Uncovering and Shaping the Latent Representation of 3D Scene Topology in Vision-Language Models

Haoming Wang **Wei Gao**
Department of Electrical and Computer Engineering
University of Pittsburgh
Pittsburgh, PA 15261
{hw.wang, weigao}@pitt.edu

Abstract

Decades of cognitive science establish that humans navigate environments by forming cognitive maps, defined as allocentric and topology-preserving representations of 3D space. While modern Vision-Language Models (VLMs) demonstrate emergent spatial reasoning from 2D egocentric inputs, it remains unclear whether they construct an analogous 3D internal representation. In this paper, we demonstrate that current VLMs do possess a latent topological map of 3D scenes, but it is heavily overshadowed by non-geometric visual semantics, such as color and shape. By isolating this spatial subspace through cross-scene linear feature extraction, we extract a clean spatial subspace that causally controls the model's spatial outputs. We mathematically shape this latent representation and prove its correspondence to the Laplacian eigenmaps of the scene's 3D Gaussian-kernel graph, aligning with physical coordinates under continuous-limit assumptions. Motivated by this geometric identification, we further introduce a mathematically principled latent regularization method for VLMs, based on Dirichlet energy. Applying this single-term regularizer to a minimal 500-step supervised VLM fine-tuning (SFT) on simple synthetic data yields significant improvements on real-world spatial benchmarks, outperforming the matched SFT baselines, including a 12.1-percentage-point gain on easy relative-direction questions.

1 Introduction

Decades of cognitive science establish that humans and other animals navigate complex 3D environments by forming an internal *cognitive map* — an allocentric representation of the spatial relations among objects that can be queried without visually perceiving the scene [1, 2]. Rather than individual pixels, image patches or metric distances, what such a cognitive map preserves is primarily the *scene topology*, defined as the qualitative graph of "what is next to what" that encodes relative positions and neighborhood structures [3, 4], as shown in Figure 1. By constructing the cognitive map, the human brain compresses complex visual inputs into a low-dimensional representation, hence supporting efficient but precise spatial understanding and reasoning.

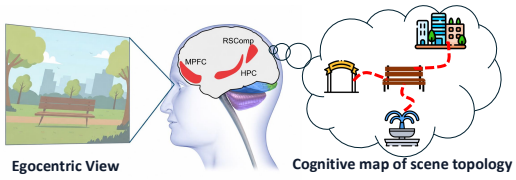


Figure 1: The human brain compresses egocentric visual inputs into a low-dimensional and topology-preserving cognitive map

Nowadays, Vision-Language Models (VLMs) [5, 6] can succeed on simple spatial reasoning tasks given only egocentric visual inputs instead of dedicated 3D geometric information (e.g., depth estimation) [7, 8, 9], but they fail on more difficult tasks in complex scenes with many objects, partial views and cluttered layouts [10, 11, 12]. The natural question, then, is whether VLMs have topology-preserving latent representations of the scene's 3D structure that are analogous to the cognitive map

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

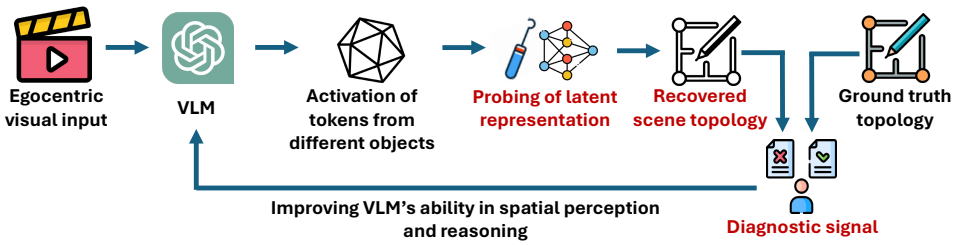


Figure 2: Uncovering and shaping the latent representation of 3D scene topology using a feature extraction probe as diagnostic signal

in human brain. It is important to know *where* such representations reside, and whether we can mathematically or experimentally *shape* them to improve the model's spatial reasoning abilities.

Existing diagnostic tools fall short of answering these questions. Attention-map analyses on VLMs reveal which 2D image patches the answer token attends to, but cannot reveal whether the model internally binds those image patches to a 3D allocentric representation [13, 14, 15]. Mechanistic interpretability for multimodal models has identified linear circuits for relational reasoning [16, 17, 18], but these analyses remain anchored to the 2D image grid or text features, and stop short of asking whether the residual stream encodes the scene's intrinsic 3D topology. End-to-end VQA accuracy, in particular, is itself an unreliable proxy, as VLMs are known to exploit caption-style co-occurrence priors, distractor-set artifacts, and language shortcuts that may result in good accuracy on certain tasks but lack generalizability [19, 20, 21, 22]. Instead, we need a direct probe of *latent* representations in model's activations, which (i) does not rely on the model's text output, (ii) is benchmark-agnostic, and (iii)establishes a precise mathematical equivalence to the physical 3D scene layout.

Our findings showed that the existing VLMs do encode information about the 3D scene topology, but this representation is severely overshadowed by spatially irrelevant visual semantics (§3). To bypass this semantic entanglement, as shown in Figure 2, this paper introduces a linear feature extraction technique that isolates visual semantics from latent representations of 3D scene topology. More specifically, we consider the same object's non-spatial attributes (e.g., color, shape, etc) as constant across different scenes, and average the object-token activations to project out the irrelevant visual semantics and extract a dominant spatial identity subspace. Experiment results validated that principal components of the extracted spatial subspace recover the ground-plane layout of the rendered 3D scenes (§4).

Furthermore, we analytically proved that this extracted spatial subspace optimally aligns with the Laplacian eigenmaps of the scene's 3D graph (Theorem 1) and converges to 3D coordinate space (Theorem 2). Consequently, we introduce a Dirichlet-energy regularizer to fine-tune the VLM's "spatially relevant" layers. This regularizer can be theoretically ensured to guarantee the task reliability, drastically reduce the sample complexity and minimize risk (Theorems 3–5). Our experiment results show that just 500 supervised fine-tuning (SFT) steps with LoRA on simple synthetic data could successfully reshape the internal geometry in VLM's latent representations, and yield significant improvements on real-world spatial benchmarks such as VSI-Bench [23] and MindCube [24], without architectural changes or introducing extra 3D modalities to the model itself (§5). For example, the gain over naive LoRA is 12.1 percentage points on easy relative-direction questions.

2 Related Work

2.1 Spatial Reasoning in Vision-Language Models

While cognitive science highlights that biological agents navigate using topology-preserving "cognitive maps" [1, 2, 4, 3], imparting such allocentric 3D understanding to VLMs has largely been approached as an explicit architectural or brute-force data scaling problem. Recent approaches attempt to force 3D awareness by appending depth modalities [24, 25] or fusing multi-vision encoders [5, 6, 26]. Other efforts scale up spatial capabilities by pretraining on large amounts (e.g., billions) of synthesized spatial QA samples [7, 9, 10, 27], or utilizing reinforcement learning (RL) with programmatic spatial rewards [28, 10, 29]. While highly effective, these approaches rely heavily on injecting metric 3D data or altering standard input pipelines. In contrast, guided by the topological nature of biological spatial representation, our work explores whether standard VLMs naturally form latent 3D representations [15, 16] solely from 2D egocentric visual inputs, avoiding the need for dedicated external 3D modules.

2.2 Mechanistic Interpretability of Geometry in VLMs

Mechanistic interpretability seeks to reverse-engineer the latent algorithms learned by neural networks [30, 31, 32]. In multimodal models, researchers have utilized attention-map analyses to explore how VLMs ground textual concepts in image patches [13, 19, 20, 14], and recent works strived to directly map between spatial and relational representations [18, 33, 32, 34]. [35] demonstrated that ViT patch embeddings encode an *is-same-object* quadratic form to bind visual features, and Kang et al. [17] identified linear "spatial-ID" vectors that VLMs use to bind spatial locations to text tokens, showing that causal interventions on these IDs can flip a model's spatial output. Despite these advances, existing analyses remain largely anchored to the 2D image grid (i.e., correspondence between 2D image patches and objects in the scene) or isolated within the vision encoder's early binding mechanisms [21]. This leaves an open question about whether the downstream language modeling residual stream constructs a cohesive 3D allocentric understanding of the physical scene.

3 VLMs' Latent Representations of 3D Scene Topology

In this section, we conduct preliminary experiments to determine *whether* pretrained VLMs already represent 3D scene topology in their latent space and *how strong* it is relative to the irrelevant semantics representations that the model also carries.

Experiment Setup: To rigorously evaluate the internal representations of VLMs without the confounding variables of real-world images (e.g., lighting, textures, background context, occlusion, etc) [36, 37, 38], we built and used a fully controllable synthetic spatial dataset, namely **SynSpat3D**, to decouple a model's purely geometric understanding from its semantic visual priors. As shown in Figure 3, each synthetic sample is generated by a two-step pipeline: (1) *Scene generation*: 6 to 12 objects vary in color, shape, size, and ground-plane position (x, y); height z equals the object radius. (2) *Frame generation*: We render the scene using randomly generated camera paths to provide egocentric 2D views. The probing corpus contains 1,000 distinct scenes, each of which is paired with 5 programmatically generated spatial questions, yielding a total of 5,000 video-question pairs. More details can be found in Appendix A.

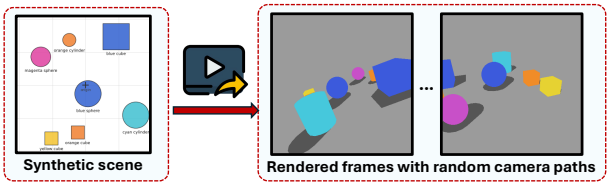


Figure 3: Synthetic spatial data

Using this dataset, we examine the residual-stream activations at every decoder layer of two pretrained VLMs, namely Qwen2.5-VL-7B [39] and InternVL3-8B [40], which are widely adopted open-sourced VLMs with strong baseline spatial reasoning capabilities.

Dominance of spatially irrelevant features. We first apply direct linear probing [31] at different layers to quantify the representational richness of spatially relevant features (3D positions) versus spatially irrelevant semantics (color, shape, etc). To do so, we train probes on the activations of image patches to predict each of these targets: object color, object shape, physical x- and y-coordinates, size-linked height z, and pairwise 3D distance using representational similarity analysis (RSA) that measures the relative distance between objects. As shown in Figure 4, the results reveal a stark representational imbalance. The visual identity of an object is recoverable from activations with near-perfect accuracy, but its ground-plane position is weakly recovered. Analyzing the activation variance captured by each feature's optimal low-rank subspace further highlights this discrepancy (details are in Appendix B).

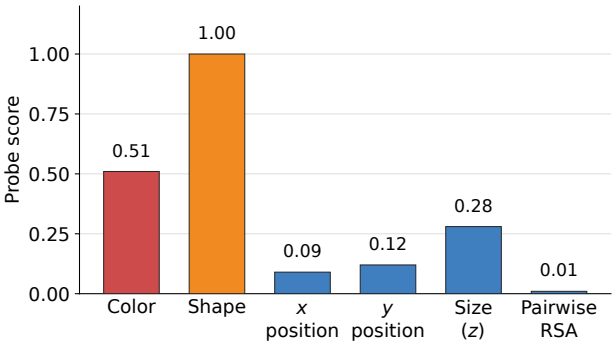


Figure 4: VLMs preferentially encode irrelevant visual features over 3D position

VLM outputs in spatial reasoning may follow semantics instead of 3D positions: To determine whether this representational imbalance dictates the VLM's downstream reasoning, we conduct a

behavioral counterfactual experiment [30]. We evaluate the model across 50 scenes under three conditions: the *original* scene, a *color-swap* variant (colors permuted, 3D positions preserved), and a *position-swap* variant (colors preserved, 3D layout scrambled). After each swap, prompts and ground-truth answers are updated to match the rendered scene (Appendix B).

Color swaps alter distance-order predictions in 41.0% of cases, compared to 48.6% for position swaps. Interpreting this difference as genuine sensitivity to visual appearance requires tracking the same physical referents and comparing answers after accounting for object renaming; otherwise, merely changing a color label can alter the ground-truth answer. Appendix B details this distinction and notes the remaining scoring criteria for verification.

4 Linear Spatial Feature Extraction via Cross-Scene Averaging

To address such representational imbalance, we aim to cleanly disentangle the latent spatial representation from the dominant non-geometric representations, and our approach is a linear feature extraction technique based on cross-scene averaging that projects out the irrelevant visual semantics. Since a VLM's latent space is fundamentally shaped by its pretraining objectives to form consistent, high-dimensional geometric structures, the extracted spatial subspace corresponds to how a VLM inherently maps 2D visual tokens into a cohesive allocentric 3D understanding, and is hence dependent only on the VLM's internal weights and architecture rather than the dataset being used in extraction.

As a result, we use the SynSpat3D dataset built in Section 3 for extraction, and validate the efficacy of the extracted spatial subspace through both qualitative geometric visualizations and quantitative formal analysis. In Section 5, we will further mathematically formalize this subspace to derive a Dirichlet-energy regularizer, demonstrating that enforcing this topology-preserving penalty in VLM fine-tuning can significantly improve the VLM's spatial reasoning abilities in real-world scenes.

4.1 Per-object Representation

Let $\mathcal{S}$ denote the set of scenes and let object identity $o \in \mathcal{O}$ index a colored geometric primitive (e.g. "red cube"). Each scene[1] $s \in \mathcal{S}$ contains m objects $\{o_i\}_{i=1}^{m}$ at 3D positions $X^{(s)} = (x_1^{(s)}, \ldots, x_m^{(s)}) \in \mathbb{R}^{3\times m}$. Unlike texts where discrete concepts map cleanly to individual tokens, visual objects in VLMs are distributed across continuous spatio-temporal patch grids, requiring a deliberate aggregation strategy to isolate their representations.

As shown in Figure 5, we feed the VLM with a sequence of rendered frames $\{I_t^{(s)}\}_{t=1}^{T}$. Each frame is patchified by the vision encoder onto a $g_h \times g_w$ patch grid, and consecutive frames are fused so that T input frames produce T/τ temporal tokens[2]. Let $h_{p,t}^{(\ell)} \in \mathbb{R}^d$ denote the activation at decoder layer ℓ of the visual token at patch indexed by $p \in \mathcal{P} := \{1, \ldots, g_h\} \times \{1, \ldots, g_w\}$ in temporal slot t. For each object o_i in scene s we form a mask–driven, coverage-weighted pool over visual tokens:

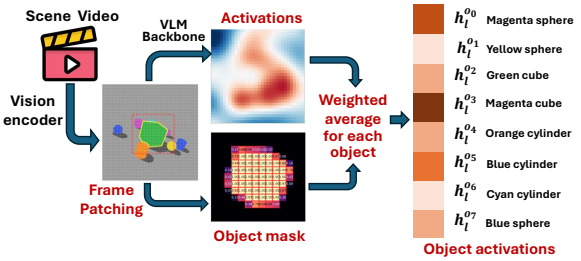


Figure 5: Computation of per-object representation. Patch activations $h_{p,t}^{(\ell)}$ are pooled with coverage weights $c_{o,p,t}^{(s)}$ into $h_\ell^{(s,o)}$ (Eqs. 1–2).

$$h_{\ell,t}^{(s,o_i)} = \frac{\sum_{p:\, c_{o_i,p,t}^{(s)} \geq \kappa} c_{o_i,p,t}^{(s)}\, h_{p,t}^{(\ell)}}{\sum_{p:\, c_{o_i,p,t}^{(s)} \geq \kappa} c_{o_i,p,t}^{(s)}}, \qquad c_{o_i,p,t}^{(s)} = \frac{\left|\, \text{patch}_p \cap \text{mask}_{o_i,t}^{(s)} \,\right|}{\left|\, \text{patch}_p \,\right|}, \tag{1}$$

where $c_{o_i,p,t}^{(s)} \in [0,1]$ is the fraction of patch p's pixels belonging to object o_i in the ground-truth mask (temporal-slot averaged), and κ is a coverage threshold that excludes patches grazed by an object's silhouette. We then average across temporal slots to obtain the *per-object activation*:

$$h_\ell^{(s,o_i)} = \frac{1}{T/\tau} \sum_{t=1}^{T/\tau} h_{\ell,t}^{(s,o_i)} \in \mathbb{R}^d, \qquad H_\ell^{(s)} = \left[h_\ell^{(s,o_1)}, \ldots, h_\ell^{(s,o_m)}\right]^\top \in \mathbb{R}^{m\times d}, \tag{2}$$

where d is the LM hidden size . $H_\ell^{(s)}$ is the extracted activation matrix for objects in different layers.

[1] Each scene contains at most one instance of a given identity.

[2] τ is used for VLM backbones with a temporal-patch merger of size τ. For example, τ=2 in Qwen2.5-VL.

4.2 Disentangling the Spatial Representation

Observation of linear decomposition: Recent works of mechanistic interpretability [41, 42, 43, 44, 45] observed that, in mainstream model architectures and both text and vision domains, high-level human-interpretable features are encoded as approximately linear directions in the latent space [17, 33, 46, 47, 48]. We approximate an object's representation additively into an identity-attribute component (carrying properties that depend only on o such as color, shape, lexical identity, etc) and a spatial component (carrying the object's 3D position in scene s):

$$h_\ell^{(s,o)} \;=\; u_{\mathrm{id}}(o) \;+\; u_{\mathrm{sp}}\big(x_o^{(s)}\big) \;+\; \varepsilon^{(s,o)}, \tag{3}$$

where $u_{\mathrm{id}}\colon \mathcal{O} \to \mathbb{R}^d$ is identity-conditional, $u_{\mathrm{sp}}\colon \mathbb{R}^3 \to \mathbb{R}^d$ is position-conditional, and $\varepsilon^{(s,o)}$ collects interactions and noise. The formulation applies to both 2D and 3D coordinates; the 2D layouts are used to visualize the extracted representations. Hence, to isolate the spatial component u_{sp} , we first compute the identity-conditional component u_{id}, by averaging object activations across randomized scenes.

