GraSP-VL: Length as a Semantic Granularity Interface for Vision-Language Representations

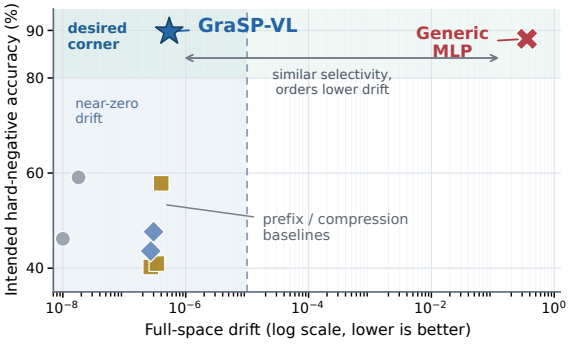
Anonymous ACL submission

Abstract

Frozen vision-language embeddings contain signals at multiple semantic resolutions, from object identity to attributes, relations, and full-caption meaning, but they expose these signals through a fixed-length vector interface. We study whether embedding length can be turned into a controllable semantic access interface. We propose **GraSP-VL**, which learns a shared near-orthogonal prefix transform over frozen VLM embeddings. GraSP-VL instantiates a **Semantic Matryoshka** interface: short prefixes are assigned coarse semantic roles, while longer prefixes progressively expose finer language-grounded distinctions. Because the transform is shared across image and text embeddings and preserves full-dimensional geometry, prefix behavior changes without rewriting the original VLM space. On a 20,147-example COCO/Flickr30K annotation pool, GraSP-VL reaches a staircase score of 53.01 and hard-negative selectivity of 89.76, while keeping full-space drift below 10^{-6}. It also transfers to SugarCrepe-clean with 86.03 object accuracy and 11.96 mean external emergence, and preserves full-dimensional zero-shot CIFAR-100 accuracy. These results show that frozen VLM embeddings can be reorganized into a truncatable semantic prefix interface rather than merely compressed. Code is available anonymously at `https://anonymous.4open.science/r/Grasp-VL-00E5`.

1 Introduction

Vision-language models (VLMs) such as CLIP make one embedding space support caption retrieval, zero-shot recognition, and prompts at very different semantic resolutions (Radford et al., 2021; Jia et al., 2021; Zhai et al., 2022; Cherti et al., 2023). This is convenient, but it hides an interface mismatch: sometimes a query only asks for an entity such as *dog*; sometimes it asks for an attributed phrase such as *brown dog*; sometimes the crucial information is a relation or event such as *dog running through snow*. A monolithic embedding gives all of these questions the same access pattern. Our premise is that frozen VLM outputs contain multi-granularity semantic signals; the missing piece is a controllable way to access them.



Teaser. GraSP-VL reaches the desired corner: high semantic selectivity with near-zero full-space drift.

The usual way to make embeddings shorter is to truncate, project, or compress them. These operations may preserve performance, but they do not tell us what the shorter vector is supposed to mean. The first small fraction of a frozen CLIP embedding is not guaranteed to be object-level; PCA may keep strong variance directions, but it does not know that a prefix should represent entities before attributes or relations. In this paper we study a different interface question: can representation length itself indicate semantic granularity?

This question is close to, but distinct from, Matryoshka-style representation learning (Kusupati et al., 2022; Yoon et al., 2024; Zhang et al., 2025), which makes multiple prefix lengths useful under different storage or latency budgets. Our target is not just usefulness under truncation: a short prefix should be coarse on purpose, and a longer prefix should add distinctions that the shorter one ignores. Put differently, length should be a semantic lens, not only a compression budget.

We propose **GraSP-VL** (**Gra**nularity-**S**elective

Prefixes for **V**ision-**L**anguage representations), a method for learning this access interface over frozen VLM embeddings. We use **Semantic Matryoshka** to denote the resulting prefix-interface principle: shorter prefixes are intentionally coarse, while longer prefixes reveal progressively finer distinctions under an explicit contract. GraSP-VL keeps the VLM frozen and learns a lightweight transform that is shared by image and text embeddings. The key constraint is that the full-dimensional geometry is preserved, while prefix geometry changes. This lets us ask a clean question: can we reorganize access to information already present in the frozen embedding so that short prefixes match coarse language and longer prefixes become increasingly compositional under an explicit interface contract? We do not claim that the chosen dimensional breakpoints are natural semantic constants or that the original coordinate order contains a discoverable semantic hierarchy. Rather, we test whether frozen VLM signals can be reorganized into a controllable prefix interface without rewriting the full embedding space.

Our evaluation is built as a falsification chain. If frozen VLMs lack multi-granularity signals, there is nothing to expose; if direct truncation, PCA, or random rotations already produce the desired order, no learned interface is needed; if Matryoshka-style training or a generic adapter gives the same behavior, the effect is ordinary compression or adaptation rather than semantic control.

Our contributions are:

- We propose **GraSP-VL**, a shared prefix transform for frozen VLM embeddings that preserves full-space similarity while changing which semantics are exposed by short prefixes.
- We formulate **Semantic Matryoshka** as the access-interface principle implemented by GraSP-VL: embedding length is assigned semantic granularity by an explicit contract rather than treated as a discovered natural boundary.
- We introduce a granularity-selective objective that combines prefix capability, cumulative retention, hard-negative discrimination, early-prefix invariance, and full-space preservation.
- We design a diagnostic protocol for prefix organization that separates semantic selectivity from raw prefix truncation, PCA compression, Matryoshka-style preservation, generic adaptation, and single-embedding compositional tuning.

2 Related Work

Nested representations. Nested representations make one embedding usable at multiple dimensionalities. Matryoshka Representation Learning trains prefixes of the same vector to remain predictive (Kusupati et al., 2022), and Matryoshka-Adaptor adapts pretrained embeddings for smaller usable dimensions (Yoon et al., 2024). Retrieval-specific work makes the compression motivation explicit: MRL-AdANNS uses Matryoshka representations for adaptive web-scale semantic search (Rege, 2023), and SMEC re-evaluates MRL for retrieval embedding compression with sequential training, adaptive dimension selection, and cross-batch memory (Zhang et al., 2025). Recent multimodal nested systems connect this idea to token-set or multi-vector retrieval (Li et al., 2024b; Cai et al., 2024; Xiao et al., 2025). These methods target prefix usability and adaptive retrieval budgets; GraSP-VL instead assigns different semantic responsibilities to different prefixes.

VLM adaptation. Many methods adapt frozen or partially frozen VLMs with a small number of trainable parameters. Prompt-learning methods such as CoOp, CoCoOp, and MaPLe tune textual or multi-modal prompts (Zhou et al., 2022b,a; Khattak et al., 2023). Feature-side methods such as CLIP-Adapter and Tip-Adapter modify or cache CLIP features for downstream recognition (Gao et al., 2021; Zhang et al., 2022), while adapters and low-rank adaptation provide general parameter-efficient machinery (Houlsby et al., 2019; Hu et al., 2022). GraSP-VL is also lightweight, but it reorganizes prefix access while preserving full-space behavior rather than adapting the VLM to a downstream classifier or domain.

Compositional VLMs. Standard retrieval scores often hide whether a VLM understands attributes, relations, or word order. SugarCrepe probes object, attribute, relation, and hard caption distractors outside standard retrieval leaderboards (Hsieh et al., 2023). Recent tuning methods such as NegCLIP, CE-CLIP, FSC-CLIP, CLIP-CAE, TripletCLIP, CLoVe, AHNPL, and DeGLA improve compositional sensitivity

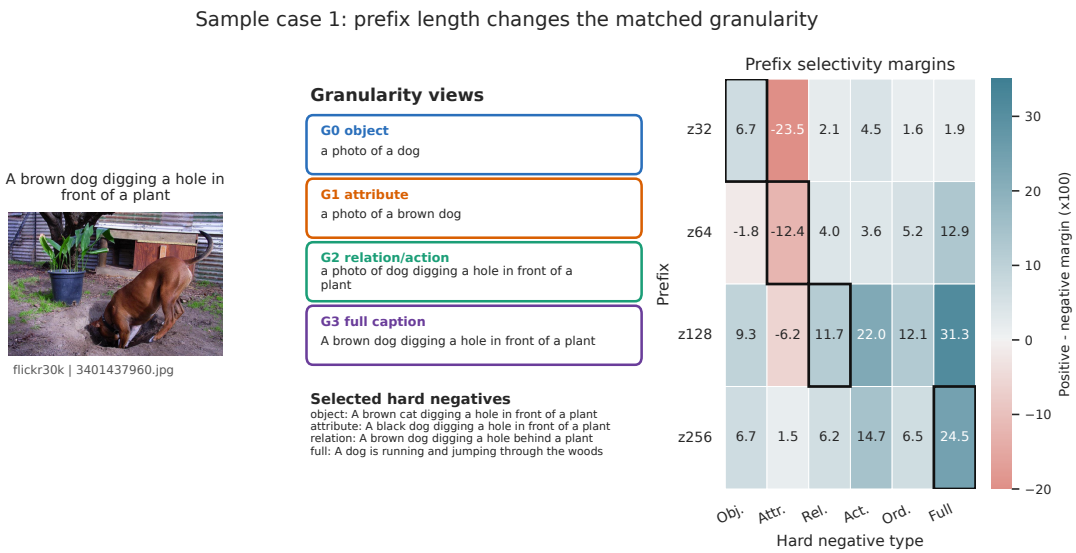


Figure 1: Sample case. Longer prefixes progressively expose attribute, relation/action, and full-caption distinctions.

through hard negatives, preservation-aware training, attribution enhancement, model patching, or global-local alignment (Yuksekgonul et al., 2023; Zhang et al., 2024; Oh et al., 2024; Li et al., 2024a; Patel et al., 2024; Castro et al., 2024; Huang et al., 2025; Hu et al., 2025). Fine-grained retrieval models such as SCAN and FILIP use region-word or token-level interactions rather than a single truncatable vector (Lee et al., 2018; Yao et al., 2022); FG-CLIP and FineLIP strengthen CLIP-style full embeddings with fine-grained data (Xie et al., 2025; Asokan et al., 2025). These methods improve compositionality in one representation, while we ask where such sensitivity should appear across prefixes.

