5-VL [2]; we extract prompt-specified vision features by following the last-token summary pooling of E5-V [13]. (4) **Open-vocabulary (OV) localization**: SAM3 [6] and GroundingDINO [19]; we use and evaluate the intermediate multimodal state. Where supported, models process images at 336×336 resolution.



(a) Feature distribution across three object types in the "kitchen" scene. **(b)** Non-salient object retrieval performance on CORE across model families.

Fig. 5: COnditional REtrieval (CORE) benchmark. Left: While DINOv2 features form scene-level clusters, appropriate prompting of SteerViT yields object-specific clusters. **Right:** Substantial differences in steerability between model families exist, with OV localization methods and SteerViT offering the greatest adaptability.

4.1 COnditional REtrieval: Steering Global Semantics with Text

We propose *CORE (COnditional REtrieval)*, a text-conditioned image retrieval benchmark to measure how well a model steers its global features with text.

We select 100 images each for three indoor and three outdoor scenes from the SUN397 dataset [32] and inpaint five objects contextually fitting each scene into each image using the FLUX.2 image editing model [17] (e.g., a fruit bowl in a kitchen; a backpack in the park). We frame the problem as one-vs-all retrieval: given a query image containing an inpainted object Ω in scene S (e.g., a fruit bowl in a kitchen), the goal is to retrieve other images of S that also contain Ω. This quantifies how well a model can steer its global features away from shared scene-level similarities (e.g., all images depict a kitchen) toward a specified non-salient object (e.g., the fruit bowl). Both query and gallery images are encoded while conditioned on the same brief description of Ω. We measure the top-1 retrieval accuracy over the remaining 495 samples of S with non-identical underlying images (pre-editing). Details about the experimental setup and results on a per-scene basis are reported in the Supplement.

Query-agnostic encoders collapse to salient concepts. Query-agnostic encoders fail at conditional retrieval because their features collapse to the dominant scene concept (Fig. 5). MAE barely exceeds random chance (20%) and DINOv2, despite its object-centric representations, achieves only 44% acc@1. Cross-modal visual encoders (CLIP, SigLIP) also perform poorly despite operating in a shared vision-language space. In contrast, SteerViT achieves 96% retrieval accuracy, confirming that text conditioning shifts the global representation from the scene level ("kitchen") to the queried concept ("fruit bowl").

Fig. 5a illustrates this on "kitchen" scenes for three objects via UMAP [23]. DINOv2 embeddings (top left) show no object separability, whereas our text-

conditioned embeddings form clusters for different text prompts. This confirms that SteerViT can reorganize its embedding space around the queried concept.

Late fusion does not enable steerability. Although CLIP and SigLIP operate in a shared vision-language space, their visual features are extracted independently of the query. Post-hoc element-wise addition of text yields a negligible 0.02% boost over their vision-only representations, confirming that late fusion cannot steer frozen visual features. In contrast, by conditioning intermediate representations on text (i.e., early fusion), SteerViT improves retrieval accuracy from 43.7% (vanilla DINOv2) to 96.0%. While FLAIR's trained attention pooling offers greater steerability (81.3%) than typical cross-modal encoders, it still falls short of SteerViT by 14.7 points.

MLLMs and OV models are steerable, but inefficient or specialized. MLLMs can moderately steer their visual features but at substantial computational cost. SteerViT outperforms both InternVL3-1B and InternVL3-2B by 49 and 20 percentage points, respectively, while only adding 21M parameters via cross-attention blocks compared to billion-parameter-scale LLMs (Tab. 1). Open-vocabulary localization models (GroundingDINO, SAM3) are highly steerable, with SAM3 nearly matching SteerViT's retrieval accuracy. However, their intermediate representations are optimized for localization and lack the generality needed for downstream transfer, as discussed in Sec. 4.3.

Conditioning on a random class breaks retrieval. To verify that steerability is genuinely text-driven rather than an artifact of training, we condition each model on a random (incorrect) object class. Doing this has no effect on CLIP and SigLIP performance, confirming their features are unconditionally visual and not steerable. However, performance degrades drastically for FLAIR and SteerViT (-29.4 and -48.3 percentage points), indicating strong text-dependence and corroborating that steerability is primarily prompt-driven. OV models see similar large drops, consistent with their deep text integration. MLLMs show only mild sensitivity, with InternVL3-1B declining by -7.6%.