Cross-scene averaging: Because the 3D layouts in our synthetic dataset are generated uniformly at random and independently of object identity (Appendix A), the cross-scene expectation of the coordinate-dependent spatial component u_{sp} converges to a fixed constant vector $\mu = \mathbb{E}_x[u_{\mathrm{sp}}(x)]$ across all scenes, with no linearity assumption required on u_{sp}. Since μ carries no information about which position an object occupies, we treat it as a global property shared by all rendered objects and absorb it into $u_{\mathrm{id}}(o)$ alongside object-specific attributes (e.g., color, shape, size, and lexical identity), allowing us to normalize $\mathbb{E}_x[u_{\mathrm{sp}}(x)] = 0$ without loss of generality [17]. Consequently, averaging an object's activations across a large corpus of scenes effectively cancels out its position-dependent variance, leaving only the invariant identity features. Thus, we have:

$$\bar{h}_\ell^{(o)} \;\triangleq\; \frac{1}{|\mathcal{S}_o|}\sum_{s\in\mathcal{S}_o} h_\ell^{(s,o)} \;=\; u_{\mathrm{id}}(o) \;+\; \underbrace{\frac{1}{|\mathcal{S}_o|}\sum_{s\in\mathcal{S}_o} u_{\mathrm{sp}}\big(x_o^{(s)}\big)}_{\to 0 \text{ as } |\mathcal{S}_o|\to\infty} + O(|\mathcal{S}_o|^{-1/2}), \tag{4}$$

Under these assumptions, the cross-scene mean activation $\bar{h}_\ell^{(o)}$ is a consistent estimator of the identity-attribute prototype $u_{\mathrm{id}}(o)$. We collect the per-identity prototypes into a matrix $\bar{H}_\ell = [\bar{h}_\ell^{(o_1)}, \ldots, \bar{h}_\ell^{(o_K)}]^\top \in \mathbb{R}^{K\times d}$ over the K unique identities in our object inventory.

Identity-attribute basis W: Having estimated the identity prototypes, we next construct an orthogonal basis for this spatial-irrelevant subspace so that it can be cleanly projected out. A thin SVD $\bar{H}_\ell^\top = U\Sigma V^\top$ gives the principal directions of the prototype cloud; we keep the top-k left singular vectors as columns of an orthonormal basis:

$$W_\ell \;\in\; \mathbb{R}^{d\times k}, \quad W_\ell^\top W_\ell = I_k, \qquad \mathrm{span}(W_\ell) \;\approx\; \mathrm{span}\{u_{\mathrm{id}}(o) : o \in \mathcal{O}\}, \tag{5}$$

and the specific choice of k is discussed in Appendix E.

Linear spatial feature extraction $\tilde{h}_\ell^{(s,o)}$: The orthogonal projector *onto* the identity-attribute subspace is $P_W = W_\ell W_\ell^\top$, and the projector *away from* it is $P_\perp \;\triangleq\; I_d \;-\; W_\ell W_\ell^\top$. We then define the *spatially-extracted* (or *spatial-feature* for short) per-object activation as

$$\tilde{h}_\ell^{(s,o)} \;\triangleq\; P_\perp\, h_\ell^{(s,o)} \;=\; h_\ell^{(s,o)} \;-\; W_\ell W_\ell^\top h_\ell^{(s,o)}, \tag{6}$$

which equals $P_\perp u_{\mathrm{sp}}(x_o^{(s)}) + P_\perp \varepsilon^{(s,o)}$ when the estimated basis spans the identity component. Spatial information overlapping that basis may also be removed, while unmodeled attributes and viewpoint effects may remain. Thus, "spatially-extracted" denotes a spatially enriched residual, not guaranteed pure geometry. The per-scene matrix is $\tilde{H}_\ell^{(s)} = H_\ell^{(s)} P_\perp$.

4.3 Results of Linear Spatial Feature Extraction

Qualitative visualization of the spatial subspace: To visualize the geometry in the spatial subspace, we collect the extracted per-object activations $\{\tilde{h}_\ell^{(s,o)}\}$ in different scenes and run PCA:

$$\tilde{\mathbf{H}}_\ell \;=\; U\Sigma V^\top, \qquad Z \;=\; \tilde{\mathbf{H}}_\ell\, V_{:,1:3} \;\in\; \mathbb{R}^{N\times 3}. \tag{7}$$

Each row of Z is the 3D PCA embedding of one object-token in one scene and $N = \sum_s m^{(s)}$. Figure 6 plots Z colored by the object's true 3D position, and demonstrates that with our linear spatial extraction, the resulting spatial subspace neatly unfolds at the middle layers of the VLM and mirrors the physical scene layout. More visualization examples are in Appendix D.

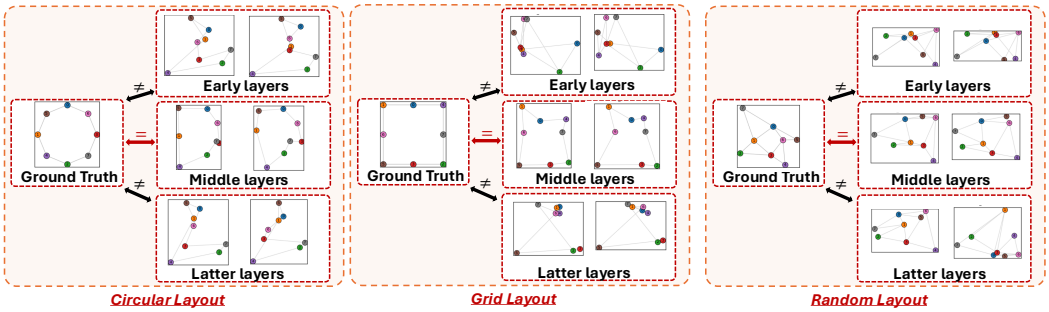
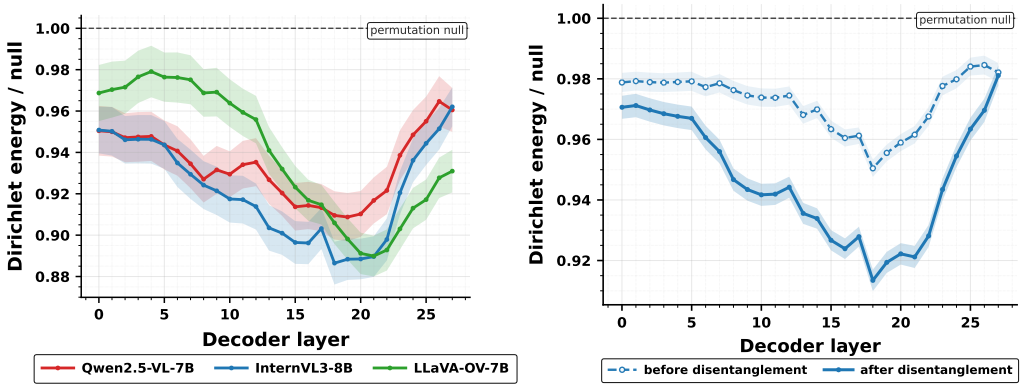


Figure 6: Visualization of extracted spatial feature in different layers of VLM



(a) Dirichlet energy in different layers

(b) Effectiveness of disentanglement

Figure 7: Dirichlet energy in different VLM layers

Quantitative metrics of topology similarity: Besides the qualitative visualization above, we also measure topology-faithfulness of $\tilde{H}_\ell^{(s)}$ to the scene's true 3D coordinates $X^{(s)}$ using Dirichlet energy [49, 50]. We define a Gaussian-kernel graph over the scene's objects as $W_{ij} = \kappa_\tau(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\tau^2}\right)$. With the degree matrix $D_{ii} = \sum_j W_{ij}$ and the graph Laplacian $L = D - W$, the Dirichlet energy of an object matrix $\tilde{H}_\ell^{(s)}$ is

$$\mathcal{E}_X(\tilde{H}_\ell^{(s)}) \;=\; \frac{1}{2}\sum\nolimits_{i,j} W_{ij}\,\|\tilde{h}_\ell^{(s,o_i)} - \tilde{h}_\ell^{(s,o_j)}\|^2 \;=\; \mathrm{tr}\big(\tilde{H}_\ell^{(s)\top} L\, \tilde{H}_\ell^{(s)}\big). \tag{8}$$

A low Dirichlet energy indicates that closer objects in the physical 3D space are mapped to proximal points in the latent space, hence better preserving the scene's physical topology. To normalize against scale and per-scene difficulty, we report the *Dirichlet ratio* against a permutation null, defined as $\rho^{\mathrm{dir}}(\tilde{H}_\ell^{(s)}) = \mathcal{E}_X(\tilde{H}_\ell^{(s)}) / \mathbb{E}_{\Pi\sim\mathcal{U}(\mathfrak{S}_m)}\big[\mathcal{E}_X(\Pi\,\tilde{H}_\ell^{(s)})\big]$, estimated with 1,000 row-shuffles of $\tilde{H}_\ell^{(s)}$ scene. $\rho^{\mathrm{dir}} < 1$ indicates the activation is more topology-faithful than a random binding of objects to rows.

Being consistent with the visualizations in Figure 6, Figure 7 shows that the Dirichlet energy is minimized at these intermediate depths. Combining these findings with prior literature on model circuits and feature binding, we deduce a distinct **functional division** across the network: *early layers* extract low-level visual features, *middle layers* construct a spatially aware 3D understanding, and *late layers* engage in linguistic reasoning to generate textual outputs. Furthermore, we validate the effectiveness of our disentanglement approach by comparing the Dirichlet energy before and after extraction. As shown in Figure 7b, the spatial localization within the middle layers becomes substantially more pronounced after the spatially irrelevant semantics are projected out.

Causal verification. To verify that the VLM actively uses the extracted spatial-aware subspace, we performed a causal intervention. As shown in Figure 8, by injecting a controlled vector along the spatial subspace's principal axis into an object's activation, the 3D-coordinate probe's prediction shifted linearly and selectively along that targeted axis. In contrast, applying a matched control vector in a null direction produced no detectable effect. This demonstrates selective control of the coordinate-probe readout; it does not by itself establish causal control of generated answers. More details can be found in Appendix C.

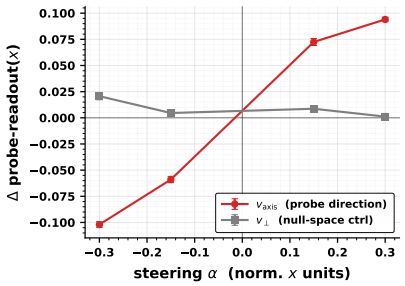


Figure 8: Causal verification on the VLM's use of the spatial subspace

5 Shaping the Spatial Subspace via Dirichlet-Energy Regularization

With the extracted spatial subspace, we further explore possibilities of deliberately shaping it to enhance the VLM's spatial reasoning capabilities. In this section, we mathematically formalize this spatial subspace to derive a quantitative penalty. We then demonstrate that enforcing this penalty in VLM fine-tuning yields measurable and theoretically guaranteed gains on spatial reasoning.

5.1 The Target Representation of Laplacian Eigenmaps

We first provide a mathematical formulation for the spatial subspace extracted in Section 4, to motivate why the *Dirichlet energy* is the ideal scalar to minimize in VLM fine-tuning. Eq. (8) described the Dirichlet energy of an object-token matrix $H \in \mathbb{R}^{m\times d}$ on the scene's 3D Gaussian-kernel graph W is $\mathcal{E}_X(H) = \mathrm{tr}(H^\top L H)$. Geometrically, a low Dirichlet energy ensures that rows of H corresponding to spatially nearby objects (large W_{ij}) are themselves nearby in $\mathbb{R}^d$. To formalize this, we seek the exact mathematical structure of the representation H^* that minimizes this energy.

Theorem 1 (Dirichlet minimization $\Rightarrow$ Laplacian eigenmap PCA)**.** *Let $X \in \mathbb{R}^{m\times 3}$ be the 3D coordinates of m scene objects, $W_{ij} = \kappa_\tau(x_i, x_j)$ their Gaussian-kernel graph, and $L = D - W$ the unnormalized graph Laplacian. Fix $s \le \min(m, d) - 1$ and consider the constrained optimization:*

$$H^* \in \arg\min_{H\in\mathbb{R}^{m\times d}} \mathcal{E}_X(H) \quad \textit{subject to} \quad \sigma_k(H) \ge \epsilon_k \quad \textit{for all } k \in [s], \tag{9}$$

where $\sigma_k(H)$ is the k-th singular value of H and $\epsilon_1 > \epsilon_2 > \cdots > \epsilon_s > 0$ are constants. Then:

(a) The minimizer is unique up to right-orthogonal transformation. The left singular vectors satisfy $u_k(H^*) = z^{(k)}$ for $k = 1, \ldots, s$, where $z^{(k)}$ is the k-th eigenvector of L ordered by smallest eigenvalues.

(b) After mean-centering $H_c^* \triangleq (I - \frac{1}{m}\mathbf{1}\mathbf{1}^\top)H^*$, the k-th principal component (left singular vector) of H_c^* exactly recovers $z^{(k+1)}$ for $k = 1, \ldots, s-1$.

This theorem proves that if we force a representation to maintain some variance (the ϵ_k constraints prevent trivial collapse to a single point) while minimizing the Dirichlet energy, the optimal representation H^* will naturally align its principal components with the eigenvectors of the scene's graph Laplacian. In practical VLM fine-tuning, the standard language loss ($\mathcal{L}_{\mathrm{LM}}$) implicitly provides this non-degeneracy constraint, as the network cannot collapse its hidden states to zero without destroying its ability to predict text tokens.

Proof Sketch. The full formal proof is in Appendix F and relies on the Ky Fan trace minimum theorem. (1) By expanding H into its Singular Value Decomposition ($H = U\Sigma V^\top$), the Dirichlet objective exactly decomposes along the orthonormal left-singular directions u_k. (2) The lower-bound constraints $\sigma_k \ge \epsilon_k$ freeze the singular values. (3) Applying the weighted Ky Fan theorem proves that the optimal orthonormal vectors u_k must perfectly align with the eigenvectors of L corresponding to its smallest eigenvalues. (4) Mean-centering simply projects out the trivial uniform mode $z^{(1)} \propto \mathbf{1}$.

This identifies the representation as a set of abstract graph eigenvectors. To connect back to physical vision, we consider the continuous limit:

Theorem 2 (Continuous Limit to 3D Space)**.** *Let $\{x_i\}_{i=1}^m$ be i.i.d. samples from a smooth density ρ on a compact Riemannian manifold $\Omega \subset \mathbb{R}^3$. Let $L^{(\tau)}$ be the Gaussian-kernel Laplacian with*

Overall accuracy (%)	Qwen2.5-VL-7B		InternVL3-8B	
	VSI-bench	MindCube	VSI-bench	MindCube
Naive LoRA fine-tuning	29.3	41.3	36.7	45.2
2D regularization [17]	28.7	39.8	36.5	44.7
Text-domain cog map [24]	28.8	38.5	36.0	43.1
Ours	**33.6**	**48.3**	**42.4**	**51.7**

Table 1: Overall accuracy on spatial tasks

bandwidth τ_m. If $\tau_m \to 0$ and $m\tau_m^{3+\alpha} \to \infty$, then for any fixed k, with probability tending to 1, the discrete eigenvector $z^{(k+1)}(x_i)$ converges uniformly to $\rho(x_i)^{-1/2}\phi^{(k)}(x_i)$, where $\phi^{(k)}$ is the k-th non-constant eigenfunction of the weighted Laplace-Beltrami operator $-\Delta_\rho$. For uniform ρ on a Euclidean cube, $\phi^{(1)}, \phi^{(2)}, \phi^{(3)}$ are exactly the Cartesian coordinates x, y, z (up to a smooth cosine reparameterization).

5.2 Loss Formulation and Realizability

Motivated by that Dirichlet energy favors representations smooth on the scene graph, we introduce a training intervention that shapes the latent space toward this target in VLM fine-tuning. We add a single scalar Dirichlet-energy term to the standard LoRA cross-entropy loss at the cognitive-map layer ℓ:

$$\mathcal{L}(\theta) = \mathcal{L}_{\mathrm{LM}}(\theta) \; + \; \lambda_{\mathrm{dir}} \cdot \mathcal{R}_X\big(H_\ell(\theta) P_\perp\big), \tag{10}$$

where $P_\perp$ is the cross-scene spatial-feature-extraction projector from Section 4 and W is the scene's 3D Gaussian-kernel graph. Training uses the normalized energy $\mathcal{R}_X$ in Appendix I.2, averaged over the selected layers.

Why does forcing the model into this specific representation improve downstream spatial reasoning? We prove that it guarantees the *expressivity* required by the LM head:

Theorem 3 (Realizability of Spatial Readouts)**.** *Let $f : \mathbb{R}^3 \to \mathbb{R}$ be any linear function of the monotonic coordinate transformations (e.g., an axis-aligned ordinal comparison encoding "is object A to the left of object B"). For the Dirichlet-minimizing representation H^* satisfying Theorem 2, and for every $\eta > 0$, there exists a linear functional $\ell : \mathbb{R}^d \to \mathbb{R}$ such that:*

$$\sup_{i \in [m]} |\ell(h_i^*) - f(x_i)| \; < \; \eta \quad \textit{with high probability as } m \to \infty. \tag{11}$$

Since the latent space is reshaped to mirror the 3D world coordinates, spatial relational questions become linearly separable. The VLM's linear language modeling head possesses the exact architectural capacity required to answer perfectly without needing deeper, non-linear geometric parsers.

5.3 Experiment Setup

To experimentally validate the efficacy of our proposed VLM training intervention, we utilize our proposed loss to fine-tune pretrained VLMs, namely Qwen2.5-VL-7B [39] and InternVL3-8B [40], and evaluate zero-shot transfer to the following real-world spatial reasoning benchmarks:

- **VSI-Bench [23]**: a comprehensive benchmark for models' capabilities in fine-grained 3D spatial reasoning, relational judgments, and absolute distance estimation from egocentric video sequences. The evaluated tasks include: object counting (*obj-count*), object and room size estimation (*obj-size*, *room-size*), absolute and relative distance estimation (*abs-dist*, *rel-dist*), relative direction across varying difficulties (*rel-d*), appearance order (*appear-ord*), and route planning (*route*).
- **MindCube [24]**: a benchmark designed to measure how well models can construct internal cognitive maps and perform spatial mental modeling given only restricted and partial 2D views.