Hierarchical semantics. Hierarchical semantics have long been central to language and vision. WordNet organizes lexical concepts taxonomically (Miller, 1995), while ImageNet and Visual Genome provide visual categories and dense semantic relations (Deng et al., 2009; Krishna et al., 2017). Representation learning has also used order embeddings, Poincaré embeddings, and hyperbolic image-text spaces to model hierarchy and asymmetric relations (Vendrov et al., 2016; Nickel and Kiela, 2017; Desai et al., 2023). Recent VLM analyses further show that CLIP-style embeddings encode fine-grained but uneven object and order information (Abbasi et al., 2025), while Insight extracts interpretable, spatially grounded semantic hierarchies from vision-language encoders using hierarchical concept representations (Wittenmayer et al., 2026). GraSP-VL is complementary: it does not extract a taxonomy, but learns a Semantic Matryoshka prefix interface in a frozen Euclidean VLM space.

3 Method

3.1 Problem Setup

Let E_I and E_T be frozen image and text encoders from a pretrained VLM. For an image x_i and a text string t, we denote their normalized embeddings as

$$e_i^x = E_I(x_i),$$
$$e(t) = E_T(t),$$
$$e_i^x, e(t) \in \mathbb{R}^D.$$

Each training pair has language views t_i^0–t_i^3 for object, object+attribute, relation/action/event, and full caption. These views can come from dataset annotations, parsers, or filtered LLM annotations; all training happens on top of frozen embeddings.

Our target prefixes are $\mathcal{K} = \{D/16, D/8, D/4, D/2, D\}$, with intended roles

$$\begin{aligned} D/16 &\to t^0, \\ D/8 &\to t^1, \\ D/4 &\to t^2, \\ D/2 &\to t^3, \\ D &\to \text{original full-space behavior.} \end{aligned}$$

For a 512-dimensional CLIP/OpenCLIP backbone, these ratios instantiate as

$\{32, 64, 128, 256, 512\}$. The full-dimensional role is important: prefix behavior should improve without destroying the frozen VLM geometry. Figure 2 summarizes the full pipeline and the contrast to direct truncation or generic adapters.

3.2 Shared Prefix Transform

GraSP-VL learns one transform $F_\theta : \mathbb{R}^D \to \mathbb{R}^D$ and applies it to both modalities:

$$z_i^x = F_\theta(e_i^x), \qquad z_i^t = F_\theta(e(t_i)).$$

The shared transform prevents the model from learning separate task-specific image and text spaces.

Our main instantiation uses an orthogonal transform,

$$F_\theta(e) = Re, \qquad R^\top R = I.$$

We parameterize R with a Cayley transform. Let $B_\theta \in \mathbb{R}^{D\times D}$ be trainable and

$$A_\theta = B_\theta - B_\theta^\top.$$

The orthogonal matrix is

$$R = (I + A_\theta)^{-1}(I - A_\theta),$$

with the linear solve implemented directly during the forward pass. This enforces orthogonality without projected-gradient steps; the non-orthogonal ablation removes this parameterization. The transform has $D^2 + |\mathcal{K}|$ trainable parameters, or 262,149 for a 512-dimensional backbone, and can be applied once to cached embeddings. For full-dimensional embeddings, this preserves cosine similarity exactly:

$$\langle Re_a, Re_b\rangle = \langle e_a, e_b\rangle.$$

Thus the full-space VLM geometry is unchanged, while truncation can expose a different prefix ordering. GraSP-VL learns a new access order for the frozen space, not a new full embedding.

For a prefix length k, we write

$$\pi_k(z) = \frac{z_{1:k}}{\|z_{1:k}\|_2},$$

and score an image-text pair with a prefix-specific temperature:

$$s_k(x_i, t_j) = \tau_k^{-1}\langle \pi_k(z_i^x), \pi_k(z_j^t)\rangle.$$

Each prefix is normalized independently and has its own learned positive temperature τ_k.

3.3 Granularity-Selective Objective

GraSP-VL uses five coupled terms:

$$\begin{aligned}\mathcal{L} = \mathcal{L}_{\text{align}} + \lambda_{\text{ret}}\mathcal{L}_{\text{ret}} + \lambda_{\text{rank}}\mathcal{L}_{\text{rank}} \\ + \lambda_{\text{inv}}\mathcal{L}_{\text{inv}} + \lambda_{\text{pres}}\mathcal{L}_{\text{pres}} + \lambda_{\text{ortho}}\mathcal{L}_{\text{ortho}}.\end{aligned}$$

$\mathcal{L}_{\text{align}}$ is a symmetric InfoNCE loss that aligns each prefix to its assigned language view. $\mathcal{L}_{\text{ret}}$ aligns longer prefixes to coarser views with smaller weights, so that finer prefixes add information rather than replace earlier semantics. $\mathcal{L}_{\text{rank}}$ is a margin loss over style-matched typed negatives, activated from the earliest prefix

$$\begin{aligned}\kappa(\text{obj.}) = D/16, \quad \kappa(\text{attr.}) = D/8, \\ \kappa(\text{rel./act./ord.}) = D/4, \\ \kappa(\text{full}) = D/2.\end{aligned}$$

For a positive–negative pair (p_i^r, n_i^r) of type r, the ranking term applies to $k \geq \kappa(r)$ and encourages $s_k(x_i, p_i^r) > s_k(x_i, n_i^r)$. $\mathcal{L}_{\text{inv}}$ applies before $\kappa(r)$ and discourages premature separation of the same pair. This makes early prefixes selective rather than merely weak: $D/16$ should detect the object without being forced to decide color, relation, or full-caption distractors. Finally, $\mathcal{L}_{\text{pres}}$ and $\mathcal{L}_{\text{ortho}}$ monitor non-orthogonal ablations; with the Cayley transform, full-space cosine preservation follows analytically.

The boundary function $\kappa(r)$ is an interface contract, not a claim that semantics has universal dimensional breakpoints. The experiments therefore test whether this contract can be realized while preserving the frozen full space, beating coordinate/permutation baselines, and transferring beyond the exact annotation format. At inference time, the user selects a prefix according to the required resolution: object, attribute, relation/action/order, or full-caption matching.

Cost and caching. GraSP-VL adds $D^2 + |\mathcal{K}|$ trainable parameters: a dense Cayley parameter $B_\theta \in \mathbb{R}^{D\times D}$ and one temperature per prefix. The transform is a post-encoder operation and can be applied offline to cached gallery embeddings at $O(ND^2)$ per gallery build or transform update. After caching, retrieval against a selected prefix costs $O(k)$ per candidate, while a new query pays one $O(D^2)$ transform, about one million multiply-adds at $D = 1024$. Appendix Table 12 gives 10M-gallery storage estimates and discusses structured orthogonal variants.

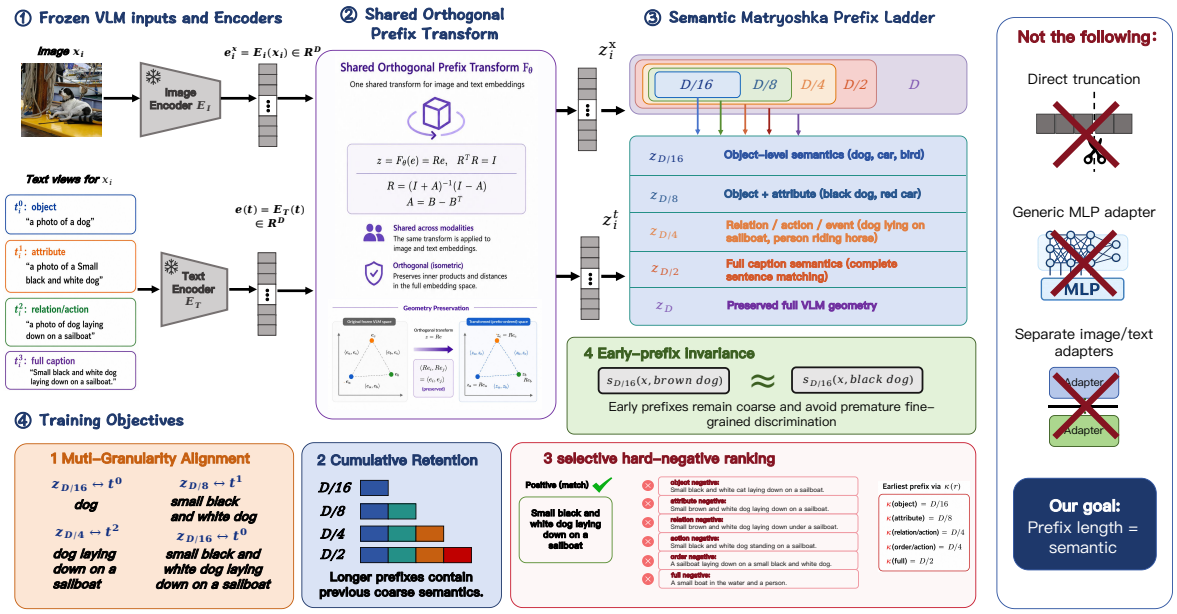


Figure 2: GraSP-VL overview. A shared near-isometric transform reorders frozen image/text embeddings into a semantic prefix ladder.