In summary, SteerViT and OV localization models can reliably be steered through text whereas standard ViTs collapse to salient concepts, with post-hoc late fusion providing no benefits. Although MLLMs offer moderate steerability, they bear significant computational cost and diminished visual feature quality. Beyond CORE, SteerViT also transfers to conditional retrieval on unaltered natural images in the GeneCIS benchmark [31], reaching 25.4% R@1 versus 9.6% for DINOv2 and 18.7% for the task-specialized baseline (details in the Supplement).

4.2 MOSAIC Localization: Text enables Targeted Attention

We examine how SteerViT routes and aggregates global representations via attention to query-relevant tokens. For this, we construct a benchmark by stitching together four images from PASCAL-VOC [7] into a single 2×2 mosaic, resulting in a total of 363 composite images with reduced saliency of each primary subject. The `[CLS]`-to-patch attention scores in the final self-attention block highlight how global attention is redirected to regions specified by text prompt.

Fig. 6: Text enables targeted attention. Attention maps on a four-image mosaic demonstrate that text conditioning with SteerViT redirects self-attention to the queried concept whereas DINOv2 attends to the most prominent objects. Note that the `[CLS]` token of SteerViT was not directly optimized for targeted attention and remains frozen in its original state.

Qualitative analysis. As revealed by Fig. 6, DINOv2's attention map focuses primarily on the prominent "pony" and "airplane" entities, confirming the saliency bias hypothesized in Sec. 1. In contrast, SteerViT conditioned on "person" routes attention to the barely visible person in the top-left. When prompted with "potted plant", it distributes attention among the class instances in the bottom-right image. Additional qualitative examples are in the Supplement.

Quantitative evaluation. We quantify this effect by measuring the area under the precision-recall curve (PR-AUC) of the attention heatmaps and the ground-truth segmentation masks for each object type appearing in the mosaic instance. DINOv2 cannot be actively steered and primarily focuses on prominent objects, resulting in a low PR-AUC of 14.3%. SteerViT can be steered with text to focus on objects of interest, achieving a substantially higher score of 50.2%.

> ***Finding 1.*** SteerViT steers its visual features with text towards queried concepts, whereas standard encoders collapse to prominent visual cues.

4.3 Preserving Visual Representation Quality while Steering

While previous sections revealed steerability for SteerViT and OV localization, this is only useful if it preserves the downstream transfer, an important property of good visual representations. A naïve solution that overwrites visual features with text yields perfect steerability but destroys the underlying representation.

To assess this trade-off, we contrast CORE performance (cf. Sec. 4.1) with vision-centric downstream tasks intended to measure representational quality. In addition to training linear probes on global features across three fine-grained classification datasets (ImageWoof [11], Waterbirds [27], StanfordCars [15]), binary object-of-interest segmentation on ADE20k [37] gauges the dense encoding performance. Here, promptable models are conditioned on the superclass (e.g., "dog" for ImageWoof) and the ADE20k class name, respectively. For Fig. 2 and 7, both scores are averaged. Additional details are in the Supplement.

Steerability $\leftrightarrow$ Representation quality trade-off. Fig. 2 (right) maps models along these two axes, revealing three regimes: (1) **Open-vocabulary localization methods** (SAM3 and GroundingDINO) achieve high steerability but produce localization-specific features that score poorly on generic vision tasks.