We train LoRA (rank-16) on the attention matrices for 500 steps. Trainable parameters are restricted strictly to the middle "spatial-aware" layers where the Dirichlet loss is applied, leaving the early visual binding and late linguistic reasoning layers untouched. We empirically identify these middle layers by selecting those with the lowest pre-intervention Dirichlet energy ratio, as established in our diagnostic probing (Figure 7). We sweep λ_{dir} to observe its effect, reporting the optimal settings in our main results and exploring the trade-off in Section 5.5. We compare our approach with the following baselines:

- **Naive LoRA fine-tuning**, which relies exclusively on the standard language modeling loss without any topological regularization.

- **2D spatial regularization** [17], which supervises the intermediate layers' spatial identifiers based on 2D pixel-space distances rather than true 3D geometry.
- **Text-domain cognitive map prediction** [24], which explicitly fine-tunes the model to generate the scene's spatial coordinates directly in the textual output space.

We compare training objectives under a matched SFT budget. Our results do not establish superiority over RL methods such as SpatialThinker [28], which differ in supervision and computational resources, or over systems using additional modalities. Our latent regularization during SFT is orthogonal to spatial RL.

Task accuracy (%)	over-all	obj-count	obj-size	room-size	abs-dist	rel-d-easy	rel-d-medium	rel-d-hard	rel-dist	appear-ord	route
Naive LoRA fine-tuning	29.3	7.4	26.3	29.3	19.5	47.9	37.1	24.5	30.4	37.3	33.6
2D regularization [17]	28.7 ↓0.6 pp	7.5 ↑0.1 pp	26.0 ↓0.3 pp	25.1 ↓4.2 pp	19.1 ↓0.4 pp	46.4 ↓1.5 pp	37.0 ↓0.1 pp	24.0 ↓0.5 pp	31.2 ↑0.8 pp	37.6 ↑0.3 pp	32.9 ↓0.7 pp
Text-domain cog map [24]	28.8 ↓0.5 pp	6.7 ↓0.7 pp	26.6 ↑0.3 pp	27.4 ↓1.9 pp	18.7 ↓0.8 pp	47.4 ↓0.5 pp	37.3 ↑0.2 pp	23.4 ↓1.1 pp	30.8 ↑0.4 pp	38.1 ↑0.8 pp	31.2 ↓2.4 pp
Ours	33.6 ↑4.3 pp	7.2 ↓0.2 pp	31.3 ↑5.0 pp	33.4 ↑4.1 pp	28.9 ↑9.4 pp	60.0 ↑12.1 pp	45.5 ↑8.4 pp	37.2 ↑12.7 pp	35.8 ↑5.4 pp	41.5 ↑4.2 pp	37.7 ↑4.1 pp

Table 2: Per-task accuracy on VSI-Bench using Qwen2.5-VL-7B. Arrows indicate percentage-point differences computed from the displayed scores relative to the case of LM loss only.

5.4 Main Results and Sample Complexity

Task accuracy (%)	over-all	obj-count	obj-size	room-size	abs-dist	rel-d-easy	rel-d-medium	rel-d-hard	rel-dist	appear-ord	route
Naive LoRA fine-tuning	36.7	10.9	28.4	43.1	80.8	49.8	33.9	26.1	26.5	34.0	33.6
2D regularization	36.5 ↓0.2 pp	10.9	28.8 ↑0.4 pp	43.5 ↑0.4 pp	81.3 ↑0.5 pp	47.4 ↓2.4 pp	33.3 ↓0.6 pp	26.0 ↓0.1 pp	26.8 ↑0.3 pp	34.0	33.5 ↓0.1 pp
text-domain cog-map pred	36.0 ↓0.7 pp	9.2 ↓1.7 pp	28.5 ↑0.1 pp	40.6 ↓2.5 pp	78.9 ↓1.9 pp	49.8	33.6 ↓0.3 pp	25.7 ↓0.4 pp	26.6 ↑0.1 pp	34.3 ↑0.3 pp	32.3 ↓1.3 pp
Ours	42.4 ↑5.7 pp	11.8 ↑0.9 pp	28.3 ↓0.1 pp	49.3 ↑6.2 pp	89.1 ↑8.3 pp	60.7 ↑10.9 pp	43.4 ↑9.5 pp	29.4 ↑3.3 pp	33.4 ↑6.9 pp	38.4 ↑4.4 pp	36.8 ↑3.2 pp

Table 3: Per-task accuracy on VSI-Bench using InternVL3-8B. Arrows indicate percentage-point differences computed from the displayed scores relative to the case of LM loss only.

As shown in Tables 1, 2, and 3, our proposed Dirichlet-energy regularization consistently outperforms all the baselines. The 2D regularization method fails to yield improvements because it constrains the model to rely on 2D pixel-space distances rather than true 3D geometry. Similarly, the text-domain cognitive map prediction falls short because it forces the language model to over-index on text semantics rather than faithfully grounding the physical visual features. Analyzing the per-task accuracy reveals that our method significantly excels in categories requiring a robust understanding of scene topology, such as relative direction (*rel-d-easy*, *rel-d-medium*, *rel-d-hard*) and absolute distance (*abs-dist*). Conversely, for tasks such as *route* and *appear-ord*, the gains of our method become smaller, as these tasks demand reasoning over temporal dynamics and sequential planning, which fall outside the scope of our purely spatial topological regularization.

Crucially, fine-tuning requires just 500 steps on simple synthetic data to generalize to complex, real-world benchmarks. We theoretically ground this extreme data efficiency:

Theorem 4 (Sample-Complexity Reduction)**.** *Let a spatial task predict $y = w^{\star\top} h + \xi$, where $w^\star$ depends only on the world-coordinate axes (spanning the top 3 PCs of H^*). For the empirical risk minimizer $\hat{w}_N$ over N samples, with probability $\geq 1-\delta$, the excess risk is bounded by:*

$$\mathbb{E}\big[(\hat{w}_N^\top h - w^{\star\top} h)^2\big] \;\leq\; \frac{3C\sigma_\xi^2}{N} \log(1/\delta), \tag{12}$$

improving upon the standard $O(d/N)$ bound of an unstructured baseline.

Standard VLM residual streams are high-dimensional ($d \approx 4000$), typically requiring massive datasets to prevent overfitting. Theorem 4 proves that the Dirichlet penalty collapses task-irrelevant variance, shrinking the effective dimension to exactly 3. This reduces sample complexity by a factor of $d/3$, enabling the LM head to learn highly generalizable spatial readouts from minimal data.

5.5 Ablation Studies

To evaluate the sensitivity of our regularization method, we vary the Dirichlet penalty weight λ during the 500-step fine-tuning. As shown in Table 4, zero-shot performance on VSI-Bench peaks at $\lambda = 3$ for Qwen and $\lambda = 1$ for InternVL. However, excessive regularization ($\lambda = 9$) reduces accuracy below the unregularized baseline.

λ	Qwen2.5-VL-7B	InternVL3-8B
0	29.3	36.7
0.3	29.5	39.1
1	32.3	**42.4**
3	**33.6**	37.2
9	28.6	35.8

Table 4: Ablation study on the Dirichlet regularization weight (λ) on VSI-Bench.

The following risk decomposition characterizes the local trade-off between the benefits of regularization and higher-order effects:

Theorem 5 (Training-Time Risk Decomposition)**.** *Let $\mathcal{L}_\lambda(\theta) = \mathcal{L}_{\mathrm{LM}}(\theta) + \lambda \mathcal{R}_X(H_\theta)$, where $\mathcal{R}_X$ is the Dirichlet ratio. Let θ^* be the population minimizer of the standard LM loss, and θ^*_λ the minimizer of the regularized loss. For a constant $\beta > 0$ correlating the spatial risk surface with the Dirichlet energy, the population spatial-task risk R_{spatial} satisfies:*

$$R_{\mathrm{spatial}}(\theta^*_\lambda) \;\le\; R_{\mathrm{spatial}}(\theta^*) - \lambda \cdot \beta\big(\mathcal{R}_X(H_{\theta^*}) - \mathcal{R}^*_X\big) + O(\lambda^2). \tag{13}$$

Under the positive gradient-alignment assumption ($\beta > 0$), the first-order term predicts a local decrease in spatial risk; realizability alone does not guarantee this alignment for every VLM or task. Higher-order effects, represented by $O(\lambda^2)$, can offset this first-order benefit as regularization increases.

Component ablations. We evaluate the effects of linear spatial feature extraction and fine-tuning layer selection on InternVL3-8B using VSI-Bench (Table 5). Table 5a compares naive LoRA, which uses only the language modeling loss, with Dirichlet regularization applied to either the original object activations H_ℓ or the spatially extracted activations $H_\ell P_\perp$ defined in Section 4.2. Spatial feature extraction improves accuracy from 37.9% to 42.4%, compared with 36.7% for naive LoRA. Table 5b fixes the number of fine-tuned layers at five and compares the first five layers, five middle layers centered at L18, and the last five layers. Fine-tuning the middle layers achieves the highest accuracy. The separate layer-count sweep is reported in Appendix J.

Setting	Accuracy (%)
Naive LoRA	36.7
Without feature extraction	37.9
With feature extraction	**42.4**

(a) Spatial feature extraction

Five-layer selection	Accuracy (%)
Early: first five	36.1
Middle: centered at L18	**42.4**
Late: last five	32.9

(b) Fine-tuning layer selection

Table 5: Component ablations on InternVL3-8B evaluated on VSI-Bench.

General multimodal capability. Preliminary MMMU [51] evaluation on InternVL3-8B shows a small decrease from 61.2 before fine-tuning to 60.5 after latent shaping. This decrease is comparable to those observed for the SFT baselines, whose scores range from 59.9 to 60.7 (Appendix J.2, Table 7).

6 Conclusion

In this paper, we isolate the latent 3D spatial subspace within VLMs and introduce a mathematically grounded Dirichlet-energy regularizer to directly shape it. By demonstrating that a minimal, topology-preserving fine-tuning stage significantly enhances zero-shot spatial reasoning performance, we provide a principled methodology for developing VLMs with robust, intrinsic cognitive maps.

Acknowledgments

We thank the reviewers and the area chair for their insightful comments and feedback. This work was supported in part by National Science Foundation (NSF) under grant number IIS-2205360 and CCF-2215042, and National Institutes of Health (NIH) under grant number R01HL170368, R01HL184168 and R01NR022545.

References

[1] Edward C Tolman. Cognitive maps in rats and men. *Psychological review*, 55(4):189, 1948.

[2] John O'keefe and Lynn Nadel. *The hippocampus as a cognitive map*. Oxford university press, 1978.

[3] William H Warren. Non-euclidean navigation. *Journal of Experimental Biology*, 222(Suppl_1):jeb187971, 2019.

[4] Michael Peer, Iva K Brunec, Nora S Newcombe, and Russell A Epstein. Structuring knowledge with cognitive maps and cognitive graphs. *Trends in cognitive sciences*, 25(1):37–54, 2021.

[5] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

[6] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12585–12602, 2024.

[7] Navid Rajabi and Jana Kosecka. Gsr-bench: A benchmark for grounded spatial reasoning evaluation via multimodal llms. *arXiv preprint arXiv:2406.13246*, 2024.

[8] Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. Sqa3d: Situated question answering in 3d scenes. *arXiv preprint arXiv:2210.07474*, 2022.

[9] Yuan-Hong Liao, Rafid Mahmood, Sanja Fidler, and David Acuna. Reasoning paths with reference objects elicit quantitative spatial reasoning in large vision-language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17028–17047, 2024.

[10] Weichen Liu, Qiyao Xue, Haoming Wang, Xiangyu Yin, Boyuan Yang, and Wei Gao. Spatial reasoning in multimodal large language models: A survey of tasks, benchmarks and methods. *arXiv preprint arXiv:2511.15722*, 2025.

[11] Haoming Wang, Qiyao Xue, and Wei Gao. Infinibench: Infinite benchmarking for visual spatial reasoning with customizable scene complexity. *arXiv preprint arXiv:2511.18200*, 2025.

[12] Qiyao Xue, Haoming Wang, Weichen Liu, Shiqi Wang, Yuyang Wu, and Wei Gao. Reasoning path and latent state analysis for multi-view visual spatial reasoning: A cognitive science perspective. In *European Conference on Computer Vision*, pages 512–531. Springer, 2026.

[13] Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 397–406, 2021.

[14] Shiqi Chen, Tongyao Zhu, Ruochen Zhou, Jinghan Zhang, Siyang Gao, Juan Carlos Niebles, Mor Geva, Junxian He, Jiajun Wu, and Manling Li. Why is spatial reasoning hard for vlms? an attention mechanism perspective on focus areas. *arXiv preprint arXiv:2503.01773*, 2025.

[15] Jianing Qi, Jiawei Liu, Hao Tang, and Zhigang Zhu. Beyond semantics: Rediscovering spatial awareness in vision-language models. *arXiv preprint arXiv:2503.17349*, 2025.

[16] Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. The geometry of categorical and hierarchical concepts in large language models. *arXiv preprint arXiv:2406.01506*, 2024.

[17] Raphi Kang, Hongqiao Chen, Georgia Gkioxari, and Pietro Perona. Linear mechanisms for spatiotemporal reasoning in vision language models. *arXiv preprint arXiv:2601.12626*, 2026.

[18] Rim Assouel, Declan Campbell, Yoshua Bengio, and Taylor Webb. Visual symbolic mechanisms: Emergent symbol processing in vision language models. *arXiv preprint arXiv:2506.15871*, 2025.

[19] Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. Analyzing the behavior of visual question answering models. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1955–1960, 2016.

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

[21] Cheng Shi, Yizhou Yu, and Sibei Yang. Vision transformers need more than registers. *arXiv preprint arXiv:2602.22394*, 2026.

[22] Haoming Wang, Boyuan Yang, Xiangyu Yin, and Wei Gao. Never start from scratch: Expediting on-device llm personalization via explainable model selection. In *Proceedings of the 23rd Annual International Conference on Mobile Systems, Applications and Services*, pages 154–168, 2025.

[23] Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10632–10643, 2025.

[24] Qineng Wang, Baiqiao Yin, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, et al. Mindcube: Spatial mental modeling from limited views. *arXiv e-prints*, pages arXiv–2506, 2025.

[25] Haoming Wang, Qiyao Xue, Weichen Liu, and Wei Gao. Mosaicthinker: On-device visual spatial reasoning for embodied ai via iterative construction of space representation. *arXiv preprint arXiv:2602.07082*, 2026.

[26] An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision-language models. *Advances in Neural Information Processing Systems*, 37:135062–135093, 2024.

[27] Zhongang Cai, Ruisi Wang, Chenyang Gu, Fanyi Pu, Junxiang Xu, Yubo Wang, Wanqi Yin, Zhitao Yang, Chen Wei, Qingping Sun, et al. Scaling spatial intelligence with multimodal foundation models. *arXiv preprint arXiv:2511.13719*, 2025.

[28] Hunar Batra, Haoqin Tu, Hardy Chen, Yuanze Lin, Cihang Xie, and Ronald Clark. Spatialthinker: Reinforcing 3d reasoning in multimodal llms via spatial rewards. *arXiv preprint arXiv:2511.07403*, 2025.

[29] Junfei Wu, Jian Guan, Kaituo Feng, Qiang Liu, Shu Wu, Liang Wang, Wei Wu, and Tieniu Tan. Reinforcing spatial reasoning in vision-language models with interwoven thinking and visual drawing. *arXiv preprint arXiv:2506.09965*, 2025.

[30] Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. *Advances in neural information processing systems*, 34:9574–9586, 2021.

[31] Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. *Advances in neural information processing systems*, 33:12388–12401, 2020.

[32] Neel Nanda, Andrew Lee, and Martin Wattenberg. Emergent linear representations in world models of self-supervised sequence models. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 16–30, 2023.

[33] Xinting Huang and Michael Hahn. Decomposing representation space into interpretable subspaces with unsupervised learning. *arXiv preprint arXiv:2508.01916*, 2025.

[34] Haoming Wang, Qiyao Xue, Weichen Liu, Boyuan Yang, Xiangyu Yin, and Wei Gao. Spatialmind: Spatially aware on-device embodied ai via viewpoint integration. In *Proceedings of the 32nd Annual International Conference on Mobile Computing and Networking*, pages 983–1001, 2026.

[35] Yihao Li, Saeed Salehi, Lyle Ungar, and Konrad P Kording. Does object binding naturally emerge in large pretrained vision transformers? *arXiv preprint arXiv:2510.24709*, 2025.

[36] Peiran Xu, Sudong Wang, Yao Zhu, Jianing Li, Gege Qi, and Yunjian Zhang. Spatialbench: Benchmarking multimodal large language models for spatial cognition. *arXiv preprint arXiv:2511.21471*, 2025.

[37] Jiahao Wang, Yufeng Yuan, Rujie Zheng, Youtian Lin, Jian Gao, Lin-Zhuo Chen, Yajie Bao, Yi Zhang, Chang Zeng, Yanxi Zhou, et al. Spatialvid: A large-scale video dataset with spatial annotations. *arXiv preprint arXiv:2509.09676*, 2025.

[38] Qiyao Xue, Xiangyu Yin, Boyuan Yang, and Wei Gao. Phyt2v: Llm-guided iterative self-refinement for physics-grounded text-to-video generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18826–18836. IEEE, 2025.

[39] Qwen Team. Qwen2.5-vl, January 2025.

[40] 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.

[41] Curt Tigges, Oskar John Hollinsworth, Atticus Geiger, and Neel Nanda. Linear representations of sentiment in large language models. *arXiv preprint arXiv:2310.15154*, 2023.

[42] Matthew Trager, Pramuditha Perera, Luca Zancato, Alessandro Achille, Parminder Bhatia, and Stefano Soatto. Linear spaces of meanings: compositional structures in vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15395–15404, 2023.

[43] Haoming Wang and Wei Gao. Deciphering personalization: Towards fine-grained explainability in natural language for personalized image generation models. *arXiv preprint arXiv:2511.01932*, 2025.

[44] Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*, 2023.

[45] Yibo Jiang, Goutham Rajendran, Pradeep Ravikumar, Bryon Aragam, and Victor Veitch. On the origins of linear representations in large language models. *arXiv preprint arXiv:2403.03867*, 2024.