Figure 3 visualizes the effect of this shared orthogonal rotation. Before GraSP-VL, direct prefixes expose a mixture of semantic distinctions without the intended coarse-to-fine contract. After rotation, object, attribute, and relation/action/order selectivity move toward their assigned prefixes while the full-dimensional row remains the frozen VLM geometry.

4 Experiments

Our experiments are diagnostic rather than a standard retrieval leaderboard. They are organized as a falsification chain: we first document the weak supervision, then ask whether frozen VLMs already contain multi-granularity signals, whether raw coordinates expose them, whether GraSP-VL induces the intended staircase, and whether the effect can be explained by compression, generic adaptation, prompt templates, or a fragile training schedule.

4.1 Experimental Setup

Data and annotation. We build a 20,147-example image-caption pool from COCO and Flickr30K (Lin et al., 2014; Young et al., 2014): 16,028 examples from COCO and 4,119 from Flickr30K. The split is 16,785/1,861/1,501 for train/validation/test. For each example, DeepSeek-V4-Flash produces four text views: object (G_0), object-attribute (G_1), relation/event (G_2), and the original full caption (G_3). It also generates typed hard negatives for object, attribute, relation, action, order, and full-caption discrimination. Because these annotations are weak supervision, we normalize prompt style and discard malformed or structurally inconsistent generations. Appendix Table 6 reports the deterministic validation pass: 20,147 of 22,710 attempted annotations are accepted, and accepted rows are required to pass structural, grounding, and negative-consistency checks before training. We additionally manually audit 500 randomly sampled accepted examples; Appendix Table 7 reports high validity for object, attribute, relation/event views and typed negatives, with relation/event annotations being the most ambiguous. This is not meant to replace human semantic auditing; rather, it makes the filtering rules explicit before the model sees the data. The remaining likely failure modes are hallucinated attributes, underspecified entities, implausible relations, and negatives that are grammatical but still compatible with the image.

Baselines and metrics. Coordinate baselines use direct prefixes, PCA prefixes fitted on the training split, and random orthogonal rotations. Compression baselines follow MRL, Matryoshka-Adaptor, and SMEC-style objectives (Kusupati et al., 2022; Yoon et al., 2024; Zhang et al., 2025). We also compare with unconstrained adapters, learned coordinate permutations, and a backbone sweep over OpenAI CLIP, OpenCLIP, EVA-CLIP,

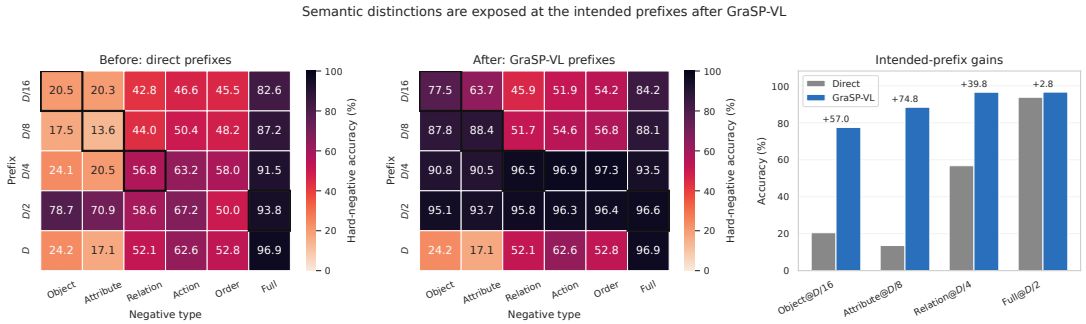


Figure 3: Effect of the shared orthogonal rotation: prefix selectivity moves toward the intended semantic ladder while full-space geometry is preserved.

and SigLIP (Radford et al., 2021; Sun et al., 2023; Zhai et al., 2023). These are training-matched prefix-interface baselines: they operate on the same frozen-feature protocol and can therefore be compared using staircase score, emergence, and full-space drift. To avoid limiting the comparison to adapters, we separately include external fine-grained and compositional VLMs such as FG-CLIP, NegCLIP, TripletCLIP, CLIP-CAE, CE-CLIP, CLoVe, and DeGLA under the same held-out diagnostic protocol when released checkpoints can be encoded. Architectural retrieval models based on region-word or token-level late interaction, such as SCAN and FILIP, are important but not directly prefix-compatible: their scores are computed from local interactions rather than a single global vector with truncatable prefixes, so full-space drift and prefix length are not defined in the same way. We therefore treat them as architectural competitors for retrieval, not as substitutes for the prefix-interface baselines. All encoders are frozen; GraSP-VL trains only the shared Cayley orthogonal prefix transform and prefix temperatures.

For diagnostic retrieval, the 1,501 held-out test images query the full 20,147-example text pool; train and validation captions are distractors, not credited positives. This makes R@1 a large-candidate prefix-organization diagnostic rather than a standard retrieval leaderboard score. Appendix Table 14 shows that a test-only pool raises absolute R@1 but preserves the same ordering. Hard-negative selectivity and out-of-objective transfer probes are therefore the primary evidence for prefix granularity:

$$\mathrm{Sel}(k, r) = \Pr_{i \in \mathcal{Q}_r} \left[s_k(x_i, t_i^+) > s_k(x_i, n_i^r) \right],$$

where r is the negative type. We summarize the interface with

$$\mathrm{Stair} = \frac{1}{2} \left(\mathrm{RetAvg} + \mathrm{HardAvg} \right).$$

We also report emergence gap, pre-κ leakage for schedule analyses, and full drift, the maximum absolute change in full-dimensional image–text similarities relative to the frozen VLM.

4.2 Frozen Signals and Coordinate Baselines

Table 2 first checks the premise. The frozen full embedding contains multi-granularity signal, but direct, PCA, and random-coordinate prefixes do not expose a stable semantic ladder. Low object R@1 is expected because many images share the same short object phrase.

4.3 Main Results: Prefix Staircase

Table 1 is the central diagnostic result. Hard negatives show the intended order: object at $D/16$, attribute at $D/8$, relation at $D/4$, and full-caption discrimination at $D/2$. Retrieval follows the same trend, while object/attribute R@1 remain modest because short phrases are shared by many candidates. The full D row is only a preservation diagnostic: under an orthogonal transform, full-space cosine scores reproduce the frozen VLM. Thus low full-D object or attribute selectivity should not be read as drift; it reflects the frozen full-space decision surface. GraSP-VL changes what prefixes expose, not what the full embedding knows.

4.4 Method Comparison

Table 2 rules out compression and generic adaptation. Direct, PCA, MRL-style, Matryoshka-style, and SMEC-style baselines either lack the staircase or trade it for caption preservation. This is the key distinction: making every prefix useful is not the same as assigning semantic responsibility to different prefix lengths. The generic MLP reaches high selectivity but rewrites the embedding space, while learned permutations preserve geometry but remain far below Cayley GraSP-VL.

Prefix	Obj. R	Attr. R	Rel. R	Cap. R	Obj. Neg.	Attr. Neg.	Rel. Neg.	Full Neg.
$D/16$ (32)	**9.66**	2.07	1.47	2.93	**77.48**	63.69	45.90	84.21
$D/8$ (64)	6.93	**2.47**	1.93	4.80	87.81	**88.41**	51.70	88.07
$D/4$ (128)	6.93	2.86	**16.32**	20.52	90.81	90.47	**96.54**	93.54
$D/2$ (256)	8.19	4.46	16.19	**34.91**	95.07	93.74	95.80	**96.60**
D (512)	6.20	5.53	35.11	46.50	24.18	17.06	52.10	96.87

Table 1: GraSP-VL prefix staircase. Bold marks the intended prefix-view or prefix-negative pairing; full D is a preservation row.

The effect therefore requires dense near-isometric mixing rather than simple dimension sorting.

Is the transform just a permutation? No: learned permutation baselines are much weaker, and the best permutation explains only 25.23% of the learned Cayley energy. Figure 4 makes this compactly visible. Appendix Tables 15 and 16 unpack staircase and emergence scores.

4.5 Transfer and Backbone Generalization

Table 3 summarizes two robustness checks. SugarCrepe-clean is external to our annotation pipeline and is never used for training or validation; it provides partial transfer evidence rather than a strong generalization proof, with high object accuracy already at $D/16$ and relation/binding contrasts improving by $D/4$. The same ratio-prefix formulation also transfers across CLIP-style backbones with different dimensions while keeping near-zero full drift. Appendix Tables 18, 19, 20, 21, 22, and 23 give the detailed transfer, ablation, schedule, and $\kappa(r)$ checks.