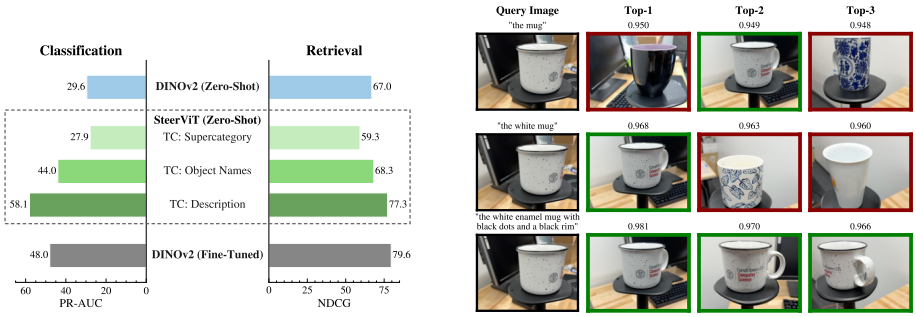


Fig. 8: Text controls feature granularity. Left: The quality of personalized representations produced by SteerViT improves drastically with more detailed prompts for text conditioning (TC), even surpassing a supervised fine-tuned DINOv2. **Right:** Retrieval improves when SteerViT is given more detailed descriptions.

(2) **MLLMs** perform reasonably on classification but falter on dense prediction tasks and require billions of parameters for moderate steerability. (3) **Query-agnostic encoders** (DINOv2, SigLIP) produce rich transferable features but cannot be steered. SteerViT bridges this gap, achieving high steerability while fully preserving the representation quality of the underlying ViT.

Cross-attention gate as a continuous control knob. The tanh-gated cross-attention mechanism (Eq. (2)) provides an implicit control knob for text conditioning strength. At inference, we can scale the learned gating parameters α_ℓ by a factor $\omega \in [0, 1]$ to smoothly interpolate between the unaltered ViT subspace and a fully text-conditioned state. Plotting this trajectory (Fig. 2, Fig. 7) reveals a clear Pareto frontier with an optimal operating point at a scaling factor of ω=0.6 for DINOv2 and SigLIP, where both slightly exceed the original ViT's representation quality while unlocking high steerability. Most strikingly, for MAE, representation quality monotonically improves as ω increases, rising from 40 at ω=0.0 to 50 points at ω=0.6. Text conditioning enriches MAE's features with semantic structure, making them more transferable.

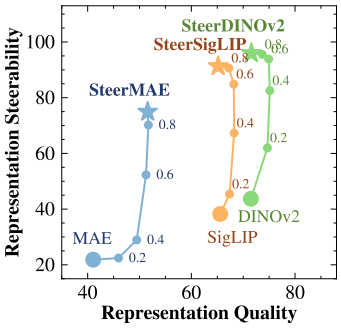


Fig. 7: Post-hoc modulation. Scaling CA gates α_ℓ at inference for continuous interpolation between vanilla ViT and SteerViT properties. A factor of 0.6 provides the optimal steerability-quality trade-off for SigLIP- and DINOv2-based models.

Finding 2. SteerViT produces the first family of visual representations that can be steered with text without sacrificing the representation quality of the underlying vision encoder.

4.4 Text Specificity Guides Semantic Granularity

Next, we explore the role of text in forming high-fidelity visual representations. For this, we choose Personal Object Discrimination Suite (PODS) [28], a benchmark evaluating the formation of instance-aware feature spaces via the ability of models to recognize particular objects (e.g., *your* mug vs. all other mugs). Given a small set of reference images, PODS computes cosine similarities on frozen global features and performs: (i) one-vs-all instance classification as well as (ii) similarity-based retrieval over a set of test images. Following [28], we report PR-AUC for classification and NDCG for retrieval.

From coarse to fine-grained steering. We vary the amount of detail provided to SteerViT via the text prompt from coarse supercategories (e.g., "mug", "shoe"), object names (e.g., "white ECCV mug"), to comprehensive MLLM-generated descriptions of each instance's visual appearance. As shown in Fig. 8, the semantics of the steered representations are sensitive to the prompt specificity. When conditioning is too coarse, the model overlooks fine-grained cues necessary for discriminating instances within an object category, yielding slightly worse performance than vanilla DINOv2 (27.9% vs. 29.6% PR-AUC). Enriching the prompt with instance-level descriptions substantially boosts performance to 58.1% PR-AUC, surpassing custom DINOv2 variants fine-tuned on synthetic task-specific data (48.0% PR-AUC) and nearly closing the gap between zero-shot and fine-tuned DINOv2 on retrieval (77.3% vs. 79.6% NDCG). This result is particularly noteworthy given that DINOv2 fine-tuning is object-specific, requiring a separate model per object class: 100 models in this setting compared to a single SteerViT model. These results demonstrate that SteerViT does not simply *add* information to the visual encoder but precisely controls the granularity of its visual features through the level of detail in the text prompt.