[46] Achyuta Rajaram, Sarah Schwettmann, Jacob Andreas, and Arthur Conmy. Line of sight: On linear representations in vllms. *arXiv preprint arXiv:2506.04706*, 2025.

[47] Usha Bhalla, Alex Oesterling, Suraj Srinivas, Flavio P Calmon, and Himabindu Lakkaraju. Interpreting clip with sparse linear concept embeddings (splice). *Advances in Neural Information Processing Systems*, 37:84298–84328, 2024.

[48] Matthieu Tehenan, Christian Bolivar Moya, Tenghai Long, and Guang Lin. Linear spatial world models emerge in large language models. *arXiv preprint arXiv:2506.02996*, 2025.

[49] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

[50] Core Francisco Park, Andrew Lee, Ekdeep Singh Lubana, Yongyi Yang, Maya Okawa, Kento Nishi, Martin Wattenberg, and Hidenori Tanaka. Iclr: In-context learning of representations. *arXiv preprint arXiv:2501.00070*, 2024.

[51] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[52] Aditya Ravuri and Neil D Lawrence. Transformers as unrolled inference in probabilistic laplacian eigenmaps: An interpretation and potential improvements. *arXiv preprint arXiv:2507.21040*, 2025.

[53] Alex Lewandowski, Michał Bortkiewicz, Saurabh Kumar, András György, Dale Schuurmans, Mateusz Ostaszewski, and Marlos C Machado. Learning continually by spectral regularization. *arXiv preprint arXiv:2406.06811*, 2024.

[54] Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7(85):2399–2434, 2006.

[55] William L Hamilton, Rex Ying, and Jure Leskovec. Representation learning on graphs: Methods and applications. *arXiv preprint arXiv:1709.05584*, 2017.

[56] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[57] Dengyong Zhou, Olivier Bousquet, Thomas Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. *Advances in neural information processing systems*, 16, 2003.

A Data generation details

This appendix expands on the two-step synthetic-scene pipeline summarized in §3 and used both for probing (§3–§4) and for Dirichlet-regularized fine-tuning (§5).

Step 1: 3D scene generation. Each scene places between 6 and 12 rigid objects on a flat floor inside a bounded working volume of $[-4, 4] \times [-4, 4]$ in world units, with z pinned to the object's radius so all objects rest on the ground plane. The independent latent factors per object are: *shape* drawn uniformly from {*cube*, *sphere*, *cylinder*}, *color* drawn uniformly from an 8-way palette {*red*, *green*, *blue*, *yellow*, *cyan*, *magenta*, *orange*, *purple*}, *size* drawn from $\{0.4, 0.6, 0.8\}$ world units, and *position* sampled by rejection until pairwise edge-to-edge separation exceeds 0.2 world units (no overlapping objects, no occluded clusters). Color and shape are sampled independently of position, so *at the population level the visual identity of an object carries no information about where it is*; this refers to the sampled ground-plane coordinates, not the size-linked height z. Rejection constraints may still induce dependencies, so the averaging argument is an approximation. Each scene receives a unique hash-based ID (`s_<hash>`) and is written to disk as `scene.json` containing the full per-object table {*name*, *shape*, *color*, *size*, *world_position*}.

Step 2: egocentric frame rendering. For each canonical scene we render $T=16$ frames at 448×448 resolution under one of three camera regimes. (i) *Orbit* (1-DoF planar arc): the eye traces an 180° arc at fixed altitude (sampled from $\{3.5, 4.5\}$) and radius (sampled from $\{8.0, 9.0\}$), with the look-at point fixed at $z=0.5$ in the scene center. (ii) *Free6DoF* (the default for all probing/training runs reported in the body): both the eye and the look-at point drift in $\mathbb{R}^3$, the camera rolls, and the orbit radius/altitude jitter independently of arc angle (`eye_jitter`=1.2, `radius_jitter`=1.0, `altitude_jitter`=0.8, `target_jitter`=1.5; smooth-noise basis with 3 frequency modes). (iii) *Person-walk* (used as a generalization control in §3 and Appendix B): instead of orbiting around the working volume, the camera is placed *inside* it at human eye-height ($z = 1.5$) and walks forward along a smooth-noise path with stochastic per-frame yaw and pitch. The dynamics are: forward step length sampled around `speed_mean` $= 0.5$ world units with `speed_amp` $= 0.25$ smooth-noise modulation; per-frame yaw drift up to $\pm25^\circ$; per-frame pitch up to $\pm12^\circ$; collision avoidance with `object_margin` $= 0.6$ around every object and `bounds_margin` $= 0.3$ from the working-volume edges; up to 40 retries on the whole trajectory if a step cannot be repaired without violating the margins. This regime produces partial-view egocentric trajectories where individual frames frequently see zero or one object, so the corresponding spatial reasoning task requires *cross-frame* integration (the body's "cognitive-map" setting). Across all three regimes, each frame is post-validated to ensure (orbit / free6dof) at least one object remains visible or (person-walk) the trajectory satisfies its margin constraints; failed frames trigger trajectory repair. The horizontal field of view is 60° and the background is uniform mid-gray (RGB 155). To dissociate the scene layout from any single trajectory, every scene is re-rendered under 4 independent trajectories (indexed `traj_idx` $\in \{0, 1, 2, 3\}$); we use `traj_idx`=1–2 for probing and Dirichlet training.

Spatial QA generation. Per scene, 5 questions are programmatically synthesized from the ground-truth {*name*, *world_position*} table. The probing questions include: (a) *Relative-position binary* — "Is the [COLOR] [SHAPE] [IN FRONT OF/BEHIND/TO THE LEFT OF/TO THE RIGHT OF] the [COLOR] [SHAPE]?" answered against the camera-aligned xy-plane of the first frame; (b) *Distance-order ternary* — "Which is closer to the [REFERENCE], the [A] or the [B]?" Distractor objects are guaranteed present in the rendered frames but not always in the question. Each question stores the in-question object names alongside their 3D coordinates so the same JSONL can drive both standard LM cross-entropy and the Dirichlet-energy regularizer (which reads the coordinates directly). The SFT subset instead contains 2,000 relative-position and 1,000 appearance-order questions, as described below.

Probing dataset (Free6DoF $1{,}000$ scenes). The headline probing experiments in §3 draw from a $1{,}000$-scene Free6DoF corpus, each with 5 questions $\rightarrow$ $5{,}000$ video-QA pairs, plus extended controls (*Circle8* and *Grid3×3* layouts, $\sim5{,}000$ scenes total) used for the topology-vs-shortcut robustness checks reported in Appendix B. The 400-scene subsets used for layer-wise Dirichlet-energy curves (Fig. 7) and the residualized variant (Fig. 7b) are the first 400 scenes by scene-ID hash so the same split is reused across all backbones.

Counterfactual perturbation set (§3). For the color/position swap experiment we draw 50 base scenes (a random subset of the probing corpus) and produce three variants per scene by re-rendering with the same trajectory but: (i) *original*; (ii) *color-swap* — the k assigned colors of the k objects are permuted by a random derangement, leaving every 3D position untouched; (iii) *position-swap* — the k positions are permuted, leaving every color untouched. We regenerate prompts and ground truths for the same question templates and object tuples after each perturbation, using $\sim$9 questions per scene, yielding 1,290 QA pairs per seed used in the flip-rate analysis.

More examples of generated data See Figure 9, 10, 11, 12

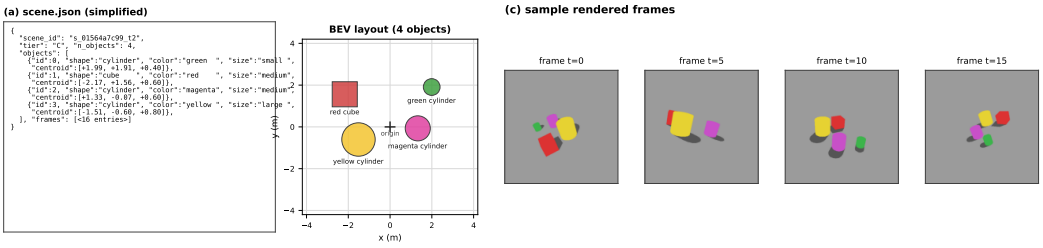


Figure 9: Generated data, example 1

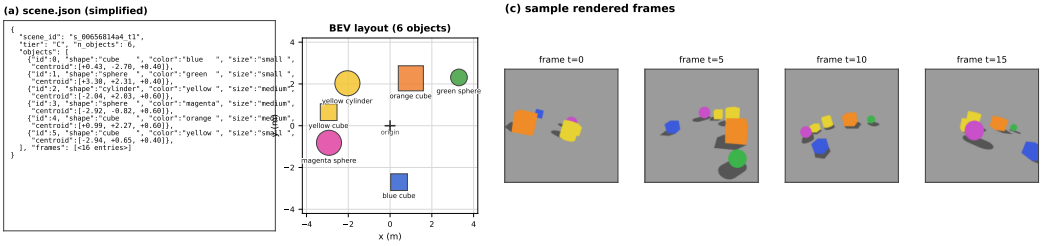


Figure 10: Generated data, example 2

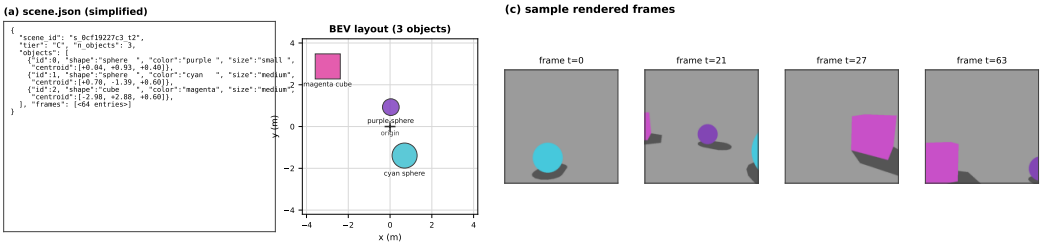


Figure 11: Generated data, example 3

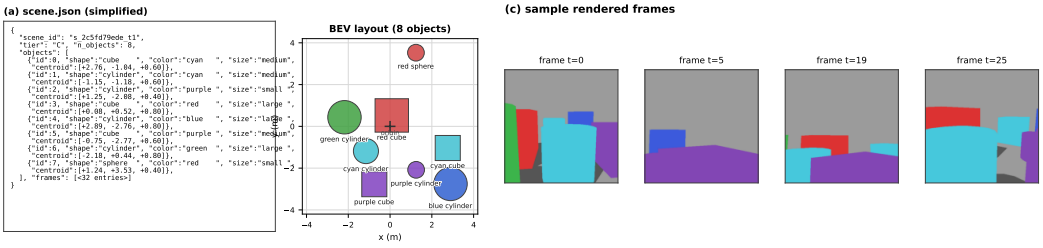


Figure 12: Generated data, example 4

B Diagnostic experiments

This appendix expands on the diagnostic experiments referenced from §3

Linear probing setup and full results. For each (model, layer) pair we extract residual-stream activations $h_\ell^{(s,o)} \in \mathbb{R}^d$ at every object-token position over the 1,000-scene Free6DoF probing corpus (Appendix A), pooled across all $T = 16$ frames per scene with the same mask-driven, coverage-weighted scheme used in §4. We then fit five independent probes via 5-fold cross-validation (folds at the scene level, so no scene appears in both train and test) ridge regressions ($\alpha = 1.0$) on ground-plane x and y, and size-linked height z, and an RSA computation comparing the activation cosine-similarity matrix to the Gaussian-kernel similarity on 3D coordinates (§3, Eqs. 18–19). Probe metrics for Qwen2.5-VL-7B at the cognitive-map layer L17 (Fig. 4):

target	color	shape	x (R^2)	y (R^2)	size/z (R^2)	RSA (ρ)
results	0.51 ± 0.02	1.00 ± 0.00	$+0.09$	$+0.12$	$+0.28 \pm 0.02$	$+0.01 \pm 0.01$

Activation-variance subspace analysis. Because a linear-probe-accuracy gap could in principle hide a representation that is encoded faithfully but in fewer dimensions, we additionally compare the *activation variance* carried by each feature's optimal low-rank subspace. Concretely, the color subspace is the 7-dimensional class-mean-difference basis spanning the 8-color affine hull (i.e., the orthogonal complement of the global mean within the 8-class centroid subspace); the shape subspace is the 1-dimensional axis discriminating the 4 shapes after class-mean centering (rank deficiency from the categorical leaves a 3-dim residual, but the leading axis carries the bulk of the variance); the position subspace is the 3-dim basis recovered by ridge regression of 3D coordinates on H and orthogonalized against the color subspace. The fraction of the layer's Frobenius energy $\|H\|_F^2$ captured by each subspace, on Qwen2.5-VL-7B at L17:

$$v_{\text{color 7d}} = 11.86\% \pm 0.24, \quad v_{\text{shape 1d}} = 0.88\% \pm 0.05, \quad v_{\text{pos 3d}} = 0.085\% \pm 0.003.$$

Color and shape do not just appear in more dimensions; *each* of those dimensions is louder than its position counterpart — a $\sim 60\times$ gap per dimension and a $\sim 140\times$ aggregate gap between color and position variance (values reported above). The qualitative pattern replicates on InternVL3-8B at L18 (color 1.54%, shape 1.08%, position 0.066% — a $\sim 23\times$ aggregate gap with the matched small-basis estimator).

Counterfactual color/position swap. The counterfactual experiment summarized in §3 perturbs the input scene rather than the model's activations. We sample 50 base scenes from the Free6DoF corpus that have ≥ 4 distinct (color, shape) identities so the swap permutations are non-trivial. From each base scene we generate three rendered variants:

- *Original*: the scene rendered as-is.
- *Color-swap*: the per-object *colors* are permuted by a random derangement (no object keeps its original color), but every 3D position, shape, size, and camera trajectory is held fixed; the scene is re-rendered.
- *Position-swap*: the per-object *positions* are permuted by a random derangement, leaving every color, shape, size, and camera trajectory fixed; re-rendered.

Per scene we generate ~ 9 spatial questions across two kinds: *distance-order* ("Which is closer to the magenta cube: the purple sphere or the cyan cube?") and *relative-position* ("Is the yellow cylinder to the left of the red cylinder?"). Prompts and ground-truth answers are regenerated after each swap, yielding 1,290 QA pairs per random seed. The QA jsonl and pre-rendered variant frames live at `data/counterfactuals/`. We measure the per-(scene, question) *answer-flip rate* A color swap renames physical objects, so answer strings must be mapped back to the original object identities before comparing distance-order choices. Without this check, string changes alone cannot establish geometric reasoning failures.

$$\text{flip}_v \;=\; \Pr\big(\hat{y}_v \neq \hat{y}_{\text{orig}}\big), \qquad v \in \{\text{color-swap}, \text{position-swap}\},$$

under greedy decoding at `bf16`, averaged over $n = 4$ seeds. Per-question-kind breakdown on Qwen2.5-VL-7B (lam0 baseline):

question kind	color-swap flip	position-swap flip	expected (position-driven head)
distance-order	$\mathbf{41.0}\%$	48.6%	color-swap $\to 0\%$
relative-position	3.5%	7.3%	color-swap $\to 0\%$

The reported distance-order flip rates are 41.0% for color swaps and 48.6% for position swaps. Under the referent-preserving scoring protocol described above, a purely geometric decision should be invariant to a color renaming. A position swap can change the correct spatial answer, so its flip rate is a sensitivity comparison, not an error rate or a quantity expected to vanish.

C Causal verification by activation steering

This appendix expands on the "Causal verification" paragraph in §4 (Fig. 8). The motivation: linear probing and the variance-subspace analysis (Appendix B) only tell us what is *decodable* from the residual stream; they cannot rule out the possibility that the LM head simply ignores the probe's subspace and reads from a different code. To close that gap we ran an interventional experiment in which we directly inject a controlled vector along the probe's principal axis into the residual stream and measure whether the model's internal 3D readout shifts as predicted.

Setup.

backbone	Qwen2.5-VL-7B-Instruct
data	Tier-C Free6DoF, 5 scenes $\times$ 3 objects
steered layer	$\ell = 12$ (chosen by held-out probe R^2)
targeted axis	world-frame x (per-scene-normalized)
intervention magnitudes	$\alpha \in \{-0.30, -0.15, +0.15, +0.30\}$ (norm-x units)
trials per (direction, α)	15
control direction	length-matched null-space vector $v_\perp$

Probe-axis and control construction. We fit a Ridge probe at layer $\ell = 12$ that maps the pooled object-token residual-stream activation to per-scene-normalized 3D coordinates: $\hat{x} = Wh$ with $W \in \mathbb{R}^{3\times d}$ and ℓ_2 regularization $\alpha_{\text{ridge}} = 1.0$. The steering direction $v_{\text{axis}} \in \mathbb{R}^d$ is the column of $W^\dagger$ corresponding to the world-frame x axis (so injecting $\alpha \cdot v_{\text{axis}}$ shifts the probe's $\hat{x}$ readout by α to first order, while leaving $\hat{y}$ and $\hat{z}$ unchanged). The length-matched control $v_\perp$ is the projection of a random Gaussian vector onto the null space of W (so $Wv_\perp = 0$ to numerical precision), rescaled to $\|v_\perp\| = \|v_{\text{axis}}\|$. By construction $v_\perp$ encodes "a direction in the residual stream of the same magnitude as v_{axis} that the probe cannot read."

Intervention. For a chosen object in a chosen scene, we register a forward-pre-hook on layer 12 that adds $\alpha \cdot v$ to that object's residual at every visual-token position belonging to that object (mask-driven, same indices used for object-token pooling in §4); other tokens are untouched. We then re-run the model and re-read the probe's 3D-coordinate prediction $(\hat{x}, \hat{y}, \hat{z})$ for the same object. The diagnostic statistic is $\Delta\hat{x} = \hat{x}_{\text{intervention}} - \hat{x}_{\text{baseline}}$ (and analogously for $\hat{y}, \hat{z}$).