Appendix A contains baseline implementation details, prompt-template, candidate-pool, transform, ablation, and schedule diagnostics.

5 Discussion

GraSP-VL is interface engineering over a frozen VLM, not a proof that language has universal breakpoints at $D/16$, $D/8$, or $D/4$. The evidence is about contract realization: the full space is preserved, raw coordinates and permutations do not expose the interface, and external probes show partial transfer beyond matched LLM supervision. The method can only expose information already present in the frozen embedding; weak base relation/action/order sensitivity cannot be created from nothing. LLM views remain weak supervision after filtering and audit, so the results should be read as diagnostic evidence for a controllable interface rather than as a claim that every generated view is semantically perfect. Dense Cayley transforms are practical for the studied backbones, but structured orthogonal variants are the natural next step for very large deployments.

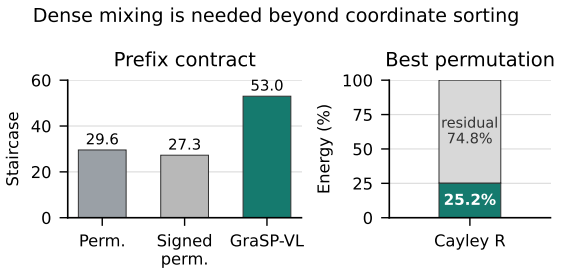


Figure 4: Permutation summary. GraSP-VL improves the prefix staircase over learned coordinate permutations, and the learned Cayley rotation is not dominated by a single hard permutation.

This framing is important for interpreting $\kappa(r)$. The mapping from negative type to earliest prefix is a user-facing interface contract: it specifies where the system should begin exposing a distinction, not where the VLM "naturally" stores that distinction. The shifted-κ experiments therefore test controllability rather than search for a unique semantic boundary: frozen VLM embeddings contain enough multi-granularity signal to be reorganized into a chosen prefix interface while preserving the full geometry.

The main remaining risk is supervision mismatch. Typed negatives and semantic views are useful because they isolate object, attribute, relation, action, order, and full-caption decisions, but they are still generated from a finite annotation pipeline. For this reason, we do not treat the in-pipeline hard-negative staircase as sufficient evidence by itself. The prompt-template, human-caption, and SugarCrepe-clean probes test whether the interface survives changes in wording, data source, and contrast construction, but future work should still evaluate unseen semantic types and stronger domain shifts.

This also explains why no single metric is sufficient. High hard-negative accuracy alone can be obtained by an unconstrained adapter that changes

Frozen and prefix-interface baselines						
Method	Stair.	Emerg.	Cap. R@1	Hard Avg.	Drift ↓	Params
Frozen full	–	–	36.91	74.05	0.0e+00	0
Direct prefix	26.48	7.84	14.46	46.17	0.0e+00	0
PCA prefix	39.24	-2.51	46.04	59.09	0.0e+00	0
Random rotation	37.34	–	23.31	66.57	0.0e+00	0
MRL-style	29.77	0.94	41.44	40.24	4.2e-07	262,149
Matryoshka-style adaptor	29.78	1.08	38.97	40.96	4.2e-07	294,918
SMEC-style compression	39.08	-3.24	41.11	57.84	4.0e-07	262,149
Generic MLP adapter	50.88	**37.50**	24.72	88.24	3.6e-01	1,052,165
Learned permutation	29.56	2.01	31.91	47.65	3.0e-07	262,149
Learned signed permutation	27.27	-1.97	29.25	43.60	2.7e-07	262,661
GraSP-VL	**53.01**	33.17	34.91	**89.76**	5.4e-07	262,149

External single-vector VLM diagnostics				
Model	Attr. Neg.	Rel. Neg.	Transfer R@1	Drift ↓
Frozen VLM	13.59	56.76	14.46	0.0e+00
FG-CLIP	20.65	54.03	36.84	–
NegCLIP	36.38	54.36	0.93	–
TripletCLIP	36.38	67.55	0.87	–
CLIP-CAE-AB	24.78	62.43	21.25	–
CE-CLIP	19.85	44.30	22.05	–
CLoVe	28.31	73.35	7.13	–
DeGLA	31.18	79.21	21.05	–
GraSP-VL	**88.41**	**96.54**	34.91	5.4e-07

Table 2: Method comparison and external encoder diagnostics. Frozen, direct, PCA, and random rows test whether the signal is already ordered; GraSP-VL is strongest among geometry-preserving prefix-interface methods.

SugarCrepe-clean external probe				
Method	Obj.	Attr. Δ	Rel. Δ	Mean
Direct prefix	65.70	-0.47	5.42	2.47
PCA prefix	82.93	10.27	6.56	8.42
Generic MLP	81.64	10.07	**15.82**	**12.94**
Learned perm.	71.54	0.95	4.75	2.85
GraSP-VL	**86.03**	**10.66**	15.66	11.96

Backbone generalization					
Backbone	Stair.	Emerg.	Hard	Cap.	Drift ↓
OpenCLIP B/16	53.01	33.17	89.76	34.91	5.4e-07
OpenAI B/32	46.77	28.80	83.60	29.63	5.1e-07
OpenAI B/16	50.27	31.20	86.98	32.53	5.7e-07
OpenCLIP L/14	56.32	35.40	92.70	38.19	5.2e-07
EVA-CLIP B/16	54.62	34.50	91.18	36.18	5.4e-07
SigLIP B/16	51.89	32.50	88.50	33.64	5.6e-07

Table 3: External transfer and backbone generalization. SugarCrepe-clean is not generated by our annotation pipeline; ratio prefixes remain effective across CLIP-style encoders.

the full embedding space; low leakage alone can mean that short prefixes are simply weak; and high caption retrieval alone can be achieved by compression-style objectives that make every prefix approximate the same full-caption task. We therefore interpret the results jointly: intended-prefix capability should be high, gains should emerge near the assigned boundary, and full-dimensional geometry should remain stable. The contribution is not a unique semantic schedule, but a way to realize a user-specified one without converting the frozen VLM into a new model.

The practical use case is also diagnostic rather than a replacement for all retrieval systems. A short prefix can serve as a coarse semantic filter when the user only needs entities, while longer prefixes can be used for attribute- or relation-sensitive matching and final caption-level reranking. Gallery embeddings can be transformed and cached once, after which shorter prefixes reduce retrieval memory and dot-product cost. At the same time, coarse prefixes may suppress task-relevant fine attributes, while fine prefixes may expose sensitive attributes already encoded by the base VLM.

6 Conclusion

We introduced GraSP-VL, a geometry-preserving method that turns frozen VLM length into a controllable semantic access interface. Across diagnostic retrieval, transfer, and preservation checks, GraSP-VL exposes coarse-to-fine prefixes while keeping full-dimensional similarities fixed. The result is interface control rather than natural-coordinate discovery: frozen VLMs contain multi-granularity signals, and a shared transform can make their access order explicit.

Limitations

GraSP-VL should be understood as a controllable interface over frozen VLM embeddings rather than evidence that any particular prefix

schedule constitutes a universal semantic partition. The boundary schedule is user-specified, and the method can only expose semantic distinctions already encoded by the underlying VLM. This design is intentional: our goal is to reorganize and make accessible existing frozen VLM semantics, not to inject new visual-linguistic knowledge into the backbone. Our supervision uses filtered LLM-generated views and typed negatives, so annotation artifacts, hallucinated attributes, underspecified relations, and prompt-style biases may remain despite deterministic filtering and human audit. External probes such as SugarCrepe-clean and human-caption transfer reduce this circularity concern but do not exhaust unseen semantic types, domains, or caption styles. Finally, dense orthogonal transforms can be cached for fixed galleries, but very large dynamic deployments may require structured orthogonal variants with lower application cost.

References

Reza Abbasi, Ali Nazari, Aminreza Sefid, Mohammadali Banayeeanzade, Mohammad Hossein Rohban, and Mahdieh Soleymani Baghshah. 2025. CLIP under the microscope: A fine-grained analysis of multi-object representation. *arXiv preprint arXiv:2502.19842*.

Mothilal Asokan, Kebin Wu, and Fatima Albreiki. 2025. FineLIP: Extending CLIP's reach via fine-grained alignment with longer text inputs. *arXiv preprint arXiv:2504.01916*.

Mu Cai, Jianwei Yang, Jianfeng Gao, and Yong Jae Lee. 2024. Matryoshka multimodal models. *arXiv preprint arXiv:2405.17430*.

Santiago Castro, Amir Ziai, Avneesh Saluja, Zhuoning Yuan, and Rada Mihalcea. 2024. CLoVe: Encoding compositional language in contrastive vision-language models. *arXiv preprint arXiv:2402.15021*.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. 2023. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Karan Desai, Maximilian Nickel, Tanmay Rajpurohit, Justin Johnson, and Shanmukha Ramakrishna Vedantam. 2023. Hyperbolic image-text representations. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 7694–7731. PMLR.