Finding 3. The level of detail in text conditioning directly controls the granularity of the *steerable visual representations*.

4.5 Visualizing and Analyzing the Steered Embedding Space

The results on PODS (Sec. 4.4) indicate that query specificity dictates the level of semantic abstraction of *steerable visual representations*. Next, we explore how text conditioning shapes the embedding space topology of SteerViT using two complementary modes: steering along the *semantic hierarchy* and steering by *compositional attributes*.

Here, we select 500 images containing exactly one out of eight curated classes from PASCAL-VOC [7]. Encoded features are reduced to 2D using UMAP [23].

Steering Along the Semantic Hierarchy. As shown in Fig. 9(1), DINOv2 embeddings cluster rigidly by object class. Conditioned on "animal" (2), two macro-clusters emerge for SteerViT, containing all animal and non-animal classes, respectively. Crucially, fine-grained structure is preserved (e.g., dogs and cats remain closer than birds), confirming that text controls clustering granularity

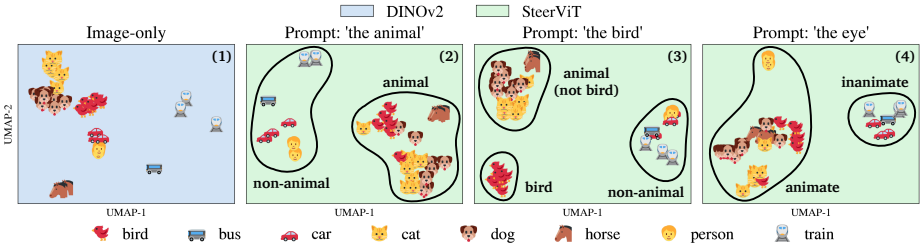


Fig. 9: SteerViT steers embedding topology via text. We show sub-sampled UMAP projections of 500 images across eight PASCAL-VOC classes. **(1)** DINOv2 clusters by object class. **(2)** Conditioning SteerViT on "animal" merges animal classes into one cluster while separating non-animals. **(3)** Conditioning on "bird" increases bird vs. non-bird separability while preserving the animal/non-animal macro-structure. **(4)** Conditioning on "eye" groups all animate classes (including "person", clustered with inanimate classes previously) together, demonstrating compositional attribute steering.

without destroying object-level semantics. When conditioned on "bird" (3) separability between bird and non-bird images increases while preserving the higher-level animal versus non-animal macro-separation. This confirms that text can steer clustering at multiple levels of the semantic hierarchy, an emergent behavior not actively encouraged during training.

Steering by Compositional Attributes. Beyond semantic categories, text conditioning enables arbitrary grouping criteria, e.g., by shared object parts, defining a grouping principle orthogonal to semantic categories. Conditioning on "eye" (Fig. 9(4)) produces two macro clusters: animate classes that possess eyes versus inanimate objects that do not. Notably, *person* images, which were farther from animals under previous conditioning, now cluster together with them as the shared attribute "has eyes" overrides the semantic category boundary. This demonstrates that we can steer representations by compositional properties, not just taxonomy.

Finding 4. SteerViT can reorganize its embedding space using text, controlling both the level of semantic abstraction and clustering criteria.

4.6 Text Facilitates Domain Transfer

The previous sections showed that SteerViT uses text to steer global features and route attention to queried objects, while preserving the representation quality of the underlying ViT. We now ask: *Does this language-conditioned flexibility translate into generalization to novel downstream tasks and unseen domains?*

Besides generalization to novel tasks like PODS (shown in Sec. 4.4), we showcase SteerViT's adaptability to extreme out-of-distribution settings by performing anomaly segmentation (AS) on the industrial MVTec AD [4] dataset. Here, models are conditioned on prompts "the anomaly in the <object>" and anomaly maps are derived by upsampling the heatmaps produced by the learned linear segmentation head (Fig. 10). We compare SteerViT to other segmentation models and dedicated zero-shot AS methods using *Per-Region-Overlap* (PRO). More