Targeted shift along v_{axis} vs. $v_\perp$. Effect of steering on the probe's x-axis readout (mean $\pm$ SEM, $n = 15$ per cell):

α	v_{axis}	$v_\perp$	ratio $\lvert axis\rvert/\lvert perp\rvert$	Welch t	p
-0.30	$\mathbf{-0.1020 \pm 0.027}$	$+0.0201 \pm 0.035$	$4.9\times$	-5.58	6.8×10^{-6}
-0.15	$\mathbf{-0.0558 \pm 0.030}$	$+0.0045 \pm 0.031$	$12.9\times$	—	—
$+0.15$	$\mathbf{+0.0706 \pm 0.033}$	$+0.0087 \pm 0.025$	$8.4\times$	—	—
$+0.30$	$\mathbf{+0.0908 \pm 0.021}$	$+0.001 \pm 0.030$	$\mathbf{82.2}\times$	$+5.04$	3.3×10^{-5}

The probe-readout shift along v_{axis} is monotonic in α and aligned with its sign; along $v_\perp$ it is at the noise floor across all four magnitudes. Welch t-tests at $\alpha = \pm 0.30$ separate the two arms at $p < 10^{-4}$.

Sign-correctness. Counting trials whose readout shift has the same sign as the injected α:

direction	trials matching $\text{sign}(\alpha)$	rate
v_{axis}	$54/60$	$\mathbf{90.0}\%$
$v_\perp$	$29/60$	48.3% (chance level)

The 90%-vs-chance gap rules out the alternative that the readout shift is a generic effect of perturbing the residual stream with any vector of the right magnitude.

Interpretation. Steering along the fitted probe axis changes its coordinate readout selectively relative to a matched null-space direction. Because the steering direction is constructed from this probe, this experiment supports controllability of the measured representation; it does not by itself show that generated spatial answers causally depend on that direction.

D Visualization examples

In Figures 13, 14, 15, 16, 17, 18, and 19, we show more examples of visualization for extracted spatial feature in different layers, similar to those in Figure 6.