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. 2021. CLIP-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. 2023. SugarCrepe: Fixing hackable benchmarks for vision-language compositionality. *arXiv preprint arXiv:2306.14610*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Xiaoxing Hu, Kaicheng Yang, Jun Wang, Haoran Xu, Ziyong Feng, and Yupei Wang. 2025. Decoupled global-local alignment for improving compositional understanding. *arXiv preprint arXiv:2504.16801*.

Xin Huang, Ruibin Li, Tong Jia, Wei Zheng, and Ya Wang. 2025. Visual perturbation and adaptive hard negative contrastive learning for compositional reasoning in vision-language models. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 5435–5443.

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

Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. 2023. MaPLe: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19113–19122.

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

Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, and Ali Farhadi. 2022. Matryoshka representation learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 30233–30249.

Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision*, pages 201–216.

Wei Li, Zhen Huang, Xinmei Tian, Le Lu, Houqiang Li, Xu Shen, and Jieping Ye. 2024a. Interpretable composition attribution enhancement for visio-linguistic compositional understanding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14616–14632, Miami, Florida, USA. Association for Computational Linguistics.

Xianming Li, Zongxi Li, Jing Li, Haoran Xie, and Qing Li. 2024b. 2D matryoshka sentence embeddings. *arXiv preprint arXiv:2402.14776*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *Computer Vision – ECCV 2014*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer.

George A. Miller. 1995. WordNet: A lexical database for english. *Communications of the ACM*, 38(11):39–41.

Maximillian Nickel and Douwe Kiela. 2017. Poincaré embeddings for learning hierarchical representations. In *Advances in Neural Information Processing Systems*, volume 30.

Youngtaek Oh, Jae Won Cho, Dong-Jin Kim, In So Kweon, and Junmo Kim. 2024. Preserving multimodal capabilities of pre-trained VLMs for improving vision-linguistic compositionality. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19060–19076, Miami, Florida, USA. Association for Computational Linguistics.

Maitreya Patel, Abhiram Kusumba, Sheng Cheng, Changhoon Kim, Tejas Gokhale, Chitta Baral, and Yezhou Yang. 2024. TripletCLIP: Improving compositional reasoning of CLIP via synthetic vision-language negatives. In *Advances in Neural Information Processing Systems*, volume 37, pages 32731–32760.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Aniket Rege. 2023. MRL-AdANNS: Matryoshka representation learning for web-scale adaptive semantic search. Master's thesis, University of Washington.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. 2023. EVA-CLIP: Improved training techniques for CLIP at scale. *arXiv preprint arXiv:2303.15389*.

Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. 2016. Order-embeddings of images and language. In *International Conference on Learning Representations*.

Kai Wittenmayer, Sukrut Rao, Amin Parchami-Araghi, Bernt Schiele, and Jonas Fischer. 2026. CFM: Language-aligned concept foundation model for vision. *arXiv preprint arXiv:2601.13798*.

Zilin Xiao, Qi Ma, Mengting Gu, Chun-cheng Jason Chen, Xintao Chen, Vicente Ordonez, and Vijai Mohan. 2025. MetaEmbed: Scaling multimodal retrieval at test-time with flexible late interaction. *arXiv preprint arXiv:2509.18095*.

Chunyu Xie, Bin Wang, Fanjing Kong, Jincheng Li, Dawei Liang, Gengshen Zhang, Dawei Leng, and Yuhui Yin. 2025. FG-CLIP: Fine-grained visual and textual alignment. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 68777–68793. PMLR.

Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. 2022. FILIP: Fine-grained interactive language-image pre-training. In *International Conference on Learning Representations*.

Jinsung Yoon, Rajarishi Sinha, Sercan O Arik, and Tomas Pfister. 2024. Matryoshka-adaptor: Unsupervised and supervised tuning for smaller embedding dimensions. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10318–10336, Miami, Florida, USA. Association for Computational Linguistics.

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

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

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

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. 2022. LiT: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18123–18133.

Biao Zhang, Lixin Chen, Tong Liu, and Bo Zheng. 2025. SMEC: Rethinking matryoshka representation learning for retrieval embedding compression. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 26209–26222, Suzhou, China. Association for Computational Linguistics.

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

Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. 2022. Tip-adapter: Training-free adaption of CLIP for few-shot classification. In *Computer Vision – ECCV 2022*, volume 13695 of *Lecture Notes in Computer Science*, pages 493–510. Springer.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022a. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022b. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130:2337–2348.

Split	COCO	Flickr30K	Total
Train	13,845	2,940	16,785
Validation	1,537	324	1,861
Test	646	855	1,501
All accepted	16,028	4,119	20,147

Table 4: Accepted examples by dataset and split.

A Reproducibility and Additional Diagnostics

This appendix provides the supporting evidence behind the compact results in the main paper. We keep the material in one continuous appendix rather than splitting each diagnostic onto a separate page, because the checks are meant to be read as a connected audit trail: annotation quality first, then reproducibility and scaling, then diagnostic decompositions and transfer probes, and finally the exact metric definitions.

Dataset splits and protocol separation. Table 4 gives the exact accepted-example counts after filtering. The split is fixed before model selection: training optimizes the prefix transform, validation selects hyperparameters and checkpoints, and test rows are used only for the final reported diagnostics. COCO and Flickr30K are both present in every split, but the test split intentionally has a larger Flickr30K share than the train split, so the reported test behavior is not simply a COCO-only retrieval check.

We use the same frozen embedding cache for GraSP-VL and the prefix-interface baselines whenever possible. This keeps the comparison focused on how a method reorganizes access to the frozen VLM space rather than on encoder differences, preprocessing differences, or candidate-pool differences. Table 5 summarizes which parts of the pipeline can influence training and which are held out for diagnosis.

Annotation validation. Table 6 reports the deterministic validation rules used for the LLM-generated semantic views and typed negatives. These checks remove malformed rows and enforce grounding/copying constraints before training, but they should be read as structural validation rather than a guarantee that every accepted view is semantically perfect. For that reason, we pair the automatic checks with the manual audit in Table 7. The two tables serve different purposes: Table 6 verifies that the dataset obeys the required schema and copying constraints, while Table 7 estimates semantic validity on 500 randomly sampled accepted examples.

Algorithm 1 GraSP-VL training

Input: frozen embeddings, semantic views $\{G_m\}$, typed negatives $\{n^r\}$, prefixes $\mathcal{K}$, boundary map $\kappa(r)$
Initialize Cayley parameter B_θ near zero and prefix temperatures $\{\tau_k\}_{k\in\mathcal{K}}$
for each minibatch $\mathcal{B}$ **do**
  $A_\theta \leftarrow B_\theta - B_\theta^\top$
  $R \leftarrow (I + A_\theta)^{-1}(I - A_\theta)$
  Apply the shared transform to all image/text embeddings: $z \leftarrow Re$
  for each prefix $k \in \mathcal{K}$ **do**
    $\pi_k(z) \leftarrow z_{1:k}/\|z_{1:k}\|_2$
    $s_k(x,t) \leftarrow \tau_k^{-1}\langle \pi_k(z_x), \pi_k(z_t)\rangle$
  end for
  Build $\mathcal{L}_{\text{align}}$, $\mathcal{L}_{\text{ret}}$, $\mathcal{L}_{\text{rank}}$, $\mathcal{L}_{\text{inv}}$, and $\mathcal{L}_{\text{pres}}$
  Update B_θ and $\{\tau_k\}$ with the weighted objective defined in Section 3.3
end for
return best validation checkpoint subject to near-zero full-space drift

We complement deterministic checks with a manual audit of 500 accepted examples. The audit separates valid, ambiguous, and invalid cases because some relation/event descriptions are plausible but underspecified rather than plainly wrong. Table 7 shows that object and attribute views are the most reliable, while relation/event views and relation negatives remain the noisiest categories.

Reproducibility details. Table 8 collects implementation choices that affect exact reproduction. All hyperparameters are selected on the validation split and kept fixed for test evaluation. Type-specific margins, invariance tolerances, loss weights, retention weights, temperatures, and the negative curriculum are part of the method rather than incidental engineering choices; we therefore treat them as release-critical configuration fields. Algorithm 1 summarizes the training loop. These details are followed by Table 12, which reports the storage and transform cost implied by the same prefix design when moving from the 512-dimensional backbone used in most experiments to larger VLM embeddings.

An anonymized implementation and evaluation scripts are available at `https://anonymous.4open.science/r/Grasp-VL-00E5/README.md`. Accepted annotation files, split metadata, and configuration files will be released with the camera-ready version. For code release, the most important reproducibility artifacts are the accepted split files, the frozen embedding cache, the GraSP-VL checkpoint, and the scripts that build each reported table. Table 9 lists the relative locations used by the experiments in this paper.

Stage	Uses training annotations?	Purpose
Frozen VLM encoding	No	Cache image/text embeddings once; encoders are never updated.
GraSP-VL training	Yes, train split only	Learn the shared prefix transform and prefix temperatures.
Hyperparameter selection	Validation split only	Choose loss weights, margins, curriculum, and checkpoint subject to drift.
Main diagnostic test	No train rows as positives	Evaluate held-out test images; train/validation captions are only distractors in the large pool.
Prompt-template transfer	No new training	Replace evaluation wording to test prompt-style sensitivity.
Human-caption transfer	No generated views/typed negatives	Use held-out human captions and real image distractors.
SugarCrepe-clean	No GraSP training/validation images	Test external compositional contrasts outside the annotation pipeline.
CIFAR-100 zero-shot	No GraSP data	Check a common classification use case with no additional tuning.