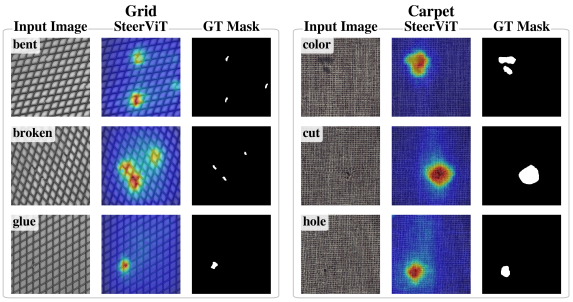


Fig. 10: Anomaly segmentation heatmap produced by SteerViT. Language-conditioning enables robust zero-shot generalization to this OOD task.

Method	PRO
MaskCLIP [38]	40.5
CLIPseg [21]	34.6
SAM3 [6]	54.5
WinCLIP [12]	64.6
DIVAD [10]	73.3
FADE [18]	84.5
SteerViT	82.1

Table 2: ZS anomaly segmentation (MVTec). Dedicated methods in gray.

details are provided in the Supplement. Tab. 2 shows SteerViT matching dedicated methods despite the extreme OOD setting. These results reinforce the pattern of text-driven steering unlocking capabilities already present in the frozen ViT and transferring them to domains beyond the training distribution.

Finding 5. SteerViT can use natural language to transfer rich unimodal vision encoders to OOD domains without task-specific training.

4.7 Analysis of SteerViT

We ablate three key architectural and training decisions in SteerViT to isolate the contribution of each. More ablations are in the Supplement. Unless otherwise noted, experiments use the DINOv2 ViT-B/14 backbone and the RoBERTa-Large language model and are trained at 336^2 resolution with a batch size of 12 for 500k iterations (∼84 H100 GPU-hours) on the full data mixture (Sec. 3.3), using AdamW with a cosine schedule that warms up to $3{\times}10^{-4}$ over 5k steps, decays to $3{\times}10^{-5}$ by 40k steps, and remains constant thereafter. We test fine-grained classification probe accuracy (FG-CLS; Sec. 4.3), referential segmentation performance (ADE; Sec. 4.3), steerability via CORE (Sec. 4.1), and PR-AUC on PODS with descriptive prompts (Sec. 4.4).

Table 3: We separately ablate SteerViT's three principal design choices (marked ✗).

Early Fusion	Tanh Gate	MLP Proj.	FG-CLS ↑	ADE ↑	CORE ↑	PODS ↑
–	–	–	89.0	53.7	43.7	29.6
✓	✓	✓	87.7	55.4	**96.0**	**58.1**
✗	✓	✓	**91.8**	**55.5**	93.3	36.6
✓	✗	✓	83.5	55.3	94.6	47.1
✓	✓	✗	86.7	54.5	95.2	56.4

✗ in Early Fusion means late fusion. ✗ in Tanh Gate means ungated cross-attention. ✗ in MLP Proj. means single linear layer. Vanilla DINOv2 (top, gray) is vision-only.

Architecture Choices. SteerViT has three principal design choices: early fusion of text within the ViT, tanh-gated CA layers, and MLP text projector. Tab. 3 ablates each component.

Early vs. late fusion. We interleave gated cross-attention within ViT layers (early fusion). Late fusion (Tab. 3, row 3) adds text only after the final ViT layer. While late fusion (row 3) also achieves high steerability (93.3) and higher FG-CLS (91.8 vs. 87.7), it reduces PODS performance dramatically (36.6 vs. 58.1). This illustrates the importance of early fusion for fine-grained tasks as the gap vanishes (26.5 vs. 27.9) when conditioning on coarse supercategory prompts.

Role of gating. Removing the zero-initialized tanh gate (row 4) reduces FG-CLS, CORE, and PODS by 4.2, 1.4, and 11.0 points below SteerViT (row 2), showing that ungated cross-attention disrupts frozen features.

Language projector. A two-layer MLP projects RoBERTa features to the ViT space (Sec. 3.1). Replacing it with a linear projector (row 5) lowers FG-CLS by 1.0 and PODS by 1.7 points, indicating that the MLP better aligns modalities.