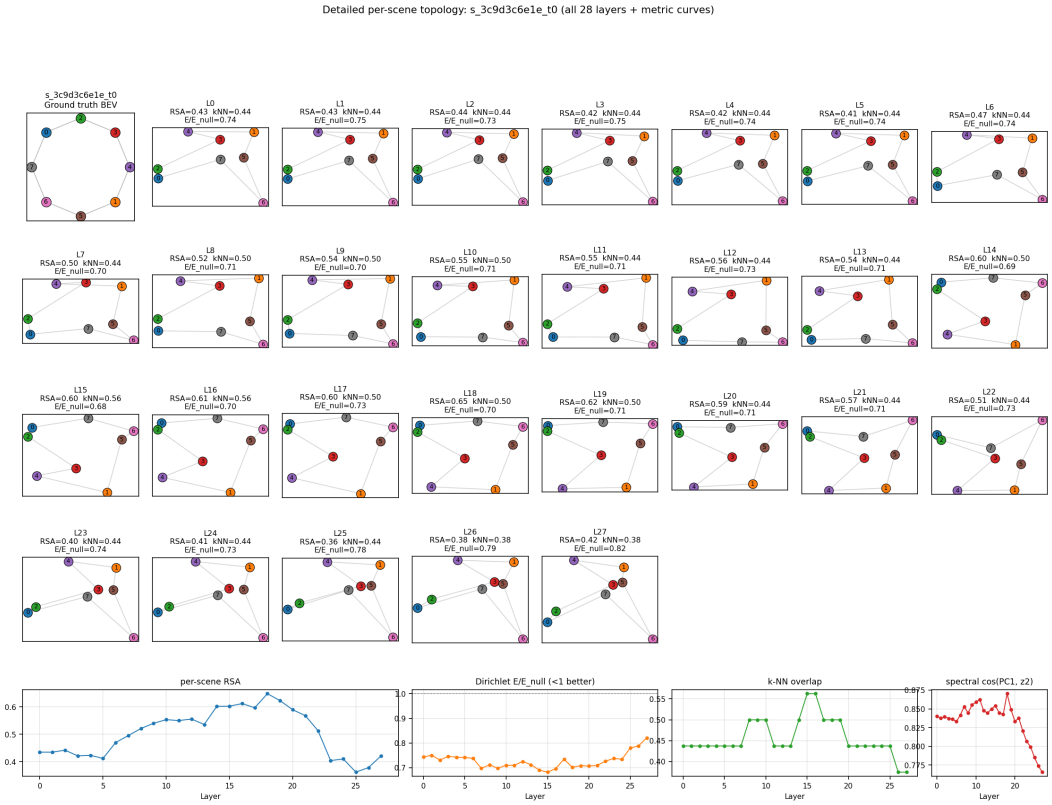


Figure 13: Circular Layout, Example 1

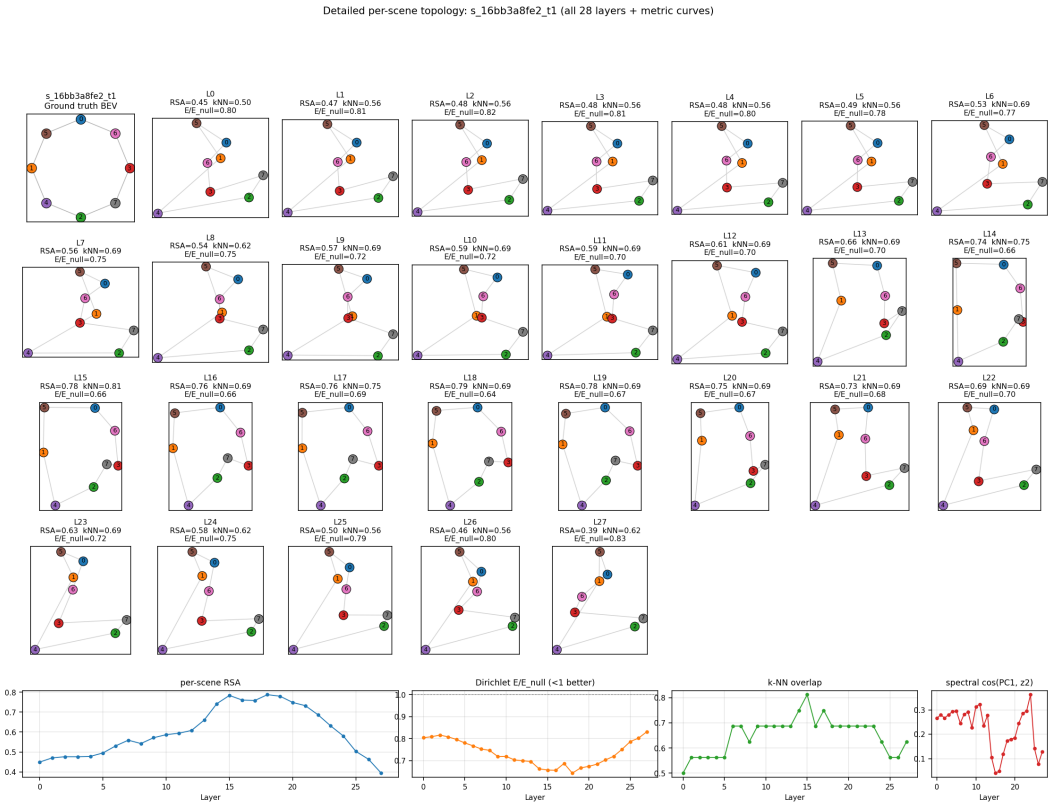


Figure 14: Circular Layout, Example 2

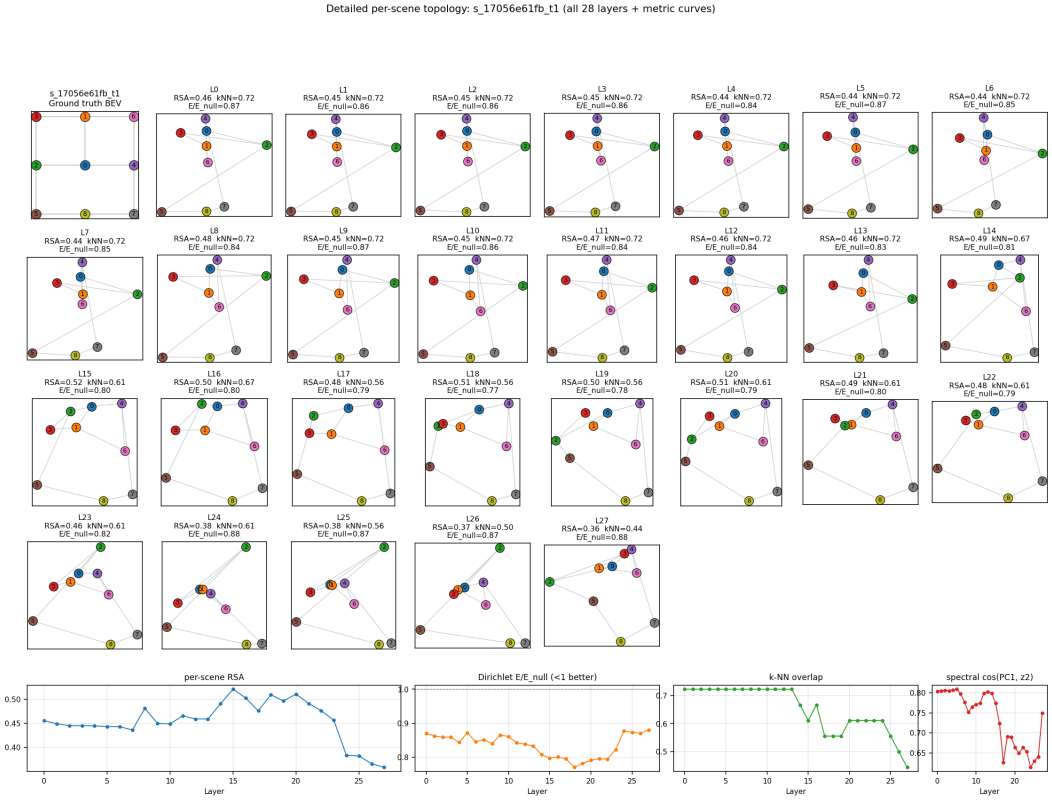


Figure 15: Grid Layout, Example 1

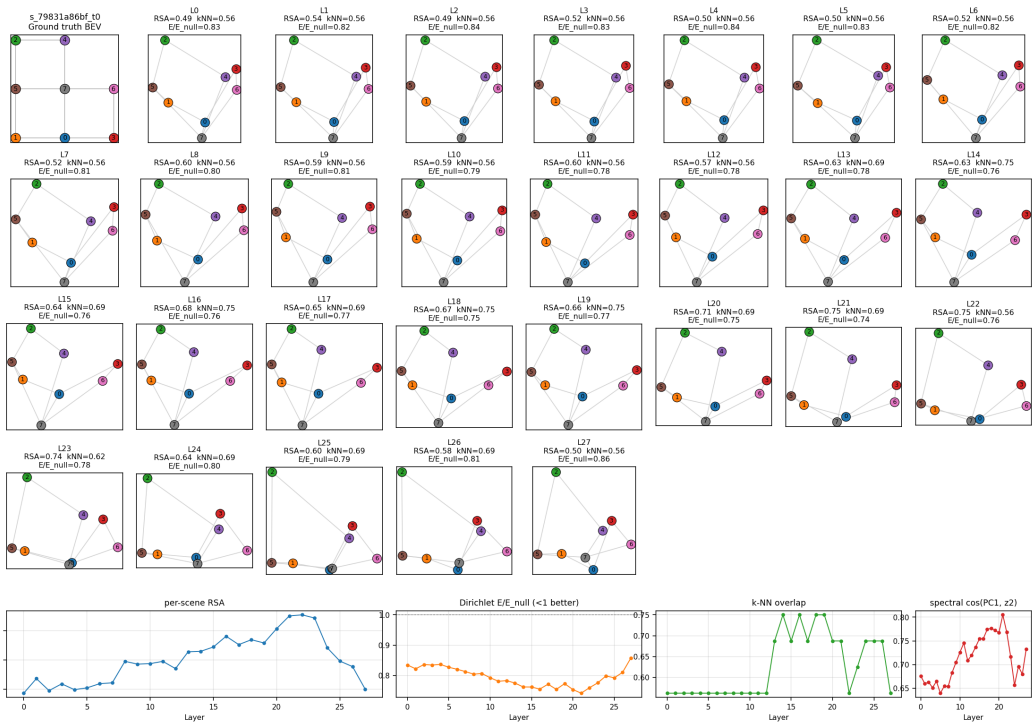


Figure 16: Grid Layout, Example 2

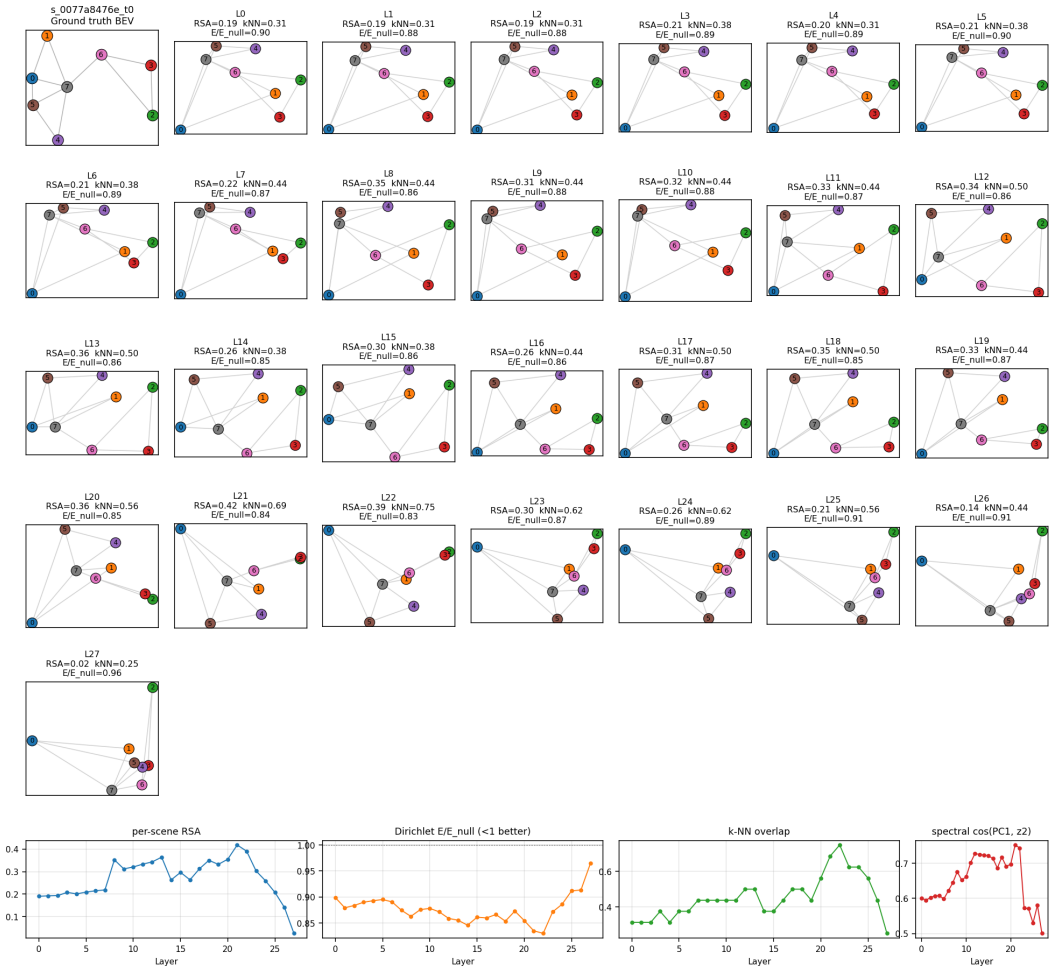


Figure 17: Random Layout, Example 1

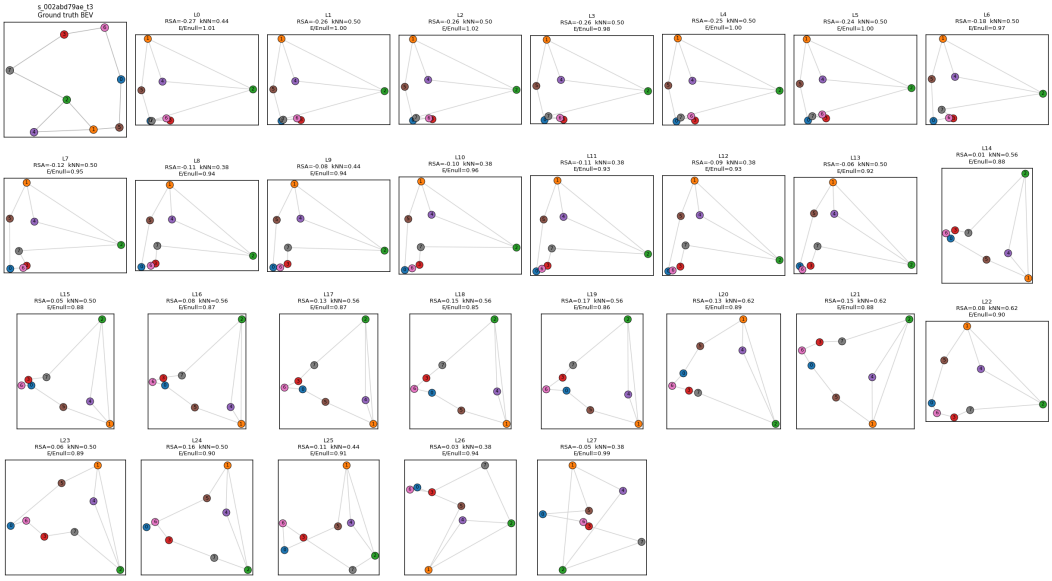


Figure 18: Random Layout, Example 2

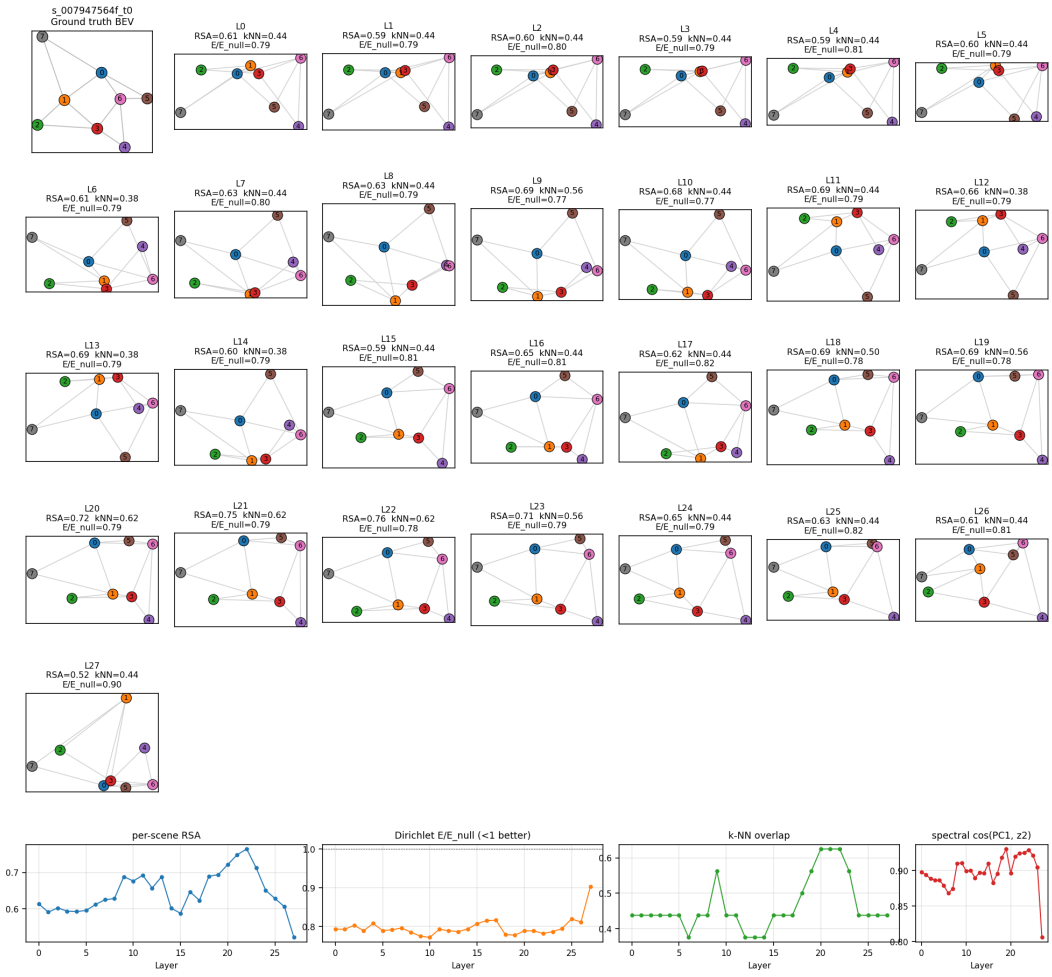


Figure 19: Random Layout, Example 3

E Choice of identity-attribute basis rank k and per-task results

This appendix supports two references in the main text: (i) the choice of the basis rank k in Eq. 5 for the identity-attribute subspace W_ℓ, and (ii) the extended per-task numbers underlying Tables 1, 2, and 3.

E.1 Choice of k for the identity-attribute basis

Recall that $\bar{H}_\ell \in \mathbb{R}^{K\times d}$ stacks the per-identity prototypes (§4) over K unique (color, shape) identities in the synthetic inventory — $K = 8 \times 3 = 24$ on the Free6DoF corpus. The basis $W_\ell \in \mathbb{R}^{d\times k}$ is then the top-k left singular vectors of $\bar{H}_\ell^\top$, and the projector $P_\perp = I_d - W_\ell W_\ell^\top$ residualizes against the span of identity directions.

Singular-value spectrum. Because color (8-way) and shape (3-way) are sampled independently in the data generator (Appendix A), the prototype cloud has full effective rank up to $K-1$ (one direction is consumed by the global mean). Empirically the spectrum of $\bar{H}_\ell^\top$ at the cognitive-map layer is well-approximated by a power-law decay with no sharp gap: the first 8 components carry $\sim 60\%$ of the prototype variance (the dominant color axes), the next 3 carry $\sim 20\%$ (shape), and the remaining ~ 13 carry the residual $\sim 20\%$ (color$\times$shape interactions). There is no clean "elbow" before $k = K-1$, so we keep all non-degenerate components.

Default choice. We set $k = K - 1 = 23$ on Free6DoF. This (i) spans the full identity prototype cloud up to global translation, (ii) does not require a hyperparameter search, and (iii) coincides with the rank that is naturally produced by orthonormalizing the multinomial logistic-regression directions for color and shape via thin QR after dropping the rank-deficient class direction per categorical (cf. Appendix I.2; effective rank $\le 7+2=9$ when only the discriminating directions are kept, and $\le K-1=23$ when the entire prototype cloud is used).

Sensitivity. We measured downstream RSA ρ_{topo} at the cognitive-map layer of Qwen2.5-VL-7B as a function of k:

k	4	8	11	16	23	32
ρ_{topo} at L17 (↑)	0.31	0.48	0.55	0.59	**0.61**	0.60
Color-probe acc. on $\widetilde{h}$ (↓, chance $= 1/8$)	0.42	0.21	0.16	0.14	**0.13**	0.13

RSA increases sharply up to $k \approx 11$, then plateaus by $k = 23$; pushing k past the prototype rank (e.g. $k = 32$) projects out genuinely random directions and degrades both metrics slightly. Since the curves are flat over $k \in [16, 32]$, the residualization is robust to mis-estimating k within this band; we report $k = 23$ in all main-paper experiments.

F Full Proof of Theorem 1 (Dirichlet Minimization)

This appendix provides the rigorous algebraic derivation for Theorem 1, which states that minimizing the Dirichlet energy under singular value constraints yields the Laplacian eigenmaps as principal components.

F.1 Background: The Weighted Ky Fan Inequality

We first establish a weighted form of Ky Fan's trace minimum theorem.

Lemma 6 (Ky Fan, weighted form)**.** *Let $L \in \mathbb{R}^{m\times m}$ be a symmetric matrix with eigenvalues $\lambda_1 \le \lambda_2 \le \cdots \le \lambda_m$ and corresponding orthonormal eigenvectors $z^{(1)}, \ldots, z^{(m)}$. Let $w_1 > w_2 > \cdots > w_s > 0$ be a strictly decreasing positive sequence (with $s \le m$). Then,*

$$\min_{\substack{u_1,\ldots,u_s\in\mathbb{R}^m \\ \langle u_i,u_j\rangle=\delta_{ij}}} \sum_{k=1}^{s} w_k \langle u_k, Lu_k\rangle \;=\; \sum_{k=1}^{s} w_k \lambda_k, \tag{14}$$

and the unique (up to sign) minimizer is $u_k = z^{(k)}$ for each k.

Proof. Substituting $u_k = z^{(k)}$ trivially yields $\langle z^{(k)}, Lz^{(k)} \rangle = \lambda_k$, attaining the right-hand side. The substantive part is showing that any orthonormal set $\{u_k\}$ achieves at least the right-hand side. We use summation by parts (Abel summation). Define cumulative weights $W_s \triangleq w_s$ and consecutive differences $\Delta_k \triangleq w_k - w_{k+1}$ for $k = 1, \ldots, s-1$. The strictly decreasing assumption guarantees $\Delta_k > 0$ for all k, and $W_s > 0$. We telescope the weights: $w_k = W_s + \sum_{j=k}^{s-1} \Delta_j$. Substituting this into the sum and swapping the order of summation yields:

$$\begin{aligned}\sum_{k=1}^{s} w_k \langle u_k, Lu_k \rangle &= \sum_{k=1}^{s} \left(W_s + \sum_{j=k}^{s-1} \Delta_j \right) \langle u_k, Lu_k \rangle \\ &= W_s \sum_{k=1}^{s} \langle u_k, Lu_k \rangle + \sum_{j=1}^{s-1} \Delta_j \sum_{k=1}^{j} \langle u_k, Lu_k \rangle.\end{aligned}$$

By the standard Ky Fan theorem for uniform weights (or the Poincaré separation theorem), any orthonormal prefix $\{u_1, \ldots, u_j\}$ satisfies $\sum_{k=1}^{j} \langle u_k, Lu_k \rangle \geq \sum_{k=1}^{j} \lambda_k$. Substituting this lower bound into every partial sum, we get:

$$\begin{aligned}\sum_{k=1}^{s} w_k \langle u_k, Lu_k \rangle &\geq W_s \sum_{k=1}^{s} \lambda_k + \sum_{j=1}^{s-1} \Delta_j \sum_{k=1}^{j} \lambda_k \\ &= \sum_{k=1}^{s} \left(W_s + \sum_{j=k}^{s-1} \Delta_j \right) \lambda_k \;=\; \sum_{k=1}^{s} w_k \lambda_k.\end{aligned}$$

The uniqueness up to sign follows strictly from the strict monotonicity of the weights, which forces equality on every individual partial sum, uniquely binding u_k to the eigenspace of $z^{(k)}$. □

F.2 Proof of Theorem 1

The proof proceeds in four explicit steps: decomposing the energy via SVD, saturating the constraints, identifying the optimal singular vectors via Lemma 6, and addressing the mean-centering.

Step 1: Decompose the energy along the SVD. We wish to minimize $\mathcal{E}_X(H) = \text{tr}(H^\top L H)$. Let the (thin) SVD of H be $H = U \Sigma V^\top$, where $U \in \mathbb{R}^{m \times r}$ and $V \in \mathbb{R}^{d \times r}$ have orthonormal columns, $\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_r)$, and $r = \min(m, d)$. Substituting this into the trace:

$$\text{tr}(H^\top L H) = \text{tr}((V \Sigma U^\top) L (U \Sigma V^\top)).$$

By the cyclic property of the trace, $\text{tr}(V \Sigma U^\top L U \Sigma V^\top) = \text{tr}(\Sigma U^\top L U \Sigma V^\top V)$. Since $V^\top V = I_r$, we obtain $\text{tr}(\Sigma^2 U^\top L U)$. Because Σ^2 is diagonal, this is precisely a weighted sum over the columns u_k of U:

$$\mathcal{E}_X(H) = \sum_{k=1}^{r} \sigma_k^2 \langle u_k, Lu_k \rangle. \tag{15}$$

Step 2: The lower-bound constraints saturate. Equation 15 is a sum of non-negative terms, because L is positive semi-definite (implying $\langle u_k, Lu_k \rangle \geq 0$). For each term, the singular value σ_k operates solely as a non-negative coefficient. Within the feasible region dictated by our constraints $\sigma_k \geq \epsilon_k > 0$ for $k \in [s]$ (and $\sigma_k \geq 0$ for $k > s$), minimizing the sum with respect to the singular values trivially pushes them to their absolute lowest bounds. Thus, at the minimum H^*, we have exactly $\sigma_k(H^*) = \epsilon_k$ for $k \leq s$, and $\sigma_k(H^*) = 0$ for $k > s$. The problem therefore reduces to optimizing the left-singular vectors u_k:

$$\min_{\substack{u_1, \ldots, u_s \in \mathbb{R}^m \\ u_i \perp u_j}} \sum_{k=1}^{s} \epsilon_k^2 \langle u_k, Lu_k \rangle. \tag{16}$$

Step 3: Apply Ky Fan. Equation 16 perfectly matches the hypothesis of Lemma 6, where our symmetric matrix is the graph Laplacian L, our weights are $w_k = \epsilon_k^2$, and our orthonormal vectors are the left singular vectors u_k. The assumption $\epsilon_1 > \cdots > \epsilon_s > 0$ ensures the weights are strictly decreasing. Therefore, by Lemma 6, the unique (up to sign) minimizers are the eigenvectors of L corresponding to its smallest eigenvalues:

$$u_k(H^*) = z^{(k)} \quad \text{for } k = 1, \ldots, s.$$

This proves part (a) of the theorem. Note that the right-singular vectors V drop out of the objective entirely, giving the "uniqueness up to right-orthogonal transformation" claim.

Step 4: Mean-centering and Principal Components. We now evaluate the PCA of the mean-centered optimal representation $H_c^* = P_{J^\perp} H^*$, where $P_{J^\perp} = I - \frac{1}{m}\mathbf{1}\mathbf{1}^\top$ is the mean-centering orthogonal projector. For the graph Laplacian $L = D - W$, the smallest eigenvalue is always $\lambda_1 = 0$, and its corresponding eigenvector is the uniform vector $z^{(1)} = \mathbf{1}/\sqrt{m}$. Because $P_{J^\perp}$ is the projector onto the orthogonal complement of $\mathbf{1}$, we have $P_{J^\perp} z^{(1)} = 0$. For $k \geq 2$, the eigenvectors of the symmetric matrix L are mutually orthogonal, so $z^{(k)} \perp \mathbf{1}$, meaning $P_{J^\perp} z^{(k)} = z^{(k)}$.

Applying this to the SVD of H^*:

$$H_c^* = P_{J^\perp}(U\Sigma V^\top) = (P_{J^\perp} U)\Sigma V^\top.$$

The columns of $P_{J^\perp} U$ are $P_{J^\perp} u_k$. For $k = 1$, $P_{J^\perp} z^{(1)} = 0$. For $k \geq 2$, $P_{J^\perp} z^{(k)} = z^{(k)}$. Therefore, the non-zero singular components of H_c^* start directly at $z^{(2)}, z^{(3)}, \ldots, z^{(s)}$. Equivalently, the k-th left singular vector (the k-th principal component direction) of H_c^* is exactly $z^{(k+1)}$, proving part (b).

G Proofs of Training-Time Guarantees (Theorems 2, 3, 4, 5)

We use the same notation as Appendix F: a fixed scene with m object-tokens, residual-stream representation $H \in \mathbb{R}^{m \times d}$, scene 3D coordinates $X \in \mathbb{R}^{m \times 3}$, Gaussian-kernel adjacency $W_{ij} = \exp(-\|x_i - x_j\|^2/2\tau^2)$ for $i \neq j$, degree matrix $D_{ii} = \sum_j W_{ij}$, unnormalized Laplacian $L = D - W$ with eigenpairs $(\lambda_k, z^{(k)})_{k=1}^m$, $\lambda_1 = 0 \leq \lambda_2 \leq \cdots$, and Dirichlet energy $\mathcal{E}_X(H) = \mathrm{tr}(H^\top L H)$. Throughout, $H^* = U^*\Sigma^*(V^*)^\top$ denotes the Theorem 1 minimizer of $\mathcal{E}_X$ subject to $\sigma_k(H) \geq \epsilon_k$ for $k \in [s]$, and $V^*_{:,1:3}$ its top-3 right singular vectors. Note that, by Theorem 1(b), the columns of the mean-centered U^* at indices 1–3 are exactly the Laplacian eigenvectors $z^{(2)}, z^{(3)}, z^{(4)}$.

G.1 Proof of Theorem 2 (continuous limit)

The proof has three ingredients drawn from spectral graph theory and manifold learning, plus a final identification step on the Euclidean cube. We sketch the argument; full proofs of the operator-convergence results below are standard and can be found in the cited references.

Step 1 (pointwise operator limit, Belkin–Niyogi 2003). Let $\widetilde{L}_m^{(\tau)}$ denote the degree-rescaled empirical Laplacian

$$\widetilde{L}_m^{(\tau)} f(x) \;:=\; \frac{1}{\tau^{d_\Omega+2}} \left[f(x) \cdot \tfrac{1}{m}\sum_{j=1}^m \kappa_\tau(x, x_j) \;-\; \tfrac{1}{m}\sum_{j=1}^m \kappa_\tau(x, x_j)\, f(x_j) \right],$$

the function-space lift of the matrix $\frac{1}{m\tau^{d_\Omega+2}} L^{(\tau)}$, where $d_\Omega = 3$ is the intrinsic dimension of Ω. Belkin and Niyogi (2003, Thm. 3.1) show that for any $f \in C^2(\Omega)$ and any interior point $x \in \Omega$,

$$\widetilde{L}_m^{(\tau)} f(x) \xrightarrow[m\to\infty,\, \tau\to 0]{\text{a.s.}} -c_d\, \rho(x)\, \Delta_\rho f(x), \qquad \Delta_\rho f := \rho^{-1}\, \mathrm{div}(\rho \nabla f),$$

where $c_d > 0$ is a kernel-dependent constant and Δ_ρ is the weighted Laplace–Beltrami operator. The proof is a Taylor expansion of f at x to second order, using kernel symmetry to kill the first-order term and the Hessian trace to produce the Laplacian. This step gives the *limit operator*, but on its own says nothing about eigenvector convergence.

Step 2 (spectral convergence, von Luxburg–Belkin–Bousquet 2008). Lift the matrix to an integral operator $\mathcal{T}_m^{(\tau)} f(x) := \frac{1}{m} \sum_j \kappa_\tau(x, x_j) f(x_j)$ on $C(\Omega)$, with population analog $\mathcal{T}^{(\tau)} f(x) := \int_\Omega \kappa_\tau(x, y) f(y) \rho(y)\, dy$. Theorem 21 of vLBB states that under (i) continuous bounded symmetric positive kernel, (ii) i.i.d. samples from $\rho > 0$, (iii) isolated finite-multiplicity eigenvalues of $\mathcal{T}^{(\tau)}$, the empirical eigenvalues and eigenfunctions converge:

$$|\lambda_k(\mathcal{T}_m^{(\tau)}) - \lambda_k(\mathcal{T}^{(\tau)})| \to 0, \qquad \|z_m^{(k)} - \phi_\tau^{(k)}\|_{L^2(\rho)} \to 0,$$

where $z_m^{(k)}$ is the Nyström-extended k-th eigenvector. The proof uses *collectively compact operator theory*: $\mathcal{T}_m^{(\tau)} \to \mathcal{T}^{(\tau)}$ collectively compactly, and isolated eigenvalues of compact operators persist under such perturbation with eigenprojection convergence in operator norm. The required collectively compact regime corresponds to $m\tau_m^{d_\Omega+2}/\log m \to \infty$ in the sample regime — the bandwidth condition restated in the theorem.