Table 5: Protocol separation between training signals, model selection, and diagnostic probes.

Group	Check	Count	Denom.	Rate	Action
Generation	All attempted annotations	22,710	22,710	100.00	–
Generation	API or malformed failures	2,301	22,710	10.13	Discarded
Generation	Deterministic quality failures	262	22,710	1.15	Discarded
Generation	Final accepted annotations	20,147	22,710	88.71	Used
Generation	Accepted among parseable annotations	20,147	20,409	98.72	Used
Accepted-row audit	Entity is non-empty	20,147	20,147	100.00	Required
Accepted-row audit	Caption is copied exactly from input captions	20,147	20,147	100.00	Required
Accepted-row audit	G_3 equals selected caption	20,147	20,147	100.00	Required
Accepted-row audit	All four semantic views are present	20,147	20,147	100.00	Required
Accepted-row audit	Entity or surface form appears in selected caption	20,147	20,147	100.00	Required
Accepted-row audit	Event view contains entity or surface form	20,146	20,147	99.995	Required
Accepted-row audit	All six negative types are present	20,147	20,147	100.00	Required
Accepted-row audit	Every negative differs from selected caption	20,147	20,147	100.00	Required
Accepted-row audit	Full negative is copied from supplied distractors	20,147	20,147	100.00	Required
Accepted-row diagnostic	Attribute view differs from object view	13,017	20,147	64.61	Reported

Table 6: LLM annotation validation. We treat LLM outputs as weak supervision and apply deterministic structural checks before training or evaluation. The final row is diagnostic rather than a filtering rule: some captions do not contain an explicit attached attribute, so G_1 can legitimately equal G_0.

Type	Valid	Ambiguous	Invalid
Object view	98.8	1.1	0.1
Attribute view	97.5	1.7	0.8
Relation/event view	97.2	1.4	1.4
Object negative	97.6	0.2	1.2
Attribute negative	98.9	0.5	0.6
Relation negative	98.4	1.1	0.5
Full-caption negative	97.3	2.4	0.3

Table 7: Human audit of 500 randomly sampled LLM-generated semantic views and typed negatives. Values are percentages. Ambiguous means the annotation is plausible but underspecified or not uniquely grounded; invalid means it contradicts the image/caption or fails the intended perturbation.

The paths are not assumptions of the method; they are included to make the released repository easier to audit.

Table 10 reports the hardware and software environment used for the main OpenCLIP ViT-B/16 experiments. Feature extraction and GraSP-VL training are GPU workloads; most table-building scripts operate on cached embeddings and can be rerun on CPU, albeit more slowly.

Baseline implementation transparency. The local baselines are protocol-matched tests of alternative explanations, not full reimplementations of every original system. Table 11 specifies the transform and objective used for each row. MRL-style, Matryoshka-style, and SMEC-style rows keep the spirit of nested retrieval/compression, so they do not use the typed $\kappa(r)$ hard-negative and early-invariance contract. Adding those losses would make them GraSP-style variants rather than independent compression baselines. Conversely, the generic MLP and learned permutation baselines do use the full GraSP objective, so they test whether capacity or coordinate sorting alone explains the result.

The prefix interface is intended for retrieval settings where gallery embeddings can be transformed and cached offline. Table 12 reports storage and dense-transform costs for several embedding dimensions. The dense Cayley transform has modest parameter memory but quadratic application cost, motivating structured orthogonal variants for very large galleries. Table 13 gives an initial empirical check with butterfly-Givens transforms, which preserve full-space geometry while retaining most of the semantic selectivity at much lower application cost.

Transform structure and candidate pools. Table 14 then addresses the retrieval-pool concern. Using only held-out test captions raises absolute R@1, as expected, but the qualitative prefix or-

Component	Symbol / field	Setting reported for reproduction
Backbone	E_I, E_T	Frozen OpenCLIP ViT-B/16 unless otherwise specified
Prefix set	$\mathcal{K}$	$\{D/16, D/8, D/4, D/2, D\}$
Semantic boundary	$\kappa(r)$	object:$D/16$; attribute:$D/8$; relation/action/order:$D/4$; full:$D/2$
Transform	R	Shared image/text Cayley orthogonal transform with dense $D \times D$ parameter
Temperature	τ_k	One learned positive temperature per prefix; initialized identically across prefixes
Ranking margins	m_r	Type-specific constants fixed before test evaluation and shared by all runs of a given protocol
Invariance tolerances	ϵ_r	Type-specific tolerances for suppressing premature positive–negative separation
Retention weights	$\alpha_{k,\ell}$	Coarser-view retention weights for longer prefixes; lower than intended-view alignment weight
Loss weights	$\lambda_{\text{ret}}, \lambda_{\text{rank}}, \lambda_{\text{inv}}, \lambda_{\text{pres}}, \lambda_{\text{ortho}}$	Fixed validation-selected constants used for all test tables
Curriculum	warmup / negative schedule	Three warmup epochs before the default hard-negative curriculum
Model selection	checkpoint	Best validation staircase subject to near-zero full-space drift

Table 8: Implementation details needed for exact reproduction. The main text specifies the objective and parameterization; the run configuration must include the numeric values for the type-specific margins, tolerances, loss weights, and retention weights.

Artifact	Relative path	Used for
Accepted splits	`results/grasp_vl_llm_20k_current/splits/*.jsonl`	Train/validation/test partition and metadata.
Frozen feature cache	`results/frozen_vlm_granularity_probe_20k_current/*.pt`	Shared OpenCLIP image/text embeddings.
GraSP-VL checkpoint	`results/grasp_vl_llm_20k_current_train/grasp_vl.pt`	Main shared orthogonal prefix transform.
Large-pool retrieval	`scripts/evaluate/evaluate_grasp_vl_large_pool_retrieval.py`	Main prefix retrieval matrix and preservation check.
Prefix baselines	`scripts/evaluate/evaluate_frozen_vlm_prefix_baselines.py`	Direct, PCA, and random coordinate diagnostics.
Method baselines	`scripts/train/train_method_comparison_baselines.py`	MRL-style, SMEC-style, MLP, and permutation comparisons.
External probes	`scripts/evaluate/evaluate_sugarcrepe_prefix_transfer.py`	SugarCrepe-clean prefix transfer.
Downstream sanity	`scripts/evaluate/evaluate_downstream_prefix_sanity.py`	CIFAR-100 zero-shot classification by prefix.

Table 9: Repository artifacts needed to reproduce the reported appendix tables.

Item	Environment
GPU	2× NVIDIA RTX 4090, 24GB each
Driver	535.86.05
CUDA runtime	12.8
Python	3.11.15
PyTorch	2.10.0+cu128
OpenCLIP	3.3.0
Transformers	4.57.6
NumPy / pandas	2.4.3 / 3.0.1
Main device	Single RTX 4090 for reported B/16 runs

Table 10: Experimental hardware and software environment.

dering and hard-negative staircase remain stable. Together, the table-based diagnostics separate two potential confounds: the learned rotation is not just sorting coordinates, and the retrieval gains are not an artifact of including training captions as distractors.

Score decompositions. The main text reports compact staircase and emergence summaries. Tables 15 and 16 unpack those summaries into retrieval, hard-negative, and type-level emergence terms. This decomposition is useful because low raw leakage is not sufficient by itself: a weak prefix can have low leakage simply because it fails on every distinction. The detailed rows also show why we use emergence gap in the main narrative: GraSP-VL is not merely suppressing early prefixes, but sharply increases attribute and relation/action/order selectivity when the assigned boundary is reached.

External and out-of-objective transfer. The following probes are designed to reduce the circularity concern that training and evaluation use the same LLM-generated view types. Table 18 changes the prompt style at evaluation time. Table 19 uses held-out human captions and real image distractors rather than generated typed negatives. Table 20 evaluates SugarCrepe-clean, which is external to our annotation pipeline. Table 21 adds a no-training CIFAR-100 zero-shot classification sanity check. These probes are intentionally not identical to the training objective. The prompt-template probe asks whether the interface survives a change in wording, the human-caption probe removes generated typed negatives entirely, SugarCrepe-clean tests compositional contrasts from a separately constructed benchmark, and CIFAR-100 checks a common classification use case. They reduce annotation-style overfitting concerns, but they should be read as partial transfer checks rather than evidence that the learned interface covers all unseen semantic types or caption styles.