Generalization Across Pretrained Backbones. To validate that our approach generalizes beyond DINOv2, we apply SteerViT to two additional ViTs: SigLIP [35] and MAE [9]. Tab. 4 compares vanilla, late fusion, and early fusion (SteerViT) variants.

Table 4: Steering different ViT pretraining approaches. SteerViT consistently outperforms late fusion, with the largest gains on weaker backbones.

Backbone	CORE ↑		
	Base	Late	SteerViT
DINOv2	43.7	93.3	**96.0**
SigLIP	38.3	75.4	**91.3**
MAE	21.8	41.0	**74.9**

Overall, DINOv2 produces the strongest steerable representations, consistent with its richer self-supervised features. SigLIP and MAE similarly benefit from text conditioning but start from weaker baselines. Early fusion consistently outperforms late fusion across all three ViT families. The advantage is more pronounced for MAE and SigLIP, where early fusion boosts steerability by 33.9 and 15.9 points, respectively. These larger early-fusion gains compared to DINOv2 (2.7) align with our layer-wise analysis in the Supplement, where SteerViT representations diverge earlier from their base ViTs for MAE and SigLIP, suggesting that language injection is beneficial when the visual features are less semantically mature.

5 Conclusion

We introduce Steerable Visual Representations (SteerViT), a new class of visual representations that equip *any* pretrained visual encoder with the ability to steer its features with natural language. SteerViT inverts the MLLM paradigm by conditioning the visual encoder on language. By interleaving lightweight cross-attention layers into frozen ViT blocks, our method steers both global and local features with text while preserving the base ViT's representation quality, achieving a Pareto improvement over prior approaches with only ∼21M trainable parameters. Across personalized object discrimination and industrial anomaly segmentation, our method matches or surpasses dedicated methods without task-specific training, generalizing across multiple ViT backbones. These results suggest that text can serve as a lightweight, post-hoc steering mechanism that extends the capabilities of rich vision encoders to new domains without fine-tuning.

Acknowledgements

The authors gratefully acknowledge the HPC resources provided by the Erlangen National High Performance Computing Center (NHR@FAU) of the Friedrich-Alexander Universität Erlangen-Nürnberg (FAU) under the BayernKI project v115be. BayernKI funding is provided by Bavarian state authorities. The authors also gratefully acknowledge the Gauss Centre for Supercomputing e.V. (www.gauss-centre.eu) for funding this project by providing computing time on the Supercomputer JUPITER at Jülich Supercomputing Centre (JSC). MT thanks Adobe Research for a research gift.

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., et al.: Flamingo: a Visual Language Model for Few-Shot Learning. In: Advances in Neural Information Processing Systems (NeurIPS) (2022)
2. Amini, A., Banaszak, A., Benoit, H., Böök, A., et al.: Lfm2 technical report. arXiv preprint arXiv:2511.23404 (2025)
3. Bai, S., Cai, Y., Chen, R., Chen, K., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
4. Bergmann, P., Fauser, M., Sattlegger, D., Steger, C.: MVTec AD — A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
5. Buswell, G.T.: How people look at pictures: a study of the psychology and perception in art. Univ. Chicago Press, Chicago, IL (1935)
6. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., et al.: SAM 3: Segment Anything with Concepts. In: International Conference on Learning Representations (ICLR) (2026)
7. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: "The Pascal Visual Object Classes (VOC) Challenge". International Journal of Computer Vision pp. 303–338 (2010)
8. Gupta, A., Dollar, P., Girshick, R.: LVIS: A dataset for large vocabulary instance segmentation. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
9. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked Autoencoders Are Scalable Vision Learners. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
10. Hicsonmez, S., Shabayek, A.E.R., Aouada, D.: Training Free Zero-Shot Visual Anomaly Localization via Diffusion Inversion. arXiv preprint arXiv:2601.08022 (2026)
11. Howard, J.: Imagewoof: a subset of 10 classes from imagenet that aren't so easy to classify (2019), `https://github.com/fastai/imagenette#imagewoof`, accessed: 2026-06-26
12. Jeong, J., Zou, Y., Kim, T., Zhang, D., Ravichandran, A., Dabeer, O.: WinCLIP: Zero-/Few-Shot Anomaly Classification and Segmentation. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
13. Jiang, T., Song, M., Zhang, Z., Huang, H., Deng, W., Sun, F., et al.: E5-V: Universal Embeddings with Multimodal Large Language Models. arXiv preprint arXiv:2407.12580 (2024)