Step 3 (combining Steps 1 and 2 as $\tau \to 0$). Step 2 gives convergence to $\phi_\tau^{(k)}$ at *fixed* bandwidth; Step 1 says that as $\tau \to 0$, the rescaled $\mathcal{T}^{(\tau)}$-eigenproblem converges to the eigenproblem of $-\Delta_\rho$. Provided $\tau_m \to 0$ at a rate slow enough to maintain the collectively compact regime ($m\tau_m^{d_\Omega+2} \to \infty$, i.e. the theorem's bandwidth condition with $\alpha > 0$), the empirical eigenvector $z_m^{(k+1)}$ converges uniformly on samples to the k-th non-constant eigenfunction of $-\Delta_\rho$, with the $\rho^{-1/2}$ factor coming from the standard weighted-Laplace–Beltrami normalization.

Step 4 (identification on the Euclidean cube). On $\Omega = [0, 1]^3$ with uniform ρ, the weighted Laplace–Beltrami reduces to the ordinary 3D Laplacian $\Delta = \partial_x^2 + \partial_y^2 + \partial_z^2$. The kernel construction induces Neumann boundary conditions in the limit (it does not impose zero Dirichlet values, only zero normal derivative). The Neumann eigenproblem $-\Delta\phi = \mu\phi$ on $[0, 1]^3$ has eigenfunctions $\cos(k_1\pi x)\cos(k_2\pi y)\cos(k_3\pi z)$ with eigenvalues $(k_1^2 + k_2^2 + k_3^2)\pi^2$. The smallest eigenvalue is $\mu_0 = 0$ (constant mode); the next three are all $\mu_1 = \pi^2$, spanned by $\cos(\pi x), \cos(\pi y), \cos(\pi z)$. Each is a strictly monotone function of the corresponding Cartesian coordinate on $[0, 1]$, so PCA on samples recovers the centered coordinates up to this monotone reparameterization. Sign-tests (the form of every axis-aligned ordinal question) are invariant under this reparameterization. □

Why three pieces, not one. Pointwise operator convergence (Step 1) does not imply spectral convergence: bounded operator sequences can have eigenvalues escaping to infinity or merging with neighbors. The substantive content is Step 2 (collectively compact + spectral gap ⇒ eigenvector convergence with quantitative control), and the precise rate $m\tau_m^{d_\Omega+2} \to \infty$ comes from the regime where Step 2's hypotheses remain valid as $\tau \to 0$.

G.2 Proof of Theorem 3 (realizability of spatial readouts)

Setup. Write the singular vectors of the Theorem-1-minimizing H^* as u_k^*, v_k^*, with singular values $\sigma_k^* = \epsilon_k$ for $k \in [s]$; assume $s \geq 4$ and the conditions of Theorem 2. By Theorem 1(b), the k-th left singular vector of the mean-centered H_c^* is $z^{(k+1)}$ for $k = 1, 2, 3$; equivalently, the k-th principal-component score of object i is

$$\xi_{i,k} \;:=\; \langle h_i^*, V_{:,k}^* \rangle / \sigma_k^* \;=\; z_i^{(k+1)}. \tag{E.1}$$

By Theorem 2, under the stated bandwidth and sample-size conditions, $z^{(k+1)}$ converges uniformly on samples to a Laplacian eigenfunction ϕ_k (Cartesian coordinate on the uniform-density cube, up to a monotone cosine reparameterization):

$$\sup_{i \in [m]} \left|\xi_{i,k} - \phi_k(x_i)\right| \;\xrightarrow[m\to\infty]{}\; 0, \qquad k = 1, 2, 3. \tag{E.2}$$

Step 1 (eigen-expansion of f). The hypothesis of Theorem 3 is that $f : \mathbb{R}^3 \to \mathbb{R}$ is a linear function of the monotonic coordinate transformations ϕ_1, ϕ_2, ϕ_3. Hence there exist constants $c_1, c_2, c_3 \in \mathbb{R}$ (computed by Galerkin projection $c_k = \int_\Omega f \cdot \phi_k \, d\rho$, finite under the stated regularity) such that

$$f(x) \;=\; \sum_{k=1}^{3} c_k\, \phi_k(x), \qquad x \in \Omega. \tag{E.3}$$

Step 2 (construction of the linear functional). Define

$$\ell(h) \;:=\; \sum_{k=1}^{3} \frac{c_k}{\sigma_k^*}\,\langle h, V_{:,k}^*\rangle, \qquad h \in \mathbb{R}^d. \tag{E.4}$$

This is linear in h and depends only on $(c_k, \sigma_k^*, V_{:,k}^*)_{k=1}^3$. Applied to h_i^* and using (E.1),

$$\ell(h_i^*) \;=\; \sum_{k=1}^{3} c_k\,\xi_{i,k}.$$

Step 3 (uniform error bound). Combining (E.2)–(E.4),

$$\big|\ell(h_i^*) - f(x_i)\big| \;=\; \left|\sum_{k=1}^{3} c_k\,(\xi_{i,k} - \phi_k(x_i))\right| \;\le\; \left(\sum_{k=1}^{3}|c_k|\right)\cdot \max_{k\in\{1,2,3\}} |\xi_{i,k} - \phi_k(x_i)|.$$

Choose $m_0(\eta)$ large enough that the right-hand side is $< \eta$ uniformly in $i \in [m]$; existence of such m_0 is exactly (E.2). □

Remark (extension to higher-degree readouts). If the readout function depends polynomially of degree $\le r$ on x, the eigen-expansion uses the leading $K(r) \sim \binom{r+3}{3}$ Laplacian eigenfunctions; Theorem 1 then needs to be invoked with $s \ge K(r)+1$. The argument generalizes term-by-term. For the axis-aligned ordinal questions used in our spatial benchmarks (front/back, left/right, distance-order), $r = 1$ suffices and $s = 4$ is the operative rank.

G.3 Proof of Theorem 4 (sample-complexity reduction)

Setup. Let $\{(\Phi_i, y_i)\}_{i=1}^N$ be i.i.d. object-token observations, $\Phi_i = h_i^* \in \mathbb{R}^d$ drawn from the population distribution induced by Theorem-1-minimizing scenes, with $y_i = w^{\star\top}\Phi_i + \xi_i$, ξ_i sub-Gaussian with variance proxy σ_ξ^2, and $w^\star \in \mathrm{span}(V_{:,1:3}^*)$. Let $\widehat{w}_N^{\mathrm{proj}}$ denote the OLS estimator obtained by first projecting features onto $V_{:,1:3}^*$ and regressing the labels on the resulting 3-dimensional features.

Step 1 (change of basis). Let $\Psi_i := V_{:,1:3}^{*\top}\Phi_i \in \mathbb{R}^3$. Stack into $\Psi \in \mathbb{R}^{N\times 3}$ and $y \in \mathbb{R}^N$. By the structural assumption there exists $a \in \mathbb{R}^3$ with $w^\star = V_{:,1:3}^* a$, so $w^{\star\top}\Phi_i = a^\top\Psi_i$. The model reduces to

$$y_i = a^\top\Psi_i + \xi_i, \qquad i = 1,\dots,N. \tag{E.5}$$

Step 2 (projected OLS).

$$\widehat{a}_N \;=\; (\Psi^\top\Psi)^{-1}\Psi^\top y, \qquad \widehat{w}_N^{\mathrm{proj}} := V_{:,1:3}^*\widehat{a}_N.$$

By construction $\widehat{w}_N^{\mathrm{proj}}$ has zero component in $\mathrm{span}(V_{:,4:d}^*)$, matching the sparsity pattern of $w^\star$, so the bias is zero.

Step 3 (variance bound). With the population PC-score covariance $\Sigma_\Psi := \mathbb{E}[\Psi\Psi^\top]$ assumed non-degenerate, the standard fixed-design analysis (Wainwright, *High-Dimensional Statistics*, Thm. 13.4) gives

$$\mathbb{E}\big[\|\widehat{a}_N - a\|_{\Sigma_\Psi}^2 \,\big|\, \Psi\big] \;=\; \sigma_\xi^2\cdot \mathrm{tr}\big((\Psi^\top\Psi)^{-1}\Sigma_\Psi\big) \;\xrightarrow[N\to\infty]{}\; \frac{3\sigma_\xi^2}{N},$$

since $(\Psi^\top\Psi/N)^{-1} \to \Sigma_\Psi^{-1}$ and $\mathrm{tr}(\Sigma_\Psi^{-1}\Sigma_\Psi) = 3$. Because $w^\star$ aligns exactly with $\mathrm{span}(V_{:,1:3}^*)$, the prediction error decomposes as pure variance:

$$\mathbb{E}_\Phi\Big[(\widehat{w}_N^{\mathrm{proj}\top}\Phi - w^{\star\top}\Phi)^2\Big] \;=\; \mathbb{E}\big[(\widehat{a}_N - a)^\top\Sigma_\Psi(\widehat{a}_N - a)\big] \;\le\; \frac{3\sigma_\xi^2}{N}\big(1 + o(1)\big).$$

Step 4 (high-probability upgrade). Local-Rademacher / Gaussian-complexity bounds for the linear class $\{\Psi \mapsto a^\top\Psi : a \in \mathbb{R}^3\}$ (Wainwright Thm. 13.4) upgrade the expectation bound to

$$\mathbb{P}\left(\mathbb{E}_\Phi\Big[(\widehat{w}_N^{\mathrm{proj}\top}\Phi - w^{\star\top}\Phi)^2\Big] \;\le\; \frac{C\,\sigma_\xi^2\,(3 + \log(1/\delta))}{N}\right) \;\ge\; 1 - \delta,$$

which is the form stated in the theorem (modulo the constant absorbing the sub-Gaussian factor and $\|\Sigma_\Psi^{-1}\|_{\mathrm{op}} \cdot \|\Sigma_\Psi\|_{\mathrm{op}}$).

Step 5 (comparison to unstructured). The unstructured OLS $\widehat{w}_N^{\mathrm{full}} = (\Phi^\top \Phi)^\dagger \Phi^\top y$ decomposes as $V^*_{:,1:3}\widehat{a}_N + V^*_{:,4:d}\widehat{b}_N$. The tail estimator $\widehat{b}_N$ regresses pure noise on $d-3$ orthogonal directions, contributing variance $(d-3)\sigma_\xi^2/N$. Total prediction error therefore scales as $d\,\sigma_\xi^2/N$ — $d/3$ times worse than the projected estimator. □

Remark. The proof presupposes knowing $V^*_{:,1:3}$ to perform the projection. In practice, this is what motivates *training* with the Dirichlet loss: by Theorem 1, the model itself learns to align its top PCs with the Laplacian eigenmaps, so the projection is realized in the model rather than imposed by a separate processing step. Theorem 5 formalizes the connection.

G.4 Proof of Theorem 5 (training-time risk decomposition)

Setup. Let $\Theta \subset \mathbb{R}^p$ be open, $\mathcal{L}_{\mathrm{LM}}, \mathcal{R}_X : \Theta \to \mathbb{R}$ both C^2, with $\nabla^2 \mathcal{L}_{\mathrm{LM}}(\theta^*) \succ 0$ at the unique population minimizer θ^* of $\mathcal{L}_{\mathrm{LM}}$. Let $R_{\mathrm{spatial}} : \Theta \to \mathbb{R}$ be C^2 and define $\theta^*_\lambda := \arg\min_\theta (\mathcal{L}_{\mathrm{LM}}(\theta) + \lambda\,\mathcal{R}_X(\theta))$.

Step 1 (implicit function theorem). Define

$$F(\theta, \lambda) \;:=\; \nabla_\theta \mathcal{L}_{\mathrm{LM}}(\theta) + \lambda\,\nabla_\theta \mathcal{R}_X(\theta).$$

Then $F(\theta^*, 0) = 0$ by first-order optimality, and the Jacobian $\partial F/\partial\theta\,|_{(\theta^*,0)} = \nabla^2 \mathcal{L}_{\mathrm{LM}}(\theta^*) \succ 0$ is invertible. By the implicit function theorem (Krantz & Parks, *The Implicit Function Theorem*), there exist a neighborhood $\mathcal{U} \ni 0$ in $\mathbb{R}$ and a unique C^2 map $\theta^*(\cdot) : \mathcal{U} \to \Theta$ with $\theta^*(0) = \theta^*$ and $F(\theta^*(\lambda), \lambda) = 0$ for all $\lambda \in \mathcal{U}$. By continuity, $\nabla^2(\mathcal{L}_{\mathrm{LM}} + \lambda\mathcal{R}_X)(\theta^*(\lambda)) \succ 0$ for λ small, so $\theta^*(\lambda)$ is a strict local minimizer; we set $\theta^*_\lambda := \theta^*(\lambda)$ for $\lambda \in [0, \lambda_0]$ where λ_0 is the largest value for which this remains the global minimizer.

Step 2 (first derivative of $\theta^*(\lambda)$). Differentiating $F(\theta^*(\lambda), \lambda) = 0$ in λ:

$$\nabla^2_\theta(\mathcal{L}_{\mathrm{LM}} + \lambda\mathcal{R}_X)(\theta^*(\lambda))\,\frac{d\theta^*}{d\lambda} + \nabla\mathcal{R}_X(\theta^*(\lambda)) \;=\; 0.$$

At $\lambda = 0$:

$$\left.\frac{d\theta^*(\lambda)}{d\lambda}\right|_{\lambda=0} \;=\; -\,[\nabla^2 \mathcal{L}_{\mathrm{LM}}(\theta^*)]^{-1}\,\nabla\mathcal{R}_X(\theta^*).$$

Taylor-expanding the smooth map $\lambda \mapsto \theta^*(\lambda)$:

$$\theta^*_\lambda - \theta^* \;=\; -\lambda\,[\nabla^2 \mathcal{L}_{\mathrm{LM}}(\theta^*)]^{-1}\,\nabla\mathcal{R}_X(\theta^*) + O(\lambda^2). \tag{E.6}$$

Step 3 (Taylor expansion of the spatial risk). Since $R_{\mathrm{spatial}} \in C^2$,

$$R_{\mathrm{spatial}}(\theta^*_\lambda) - R_{\mathrm{spatial}}(\theta^*) \;=\; \nabla R_{\mathrm{spatial}}(\theta^*)^\top(\theta^*_\lambda - \theta^*) + \tfrac{1}{2}(\theta^*_\lambda - \theta^*)^\top \nabla^2 R_{\mathrm{spatial}}(\tilde{\theta})\,(\theta^*_\lambda - \theta^*),$$

for some $\tilde{\theta}$ on the segment $[\theta^*, \theta^*_\lambda]$. Substituting (E.6) into the linear term:

$$\nabla R_{\mathrm{spatial}}(\theta^*)^\top(\theta^*_\lambda - \theta^*) \;=\; -\lambda\,\nabla R_{\mathrm{spatial}}^\top[\nabla^2 \mathcal{L}_{\mathrm{LM}}]^{-1}\nabla\mathcal{R}_X + O(\lambda^2).$$

The quadratic term in (E.6) contributes $O(\lambda^2)$ as well. Combining,

$$R_{\mathrm{spatial}}(\theta^*_\lambda) - R_{\mathrm{spatial}}(\theta^*) \;=\; -\lambda\,\nabla R_{\mathrm{spatial}}^\top[\nabla^2 \mathcal{L}_{\mathrm{LM}}]^{-1}\nabla\mathcal{R}_X \;+\; O(\lambda^2). \tag{E.7}$$

Step 4 (define β). The available improvement in Dirichlet ratio at θ^* is

$$\Delta \;:=\; \mathcal{R}_X(\theta^*) - \mathcal{R}_X^* \;\geq\; 0,$$

strictly positive whenever θ^* does not coincide with the Dirichlet minimizer (the generic case for any LM not specifically pretrained for spatial structure). Define

$$\beta \;:=\; \frac{\nabla R_{\mathrm{spatial}}(\theta^*)^\top\,[\nabla^2 \mathcal{L}_{\mathrm{LM}}(\theta^*)]^{-1}\,\nabla\mathcal{R}_X(\theta^*)}{\Delta}.$$

Substituting into (E.7):

$$R_{\mathrm{spatial}}(\theta^*_\lambda) \;=\; R_{\mathrm{spatial}}(\theta^*) - \lambda\,\beta\,\Delta + O(\lambda^2),$$

exactly the form claimed in the theorem.

Step 5 (sign of β). The numerator of β is the inner product of two vectors:

- $g := \nabla R_{\text{spatial}}(\theta^*)$, the steepest-ascent direction of the spatial risk;
- $r := [\nabla^2 \mathcal{L}_{\text{LM}}(\theta^*)]^{-1} \nabla \mathcal{R}_X(\theta^*)$, the inverse-Hessian-warped Dirichlet-ascent direction.

$\beta > 0$ iff $\langle g, r \rangle > 0$, i.e., iff the gradient direction along which the Dirichlet ratio decreases is correlated with the direction along which the spatial risk decreases, in the metric induced by $\nabla^2 \mathcal{L}_{\text{LM}}$. Theorems 3 and 4 provide the qualitative reason this correlation holds for VLMs whose top-3 PCs already partially align with world coordinates: lowering $\mathcal{R}_X$ tightens that alignment, which makes more spatial readouts realizable (Theorem 3) and reduces their estimation variance (Theorem 4). □

Remark on the validity range. The first-order term dominates whenever $\lambda\beta\Delta \gg \lambda^2 M_2 \|\nabla R_{\text{spatial}}\|$ for M_2 the bound on the Hessian remainder, i.e., $\lambda \ll \beta\Delta/(M_2\|\nabla R_{\text{spatial}}\|)$. Beyond this regime the $O(\lambda^2)$ over-regularization term takes over, predicting non-monotonic behavior in λ. The empirical inversion in Table 4 (reported settings $\lambda = 3$ for Qwen and $\lambda = 1$ for InternVL, with degradation at $\lambda = 9$; the Qwen sweep discrepancy remains marked) is the data-driven location of this crossover.

Remark on what β bundles. β is a single scalar but it carries three model-specific quantities: (i) $\|\nabla R_{\text{spatial}}(\theta^*)\|$ — how strongly spatial questions react to small parameter perturbations at the LM-loss optimum; (ii) the spectrum of $\nabla^2 \mathcal{L}_{\text{LM}}(\theta^*)$ — a flatter LM optimum makes the inverse Hessian larger, amplifying the regularizer's effect; (iii) the angle between $\nabla R_{\text{spatial}}$ and $\nabla \mathcal{R}_X$ in the whitened metric — how aligned the two losses are. A model whose pretraining leaves a flat LM optimum near a spatially-aware sub-manifold has a large β, predicting the InternVL-larger-than-Qwen ordering observed in Table 4.