Backbone, objective, and schedule sensitivity. Table 22 expands the robustness summary from the main text. The objective ablations show that hard-negative ranking is needed for fine-grained selectivity, early-prefix invariance controls premature leakage, and non-orthogonal adaptation changes the full space. Table 23 varies the assigned semantic boundaries, treating $\kappa(r)$ as an interface contract rather than a discovered natural constant. The two tables should be read to-

Baseline	Transform	Training objective	Role in comparison
Direct prefix	None	No training	Tests whether raw coordinates already expose the interface.
PCA prefix	PCA projection	Train-split variance preservation	Tests generic low-dimensional retention, not semantic control.
Random rotation	Orthogonal random R	No training	Tests whether any basis change is enough.
MRL-style	Shared orthogonal R	All prefixes align to full captions; preservation loss	Proxy for nested retrieval preservation.
Matryoshka-style adaptor	Low-rank residual adaptor	All prefixes align to full captions; preservation loss	Proxy adaptor, not pairwise/top-k objective reproduction.
SMEC-style compression	Shared orthogonal R	Multi-view alignment and retention, no typed selectivity	Proxy for capability/retention; omits sequential compression and memory.
Learned permutation	Hard coordinate permutation	Full GraSP objective	Tests whether dimension sorting alone is enough.
Learned signed permutation	Signed coordinate permutation	Full GraSP objective	Coordinate-only control with sign flips shared by modalities.
Generic MLP adapter	Unconstrained MLP	Full GraSP objective	Capacity upper bound; may rewrite full geometry.

Table 11: Baseline implementation details. Compression rows are intentionally proxy baselines for nested usability or retention, not optimized for typed semantic boundary control.

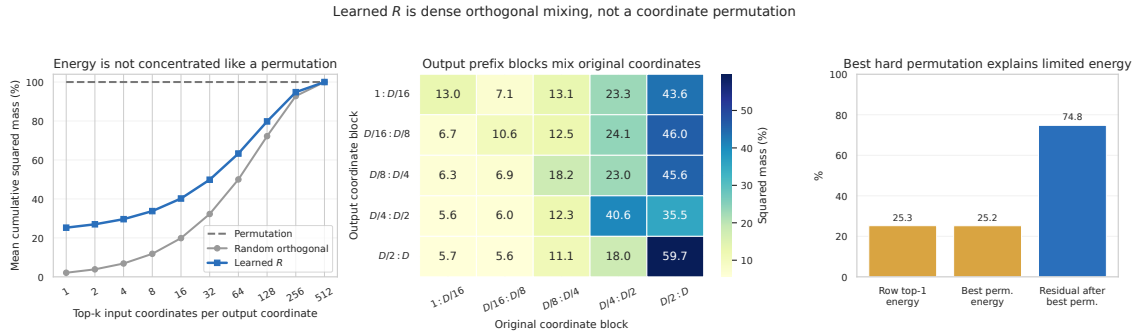


Figure 5: Full transform-structure diagnostic. The learned Cayley transform is dense and near-isometric rather than a hard coordinate permutation.

Setting	$D = 512$	$D = 768$	$D = 1024$
Query transform ops	0.26M	0.59M	1.05M
10M offline transform ops	2.62T	5.90T	10.49T
fp16 storage @ $D/16$	0.64GB	0.96GB	1.28GB
fp16 storage @ $D/2$	5.12GB	7.68GB	10.24GB
fp16 storage @ D	10.24GB	15.36GB	20.48GB
Dense transform params	0.26M	0.59M	1.05M

Table 12: Scaling estimates for a 10M-item gallery, excluding ANN index overhead. Dense Cayley has low parameter memory but an offline $O(ND^2)$ gallery transformation cost. Block-diagonal, Givens, or butterfly-style orthogonal transforms could reduce application cost to $O(Db)$ or $O(D \log D)$, at the possible cost of weaker cross-coordinate mixing.

Transform	Cost	Params	Stair.	Emerg.	Hard	Drift ↓	Cap.
Dense Cayley	$O(D^2)$	262,149	53.01	33.17	89.76	5.4e-07	34.91
Butterfly Givens, $s = 8$	$O(sD \log D)$	18,437	52.80	31.89	90.06	3.9e-07	33.11
Butterfly Givens, $s = 4$	$O(sD \log D)$	9,221	52.54	30.78	90.37	4.2e-07	31.58

Table 13: Structured orthogonal transform tradeoff.

gether: Table 22 asks whether the objective and transform constraints matter under the default interface, while Table 23 asks whether reasonable shifts of that interface still produce a coherent contract.

Prompt-template ablation. Table 24 checks whether the effect is caused only by the CLIP-style "a photo of" template used to construct short views. Removing the template mildly reduces caption retrieval but leaves the hard-negative staircase intact.

Metric definitions. For a prefix k and text view g, diagnostic retrieval reports image-to-text Re-

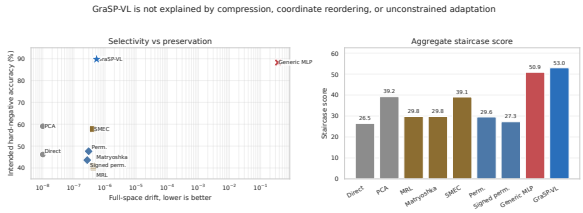


Figure 6: Selectivity–preservation trade-off across prefix-interface baselines.

call@1:

$$\mathrm{R@1}(k, g) = \frac{1}{|\mathcal{Q}|} \sum_{i \in \mathcal{Q}} \mathbf{1}[j_i^\star(k, g) = i]\,,$$
$$j_i^\star(k, g) = \arg\max_{j \in \mathcal{C}} s_k(x_i, t_j^g).$$

where $\mathcal{Q}$ is the held-out image-query set and $\mathcal{C}$ is the candidate text pool. Hard-negative selectivity for negative type r is

$$\mathrm{Sel}(k, r) = \frac{1}{|\mathcal{Q}_r|} \sum_{i \in \mathcal{Q}_r} \mathbf{1}[s_k(x_i, p_i^r) > s_k(x_i, n_i^r)]\,.$$

The intended retrieval average is

$$\begin{aligned}\mathrm{RetAvg} = \frac{1}{4}\Big[&\mathrm{R@1}(D/16, G_0)\\ &+ \mathrm{R@1}(D/8, G_1)\\ &+ \mathrm{R@1}(D/4, G_2)\\ &+ \mathrm{R@1}(D/2, G_3)\Big].\end{aligned}$$

Candidate Pool	Candidates	Obj.@$D/16$	Attr.@$D/8$	Rel.@$D/4$	Cap.@$D/2$	Ret. Avg.	Hard Neg. Avg.	Staircase	Drift ↓
Full 20,147 pool	20,147	10.86	3.13	16.32	34.91	16.31	89.76	53.03	3.6e-07
Test-only 1,501 pool	1,501	19.32	13.46	42.77	65.16	35.18	89.76	62.47	3.3e-07

Table 14: Candidate-pool sensitivity. Test images are used as queries in both rows. The full-pool setting retrieves against all 20,147 candidate texts, where train and validation captions are distractors; the test-only setting retrieves only against the 1,501 held-out test texts. Hard-negative selectivity is evaluated once on the same held-out test hard negatives and is therefore independent of retrieval candidate-pool size.

Method	Obj. R@1	Attr. R@1	Rel. R@1	Cap. R@1	Ret. Avg.	Obj. Neg.	Attr. Neg.	Rel. Neg.	Full Neg.	Hard Avg.	Staircase
Direct prefix	1.00	0.47	11.26	14.46	6.80	20.52	13.59	56.76	93.80	46.17	26.48
PCA prefix	3.06	1.33	27.12	46.04	19.39	59.56	27.51	52.50	96.80	59.09	39.24
MRL-style	6.26	2.93	26.58	41.44	19.30	16.46	9.66	38.24	96.60	40.24	29.77
Matryoshka-style adaptor	7.06	3.60	24.78	38.97	18.60	18.39	11.99	36.84	96.60	40.96	29.78
SMEC-style compression	10.79	3.46	25.92	41.11	20.32	56.83	26.05	51.97	96.54	57.84	39.08
Generic MLP adapter	10.59	2.86	15.92	24.72	13.52	71.15	87.74	98.20	95.87	88.24	50.88
Learned permutation	0.80	0.53	12.66	31.91	11.48	26.45	15.26	53.56	95.34	47.65	29.56
Learned signed permutation	0.67	0.80	13.06	29.25	10.94	23.38	16.06	39.97	95.00	43.60	27.27
GraSP-VL	10.73	3.13	16.32	34.91	16.27	77.48	88.41	96.54	96.60	89.76	53.01

Table 15: Staircase score decomposition. Retrieval and hard-negative columns use the intended prefix-view or prefix-negative pairing. Ret. Avg. and Hard Avg. are averaged to compute Staircase.