14. Kazemzadeh, S., Ordonez, V., Matten, M., Berg, T.: ReferItGame: Referring to objects in photographs of natural scenes. In: Empirical Methods in Natural Language Processing (EMNLP). pp. 787–798 (2014)
15. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3D Object Representations for Fine-Grained Categorization. In: International Conference on Computer Vision Workshops (ICCVW) (2013)
16. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., et al.: Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. International Journal of Computer Vision (IJCV) **123**(1), 32–73 (2017)
17. Labs, B.F.: FLUX.2: Frontier Visual Intelligence. `https://bfl.ai/blog/flux-2` (2025), accessed: 2026-06-26
18. Li, Y., Ivanova, E., Bruveris, M.: FADE: Few-shot/zero-shot Anomaly Detection Engine using Large Vision-Language Model. In: British Machine Vision Conference (BMVC) (2024)
19. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., Zhang, L.: Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. In: European Conference on Computer Vision (ECCV) (2024)
20. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., et al.: RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
21. Lüddecke, T., Ecker, A.: Image Segmentation Using Text and Image Prompts. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
22. Mahajan, S., Rao, S., Xie, J., Koller, A., Schiele, B.: Tevi: Text-conditioned editing of visual representations via sparse autoencoders for improved vision-language alignment. arXiv preprint arXiv:2606.07451 (2026)
23. McInnes, L., Healy, J., Saul, N., Großberger, L.: UMAP: Uniform Manifold Approximation and Projection. Journal of Open Source Software **3**(29), 861 (2018)
24. Neuhold, G., Ollmann, T., Bulò, S.R., Kontschieder, P.: The Mapillary Vistas Dataset for Semantic Understanding of Street Scenes. In: International Conference on Computer Vision (ICCV) (2017)
25. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., et al.: DINOv2: Learning Robust Visual Features without Supervision. Transactions on Machine Learning Research (TMLR) (2024)
26. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., et al.: Learning Transferable Visual Models From Natural Language Supervision. In: International Conference on Machine Learning (ICML) (2021)
27. Sagawa, S., Koh, P.W., Hashimoto, T.B., Liang, P.: Distributionally Robust Neural Networks. In: International Conference on Learning Representations (ICLR) (2020)
28. Sundaram, S., Chae, J., Tian, Y., Beery, S., Isola, P.: Personalized Representation from Personalized Generation. In: International Conference on Learning Representations (ICLR) (2025)
29. Thirukovalluru, R., Han, X., Dhingra, B., Dinan, E., Elbayad, M.: Text-Guided Semantic Image Encoder. arXiv preprint arXiv:2511.20770 (2025)
30. Torralba, A., Efros, A.A.: Unbiased look at dataset bias. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1521–1528 (2011)
31. Vaze, S., Carion, N., Misra, I.: Genecis: A benchmark for general conditional image similarity. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2023)

32. Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: SUN database: Large-scale scene recognition from abbey to zoo. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2010)
33. Xiao, R., Kim, S., Georgescu, M.I., Akata, Z., Alaniz, S.: FLAIR: VLM with Fine-grained Language-informed Image Representations. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2025)
34. Yu, L., Poirson, P., Yang, S., Berg, A.C., Berg, T.L.: Modeling Context in Referring Expressions. In: European Conference on Computer Vision (ECCV) (2016)
35. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid Loss for Language Image Pre-Training. In: International Conference on Computer Vision (ICCV) (2023)
36. Zhan, G., Liu, Y., Han, K., Xie, W., Zisserman, A.: ELIP: Enhanced Visual-Language Foundation Models for Image Retrieval. In: IEEE International Conference on Content-Based Multimedia Indexing (2025)
37. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene Parsing through ADE20K Dataset. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
38. Zhou, C., Loy, C.C., Dai, B.: Extract Free Dense Labels from CLIP. In: European Conference on Computer Vision (ECCV) (2022)
39. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., et al.: Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479 (2025)