H Training Details

This appendix expands on the training setup summarized in §5.3. The end-to-end training script is at `scripts/train_with_dirichlet.py` (included in the supplemental materials); the Dirichlet objective itself is implemented in `scripts/dirichlet_loss.py` and consumed by the script through a forward hook on the chosen residual-stream layer.

Backbones and adapter. We fine-tune two open-weights VLMs in identical configurations: Qwen2.5-VL-7B [39], InternVL3-8B [40]. The vision tower and all non-LoRA LLM weights are frozen. We attach LoRA adapters of rank r=16 (with α=32, dropout 0.05, no bias) to the four attention projections $\{q_proj, k_proj, v_proj, o_proj\}$ of every transformer block in the backbone-specific "cognitive-map band" (the five-layer InternVL3-8B ablation is centered at L18). Layers outside this band are not in the LoRA target set, so the early visual-binding layers and the late linguistic-reasoning layers see no parameter updates — this restriction is what makes "500 steps" a meaningful budget (cf. §5.3).

Loss. Given a batch of (image, question, answer, object_coords) tuples, the per-step loss is

$$\mathcal{L}_{\text{total}} \;=\; \mathcal{L}_{\text{CE}}(\text{answer} \mid \text{image}, \text{question}) \;+\; \lambda_{\text{dir}} \cdot \mathcal{R}_X\big(H_\ell(\theta) P_\perp\big),$$

where $\mathcal{L}_{\text{CE}}$ is standard next-token cross-entropy on the answer tokens, H_ℓ is the residual-stream activation at the cognitive-map layer ℓ pooled over each object's visual tokens (mask-driven, coverage-weighted; cf. §4), $P_\perp$ is the linear projection that residualizes against the cross-scene color/shape prototypes (§4), and $\mathcal{R}_X$ is the normalized Dirichlet term of Eq. 17 against the per-scene 3D-coordinate kernel graph $W_{ij} = \kappa_\tau(x_i, x_j)$ with τ set per scene to half the median pairwise distance. We use the multi-layer variant (*residMulti*) by default, which applies the Dirichlet term at 5 consecutive layers spanning the cognitive-map band (with equal weight); layer-count and placement ablations are reported in Appendix J.

Optimizer and schedule. AdamW with β_1=0.9, β_2=0.999, weight-decay 0.0 on LoRA parameters, learning-rate 1×10^{-4}, linear warmup over the first 50 steps, and constant thereafter. Gradients are clipped to global norm 1.0. We train at `bf16` mixed precision with effective batch size 4 (per-GPU batch 1, gradient accumulation 4) for 500 optimizer steps; the final split size is marked for confirmation in Appendix A. No early stopping — we keep the final checkpoint.

λ_{dir} **sweep.** For each (model, regularizer-variant) cell we sweep $\lambda_{\text{dir}} \in \{0,\ 0.3,\ 1,\ 3,\ 9\}$ where λ_{dir}=0 recovers a pure-LoRA baseline (*lam0*). The headline results in Tables 1, 2, and 3 use the λ chosen by best mean validation accuracy across n=4 seeds: λ_{dir}=3 for Qwen2.5-VL-7B, and λ_{dir}=1 for residMulti on InternVL3-8B. The full λ-vs-accuracy/RSA trade-off curve is reported in §5.5. The earlier value 0.3 came from a different layer configuration and is not the reported Qwen main-result setting.

Seeds and statistical reporting. The original main-result cells are averaged over n=4 random seeds (seeds $0, 1, 2, 3$), each controlling LoRA initialization, data shuffling, and dropout. We report mean $\pm 2\,\mathrm{SE}$ of seed-paired differences against the relevant within-cell baseline. Cells with $|\Delta| < 2$ pp at n=4 should be treated as exploratory — our project's track record is that small effects at n=4 sometimes shrink at n=8.

Baselines. (1) *Naive LoRA* (λ_{dir}=0): identical setup minus the Dirichlet term. (2) *2D-pixel regularization*: the same penalty applied with W_{ij} built from 2D pixel-space distances rather than 3D coordinates — a control that severs the connection between the loss target and the world-frame geometry. (3) *Text-domain cognitive-map prediction* [24]: the same LoRA budget redirected to a supervised loss that asks the model to emit the scene's 3D coordinates as text tokens. All baselines use identical learning rate, schedule, batch size, step count, LoRA rank and target modules, layer band, and seed pool, so the only varying axis is the training signal.

Evaluation. Checkpoints are evaluated zero-shot on the full 5,130-item VSI-Bench [23] and MindCube [24] test sets without any task-specific prompting. We use mean-relative-accuracy scoring on numeric-answer tasks (consistent with VSI-Bench's official protocol). Inference uses greedy decoding at `bf16`.

I Others

I.1 Limitation and discussion

While our findings demonstrate the efficacy of isolating and shaping latent spatial representations in VLMs, our current scope is primarily limited to scene topology. Spatial understanding is inherently multifaceted; future work will investigate how other critical aspects of spatial relationships—such as absolute metric distance and hierarchical object dependencies—are encoded within the model's residual stream. Building upon these topological foundations, we also plan to explore how applying our Dirichlet-energy regularization to fine-tuning on complex, real-world datasets can unlock even more substantial performance improvements in downstream tasks. The empirical evidence primarily concerns ground-plane relations in rendered scenes and transfer to two real-world benchmarks. Linear residualization is approximate and may discard spatial information or retain nuisance features. The method is evaluated as post-training adaptation of pretrained VLMs; applying it from scratch would require a procedure for estimating a changing identity basis and has not been tested. Preliminary model-size and MMMU checks are reported in Appendix J; they do not establish universal depth patterns or preservation of all non-spatial abilities.

I.2 Dirichlet loss implementation and residualization basis fitting

This subsection details the step-by-step computation that translates the abstract loss defined in § 5.2 into the concrete tensor operations implemented in `scripts/dirichlet_loss.py` and the offline basis-fitting procedure in `scripts/build_residualization_basis.py`. The notation follows § 5: $H_\ell \in \mathbb{R}^{B\times n\times d}$ denotes the residual stream at the cognitive-map layer, pooled over per-object visual tokens (§ 4); $X \in \mathbb{R}^{B\times n\times 3}$ represents the per-scene ground-truth 3D coordinate tensor; and $W \in \mathbb{R}^{d\times k}$ is the orthonormal nuisance basis, fitted once per backbone prior to training.

Per-step Dirichlet ratio. Given a batch of B scenes with up to n objects each, the regularizer evaluates

$$\mathcal{R}_X(H) \;=\; \frac{1}{B}\sum_{b=1}^{B} \frac{\sum_{i,j} W_{ij}^{(b)}\, \|h_i^{(b)} - h_j^{(b)}\|^2}{\sum_{i,j} \|h_i^{(b)} - h_j^{(b)}\|^2 + \varepsilon}, \qquad W_{ij}^{(b)} \;=\; \exp\!\Big(-\tfrac{\|x_i^{(b)} - x_j^{(b)}\|^2}{2\tau^2}\Big)\, \mathbb{1}[i \neq j], \tag{17}$$

with $\varepsilon = 10^{-8}$ for numerical stability. The numerator and denominator are computed in vectorized form via `torch.cdist`; valid-object masks are applied as outer products $\mathbb{1}[\text{valid}_i] \cdot \mathbb{1}[\text{valid}_j]$ to handle batches with variable n. Normalizing by the denominator (rather than using the unnormalized energy $\sum_{ij} W_{ij}\|h_i - h_j\|^2$) makes the loss *scale-invariant* in H, so the LM head is free to rescale activations without paying a Dirichlet penalty for sheer norm growth. This normalized training loss differs from the unnormalized energy used in the theoretical analysis and from the permutation-null diagnostic ratio; the unnormalized variant is available in code but never used for training runs reported in this paper.

Per-scene bandwidth. We set τ *per scene* to half the median pairwise distance among that scene's objects: $\tau^{(b)} = \frac{1}{2}\,\mathrm{median}_{i<j}\|x_i^{(b)} - x_j^{(b)}\|$. Per-scene scaling makes the kernel locally adaptive — scenes with tightly clustered objects use a smaller τ, scenes with spread-out objects use a larger one — so the same λ_{dir} has comparable effect across the batch. The half-median heuristic puts the kernel's sweet spot near the typical nearest-neighbor scale, which is what Theorem 2 requires for the Belkin–Niyogi limit to apply.

Forward-hook capture of H_ℓ. The per-token residual stream is captured by registering a forward hook on the chosen transformer block (`ResidualStreamHook` in `scripts/dirichlet_loss.py`). For the multi-layer (*residMulti*) variant, 5 hooks are registered simultaneously across the cognitive-map band, and their per-layer Dirichlet ratios are averaged with equal weight. The pooling step that translates the per-token (B, T, d) tensor into the per-object (B, n, d) tensor uses the same mask-driven, coverage-weighted pool defined in § 4, with object indices derived from the tokenizer's name spans in the prompt—so the loss only sees activations corresponding to objects the prompt actually names.

Residualization projector $P_\perp$. The projector that residualizes against the nuisance subspace is $P_\perp = I_d - WW^\top$, where $W \in \mathbb{R}^{d\times k}$ has orthonormal columns spanning the color + shape directions. We fit W once per backbone, prior to training, using the following pipeline (`scripts/build_residualization_basis.py`):

1. Extract per-object residual-stream activations $H \in \mathbb{R}^{N\times d}$ from the *base* pretrained model (no LoRA, no Dirichlet) at the cognitive-map layer over the full 1,000-scene Free6DoF probing corpus.
2. Fit two multinomial logistic-regression probes on standardized H: one for the 8-way color label and one for the 3-way shape label, both with ℓ_2 regularization $C = 0.1$ and L-BFGS to convergence (max 2,000 iterations).
3. Recover the discriminating directions in the un-standardized H-space by composing each probe's weight matrix with the inverse of the per-feature standard deviations: $\widetilde{W}_{\mathrm{color}} = W^{\mathrm{LR}}_{\mathrm{color}}\,\mathrm{diag}(1/\sigma_j)$, and analogously for shape. Each row is rescaled to unit norm.
4. Stack into $\widetilde{W} = [\widetilde{W}_{\mathrm{color}}\,;\,\widetilde{W}_{\mathrm{shape}}] \in \mathbb{R}^{(8+3)\times d}$ and orthonormalize via thin QR: $\widetilde{W}^\top = QR$, taking $W \leftarrow Q$. Drop columns with diagonal $|R_{kk}|$ below 10^{-6} of the maximum (rank-deficient class directions). The resulting W has orthonormal columns and effective rank requiring verification against the implementation (the multinomial is rank-deficient by one per categorical, so $\le 7 + 2 = 9$ free directions).
5. Save W to disk and freeze it for the entire fine-tuning run; $P_\perp = I_d - WW^\top$ is then a single $d \times d$ matrix multiply per forward pass on the captured object-token activations.

Algorithm summary. The end-to-end per-step computation is:

1. *Forward pass*, capturing $H_\ell \in \mathbb{R}^{B\times T\times d}$ at each Dirichlet-target layer via forward hooks.
2. *Object pooling*: $H \leftarrow \mathrm{pool}(H_\ell, \mathrm{name_spans}) \in \mathbb{R}^{B\times n\times d}$.
3. *Residualize*: $\widetilde{H} \leftarrow H\,P_\perp = H - (HW)W^\top$.
4. *Per-scene bandwidth*: $\tau^{(b)} \leftarrow \frac{1}{2}\,\mathrm{median}_{i<j}\|x_i^{(b)} - x_j^{(b)}\|$.
5. *Kernel*: $W_{ij}^{(b)} \leftarrow \exp(-\|x_i^{(b)} - x_j^{(b)}\|^2/(2\tau^{(b)2})) \cdot \mathbb{1}[i \neq j]$.

6. *Pairwise distances*: $D_{ij}^{(b)} \leftarrow \|\widetilde{h}_i^{(b)} - \widetilde{h}_j^{(b)}\|^2$ via `torch.cdist`.
7. *Mask invalid pairs* via outer product of validity masks.
8. *Dirichlet ratio*: per-scene $r^{(b)} \leftarrow \sum_{ij} W_{ij}^{(b)} D_{ij}^{(b)} / (\sum_{ij} D_{ij}^{(b)} + \varepsilon)$, then average over b.
9. *Total loss*: $\mathcal{L}_{\text{total}} \leftarrow \mathcal{L}_{\text{CE}} + \lambda_{\text{dir}} \cdot \mathcal{R}_X$, backward through both the Dirichlet term (which feeds gradients to LoRA via H) and the cross-entropy.

For the residMulti variant, steps 1–8 run independently at 5 layers and step 9 averages the resulting ratios.

I.3 Manifold Learning and Latent Regularization

Our methodology is heavily grounded in the mathematical foundations of spectral graph theory and manifold learning. Foundational works on Laplacian eigenmaps [49, 52] and diffusion maps established that the principal eigenvectors of a graph Laplacian yield an optimal low-dimensional continuous representation that preserves local neighborhood structures. To measure how smoothly a representation varies over a given topology, researchers compute the Dirichlet energy [50], which acts as a penalty against high-frequency geometric variations. While spectral penalties and Dirichlet-energy regularization [53, 54] have been widely used in classical graph representation learning [55] and semi-supervised learning [56, 57], their application as a targeted regularizer for the hidden states of large transformer models is largely unexplored. By formally connecting the VLM residual stream to Laplacian eigenmaps, we introduce a mathematically principled way to isolate and regularize the internal geometric structure of language models.

I.4 Representational similarity (RSA)

Build two $m \times m$ symmetric similarity matrices, one from activations and one from 3D coordinates, and rank-correlate their off-diagonal entries:

$$S_{ij}^{\text{act}} = \frac{\langle \tilde{h}_\ell^{(s,o_i)}, \tilde{h}_\ell^{(s,o_j)} \rangle}{\|\tilde{h}_\ell^{(s,o_i)}\| \, \|\tilde{h}_\ell^{(s,o_j)}\|}, \quad S_{ij}^{\text{topo}} = \kappa_\tau(x_i, x_j) = \exp\Big(-\frac{\|x_i - x_j\|^2}{2\tau^2}\Big), \tag{18}$$

$$\text{RSA}^{(s)} = \text{Spearman}\big(\, \text{vec}_{\text{off}}(S^{\text{act}}),\ \text{vec}_{\text{off}}(S^{\text{topo}}) \,\big). \tag{19}$$

τ is set per scene to half the median pairwise distance. Spearman (rank) correlation makes the metric invariant under arbitrary monotone rescalings of either similarity, removing the linearity assumption a Pearson RSA would inject.

J Additional evaluations and implementation clarifications

J.1 Residualization and layer selection

Table 5 in Section 5.5 reports the residualization and layer-placement ablations on InternVL3-8B. These support residualization and middle-layer adaptation in this setting, without implying that the extracted representation is purely spatial. Table 6 gives the layer-count sweep: five layers outperform one, three, and nine layers among the tested configurations.

Number of tuned layers	VSI-Bench accuracy (%)
1	38.7
3	41.5
5	**42.4**
9	41.8

Table 6: Layer-count ablation on InternVL3-8B.

J.2 Non-spatial performance

A preliminary InternVL3-8B evaluation on MMMU [51] gives 61.2 for the pretrained model and 60.5 after latent shaping (Table 7). Naive LoRA, 2D regularization, and text-domain cognitive-map

prediction also show small decreases. These point estimates suggest that the decrease is not specific to the Dirichlet objective, but do not establish statistical equivalence or identify its cause. Broader non-spatial evaluation remains necessary.

InternVL3-8B configuration	MMMU score
Pretrained, no fine-tuning	61.2
Naive LoRA fine-tuning	60.7
2D regularization	59.9
Text-domain cognitive-map prediction	60.4
Latent shaping (ours)	60.5

Table 7: Preliminary general multimodal reasoning evaluation.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims regarding the representational imbalance, the theoretical connection to Laplacian eigenmaps, and the empirical gains of Dirichlet-energy regularization are clearly stated in the abstract and Introduction, and directly supported by Sections 3, 4, and 5.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of our work, including the reliance on synthetic scenes for spatial probing and the task-specific scope of our empirical gains, in appendix I.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Formal statements of all theorems (Theorems 1 through 5), including their assumptions, are provided in Section 5. The complete mathematical proofs are detailed in the supplementary material (Appendices B and C).

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We describe the dataset generation pipeline (Section 3), evaluation metrics (Section 4), and provide a detailed overview of the training procedure, baseline configurations, and hyperparameters (Section 5.3 and Appendix E).

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Anonymized code, including the scripts for synthetic data generation, spatial feature extraction, and Dirichlet-regularized fine-tuning, is provided in the supplementary material along with instructions.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Key details such as the number of tuning steps (500), LoRA configuration, penalty weighting (λ), and test splits are detailed in Section 5.3, Section 5.5, and Appendix E.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report the variance bands for our probing subspace gaps in Appendix A, and we strictly average all fine-tuning performance results over $n = 4$ random seeds as noted in Section 5.3.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We explicitly note the efficiency of our approach (e.g., requiring only 500 fine-tuning steps, roughly corresponding to 30 GPU-hours per model run) in Section 5.4.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research complies entirely with the NeurIPS Code of Ethics. We utilize standard open-weights academic models and benign benchmarks without introducing novel ethical risks.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [No]

Justification: Our paper focuses on the foundational mechanistic interpretability and representational geometry of existing VLMs. Improving spatial cognition in VLMs does not introduce direct negative societal impacts that warrant specific mitigation beyond general AI safety considerations.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release any high-risk models or scraped web datasets that require specialized safeguards.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We use publicly available models (Qwen2.5-VL, InternVL3) and benchmark datasets (VSI-Bench, MindCube), properly citing their original sources in the text.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The synthetic probing datasets and generators introduced in this work are documented in Appendix D, and the generation scripts are included in our supplementary code release.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The research does not involve any crowdsourcing or human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The research does not involve human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: While our object of study involves Vision-Language Models (VLMs), we do not use LLMs as an active, non-standard component to generate the core methodology, mathematical proofs, or experimental setup.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.