The intended hard-negative average is

$$\begin{aligned}\mathrm{HardAvg} = \frac{1}{4}\Big[&\mathrm{Sel}(D/16, \mathrm{object}) \\ &+ \mathrm{Sel}(D/8, \mathrm{attribute}) \\ &+ \mathrm{Sel}(D/4, \mathrm{relation}) \\ &+ \mathrm{Sel}(D/2, \mathrm{full})\Big].\end{aligned}$$

The staircase score used in the method comparison balances retrieval utility and hard-negative selectivity:

$$\mathrm{Stair} = \frac{1}{2}\left(\mathrm{RetAvg} + \mathrm{HardAvg}\right).$$

Early leakage averages fine-grained selectivity at prefixes earlier than their designed boundary:

$$\begin{aligned}\mathrm{Leak} &= \operatorname*{mean}_{(k,r)\in\mathcal{S}} \mathrm{Sel}(k, r), \\ \mathcal{S} &= \{(k, r) : k < \kappa(r)\}.\end{aligned}$$

For a distinction type r with at least one earlier prefix, we also report a capability-adjusted emergence gap:

$$\mathrm{Emerg}(r) = \mathrm{Sel}(\kappa(r), r) - \operatorname*{mean}_{k\in\mathcal{K}_{<r}} \mathrm{Sel}(k, r),$$

where $\mathcal{K}_{<r} = \{k : k < \kappa(r)\}$. This measures whether a distinction becomes more separable when it reaches its designed prefix, rather than rewarding weak early prefixes that fail everywhere. Full drift is the maximum absolute change in full-dimensional cosine similarities:

$$\Delta_{\max} = \max_{a,b} \left|\langle F_\theta(e_a), F_\theta(e_b)\rangle - \langle e_a, e_b\rangle\right|.$$

Method	Attr.	Relation	Action	Order	Full	Mean
Direct prefix	-6.66	13.32	14.69	11.16	6.68	7.84
PCA prefix	-12.19	-3.50	1.50	-1.37	3.02	-2.51
MRL-style	-0.40	2.23	-0.03	1.73	1.18	0.94
Matryoshka-style adaptor	0.40	1.00	1.50	1.77	0.73	1.08
SMEC-style compression	-10.66	-3.63	0.80	-6.10	3.40	-3.24
Generic MLP adapter	31.78	50.23	48.97	48.23	8.31	37.50
Learned permutation	-9.13	1.67	4.96	3.76	8.79	2.01
Learned signed permutation	-6.66	-7.00	-4.30	0.13	7.97	-1.97
GraSP-VL	24.72	47.73	43.67	41.74	7.99	33.17

Table 16: Emergence-gap decomposition. Each cell measures how much a semantic distinction improves when reaching its assigned prefix boundary: attribute at $D/8$, relation/action/order at $D/4$, and full-caption matching at $D/2$.

Probe	What changes from training?	Main readout	Interpretation
Swapped prompts	Text templates are changed at evaluation time	Intended-view R@1	Tests sensitivity to wording artifacts.
Human captions	Uses held-out human captions, not generated views	Image retrieval and object purity	Tests caption compatibility outside typed negatives.
SugarCrepe-clean	Uses external positive/negative caption pairs	Prefix accuracy and emergence	Tests compositional transfer outside our annotation pipeline.
CIFAR-100 zero-shot	Uses independent classification labels and prompts	Top-1 accuracy by prefix	Tests a common VLM use case without extra training.

Table 17: Out-of-objective probes and what each controls for.

Method	Prompt Set	Object@$D/16$	Attribute@$D/8$	Relation@$D/4$	Caption@$D/2$
Direct prefix	training style	1.00	0.47	11.26	14.46
Direct prefix	swapped templates	0.67	0.40	11.73	17.52
GraSP-VL	training style	12.19	2.93	16.32	34.91
GraSP-VL	swapped templates	4.33	2.60	17.59	32.84

Table 18: Prompt-template transfer. The model is trained with the default prompt style and evaluated with either the same style or a swapped template set. Scores are large-pool image-to-text R@1 at the intended prefix-view pair.

Method	Prefix	Object Purity@10	Category mAP	Full-pool R@1	Same-object R@1	Median Rank
Direct prefix	$D/16$ (32)	15.80	7.52	1.67	8.73	528
Direct prefix	$D/8$ (64)	24.11	10.64	3.80	14.26	152
Direct prefix	$D/4$ (128)	33.38	16.26	12.79	27.05	24
Direct prefix	$D/2$ (256)	40.29	20.67	23.58	40.11	8
Direct prefix	D (512)	43.68	24.07	42.11	55.56	2
GraSP-VL	$D/16$ (32)	34.85	24.81	1.53	7.40	144
GraSP-VL	$D/8$ (64)	37.00	24.60	4.26	13.86	62
GraSP-VL	$D/4$ (128)	42.54	24.95	21.25	37.18	8
GraSP-VL	$D/2$ (256)	43.92	23.37	36.24	49.37	3
GraSP-VL	D (512)	43.68	24.07	42.11	55.56	2

Table 19: Out-of-objective validation using held-out human captions and real image distractors. No LLM-generated views or typed negatives are used at evaluation time.

Type	Subsets	Expected Prefix	Acc@$D/16$	Acc@$D/8$	Acc@$D/4$	Acc@$D/2$	Acc@D	Peak	Ext. Emergence
Object	replace_obj, add_obj	$D/16$	86.03	91.71	88.42	92.70	92.33	$D/2$	–
Attribute	replace_att, add_att	$D/8$	62.64	72.30	74.93	82.23	83.51	D	10.66
Relation/Binding	replace_rel, swap_obj, swap_att	$D/4$	56.37	58.31	71.00	70.31	69.49	$D/4$	15.66

Table 20: SugarCrepe-clean prefix transfer for GraSP-VL. SugarCrepe is not generated by our annotation pipeline and is not used for training. The clean split removes images that appear in the GraSP-VL train or validation splits; in this run, no SugarCrepe image overlaps with train/validation, so clean equals all 7,511 examples.

Task	Method	$D/16$	$D/8$	$D/4$	$D/2$	D
Coarse group	Direct prefix	25.34	37.86	57.93	67.25	80.53
Coarse group	GraSP-VL	62.98	61.08	62.48	71.20	80.53
Superclass	Direct prefix	23.87	37.78	44.55	44.73	61.63
Superclass	GraSP-VL	33.47	35.19	44.12	49.42	61.63
Fine class	Direct prefix	19.48	42.39	61.08	62.30	76.95
Fine class	GraSP-VL	31.98	46.46	60.77	67.98	76.95

Table 21: No-training downstream sanity check on CIFAR-100 hierarchical zero-shot classification. Values are top-1 accuracy; the full-dimensional prefix corresponds to the frozen VLM geometry under the shared orthogonal transform.

Panel	Setting	Staircase	Emergence Gap	Hard Avg.	Cap. R@1	Drift ↓
Backbone	OpenCLIP ViT-B/16	53.01	33.17	89.76	34.91	5.4e-07
Backbone	OpenAI CLIP ViT-B/32	46.77	28.80	83.60	29.63	5.1e-07
Backbone	OpenAI CLIP ViT-B/16	50.27	31.20	86.98	32.53	5.7e-07
Backbone	OpenCLIP ViT-L/14	56.32	35.40	92.70	38.19	5.2e-07
Backbone	EVA-CLIP ViT-B/16	54.62	34.50	91.18	36.18	5.4e-07
Backbone	SigLIP ViT-B/16	51.89	32.50	88.50	33.64	5.6e-07
Objective	Full GraSP-VL	53.01	33.17	89.76	34.91	5.4e-07
Objective	w/o retention	53.19	33.83	90.24	33.98	4.5e-07
Objective	w/o hard-negative ranking	41.53	8.39	63.74	39.44	4.2e-07
Objective	w/o early-prefix invariance	57.03	1.82	97.87	33.11	4.8e-07
Objective	non-orthogonal transform	35.56	44.16	57.71	23.65	4.4e-01
Transform	Shared orthogonal	53.17	33.17	89.76	34.91	4.8e-07
Transform	Separate orthogonal	51.53	36.17	89.71	26.58	2.8e-01
Transform	Separate orthogonal + preservation	51.59	34.39	90.11	24.78	2.3e-01
Transform	Linear adapter + drift penalty	51.63	35.56	88.47	28.91	2.4e-01
Schedule	Default curriculum	53.17	33.17	89.76	34.91	4.8e-07
Schedule	All negatives after warmup	53.71	31.85	90.31	35.24	4.8e-07
Schedule	Slow curriculum	53.49	31.22	90.69	33.11	5.1e-07
Schedule	No warmup, all negatives	53.66	32.77	91.34	32.91	4.2e-07

Table 22: Detailed backbone, objective, transform, and schedule checks. Shared orthogonal GraSP-VL combines strong selectivity with near-exact full-space preservation.

Setting	Attr. κ	Rel. κ	Attr@κ	Rel/A/O@κ	Full@κ	Contract Hard Avg.	Pre-κ Leak. ↓	Default Stair.	Cap. R@1	Drift ↓
Default κ	$D/8$	$D/4$	88.41	96.91	96.60	89.85	64.47	53.17	34.91	4.8e-07
Relation delayed	$D/8$	$D/2$	91.07	96.25	96.67	90.80	62.41	47.15	34.44	4.8e-07
Attribute delayed	$D/4$	$D/4$	91.21	97.20	96.40	91.27	65.26	50.04	34.18	5.4e-07
Compressed attr.+rel.	$D/8$	$D/8$	96.67	98.42	96.14	92.61	68.89	54.15	32.31	5.1e-07

Table 23: $\kappa(r)$ sensitivity. We vary the earliest prefix at which attribute and relation/action/order negatives should be solved. Contract Hard Avg. evaluates each model under its own assigned interface contract, whereas Default Staircase measures the same model under the default paper ladder. These results should be read as contract sensitivity rather than evidence for a unique natural semantic boundary.

View Construction	Staircase	Hard Avg.	Cap. R@1	Drift ↓
Default template	53.01	89.76	34.91	5.4e-07
No template	51.73	88.84	32.05	5.1e-07

Table 24: Prompt-template artifact check. The no-template variant removes the "a photo of" style from object, attribute, and relation views, using raw entity, attribute phrase, and event phrase strings instead.