# CONCEPTS ARE DISTRIBUTIONS: TRAINING-FREE INSERTION IN VISION-LANGUAGE MODELS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Many activation-steering methods represent a concept as a direction in activation space. We test whether this representation also supports object insertion in the visual tokens of frozen vision-language models (VLMs). We find a clear asymmetry. A single direction can suppress an object, but it does not reliably introduce an object that is absent from the image. Learned directions, text-derived directions, and individual real activations all perform poorly. We show that the limitation is distributional. Repeating one concept vector across visual tokens can match the mean while collapsing covariance. A Gaussian whose population mean and covariance match those of real object activations produces responses comparable to transplanting a real object, whereas a surrogate that preserves the marginal statistics but destroys covariance does not. Existing feature-transfer operators follow the same ordering as the second-order structure they transfer. We therefore store each concept as a low-rank Gaussian estimated from a few images. On CLIP-L/14, four covariance directions are sufficient, giving a 20 KB representation per concept. Other vision encoders require modestly higher rank. The method requires no training and no reference image at inference time, transfers across five models and three vision encoders, and works without object-location annotations. These results show that object insertion depends on second-order visual-token structure and identify the visual-to-language activation stream as a security-sensitive interface when internal activations can be modified.

## 1 INTRODUCTION

Vision-language models generate text from visual tokens produced by an image encoder (Liu et al., 2023). These tokens are exposed between the encoder and the language model, which makes them a natural point for modifying what a frozen model reports without changing the input image or any model weight. In language models, many concepts can be approximated by directions in activation space, and adding or subtracting such directions can induce or suppress the corresponding behavior (Arditi et al., 2024). The same view has been adopted for VLMs. Visual-token directions have been used to reduce hallucination (Li et al., 2025) and to remove objects from generated captions (Jiang et al., 2025).

We ask whether this directional view also supports the converse operation: introducing a concept that is absent from an image. We find that it does not. Figure 1 illustrates the asymmetry. Projecting an object direction out of the visual tokens that cover an object removes it from the caption in a substantial fraction of cases. Adding the direction of a target concept at the same positions introduces that target in only 13.8% of cases. Other single vectors perform similarly poorly, including the real activation that best represents the concept and a direction sampled from a Gaussian over concept vectors (Zhao et al., 2025). These interventions can remove the original object, but they generally do not create the target.

We argue that insertion fails because a single direction does not provide the target distribution required by the language model. An object occupies $m$ visual tokens. Replacing those tokens with $m$ copies of one vector produces a set whose mean is that vector and whose covariance is zero. Adding one vector instead preserves only the covariance already present in the source tokens. We test the role of second-order structure with surrogates that preserve progressively more statistics of a real target object. Preserving only the mean yields 6.7% injection. Preserving every per-dimension marginal



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.




Figure 1: Overview on one image. The photograph and all model weights remain unchanged. We edit only the marked visual tokens before they reach the language model. Projecting out the truck direction removes the truck from the caption. Adding the giraffe direction to the same tokens does not make the model report a giraffe. Replacing those tokens with samples from the stored 20 KB Gaussian for the giraffe concept makes the model report a giraffe and integrate it with the scene.

while destroying covariance reaches 20.0%. By contrast, a Gaussian whose population mean and covariance match the object reaches 52.7%, even though it contains no real activation. This is statistically indistinguishable from transplanting the real object at 50.0% (11:15, $p = 0.56$). Existing feature-transfer operators, including AdaIN (Huang and Belongie, 2017) and WCT (Li et al., 2017), follow the same ordering when ranked by the target second-order structure they transfer.

This observation leads directly to a practical representation. We estimate the mean and covariance of a concept once from a small set of images and store a low-rank factorization. At inference time, we sample from these statistics instead of requiring a reference image or a stored real activation. On CLIP-L/14, rank four costs 20 KB per concept and reaches 59.3% injection, exceeding a real donor object (28:14, $p = 0.044$). The same statistics transfer to three language models that share the vision encoder, and the mechanism also holds for two models with different encoders. Direct questioning gives the same qualitative conclusion. The unedited model answers "no" for the absent target in 98.0% of cases, whereas the edited model answers "yes" in 88.0%. Under the same edit, it continues to reject an unrelated absent concept.

Our results support training-free counterfactual interventions in multimodal representations and clarify why constructive editing differs from suppression. They also show that systems exposing intermediate activations should treat the visual-to-language activation path as part of the trusted inference boundary. Our main contributions are:

- We identify and explain an asymmetry between removing and inserting concepts in the visual-token stream of a frozen VLM. Removal succeeds with a rank-one modification, whereas insertion requires target-specific second-order structure.
- We show that real activations are unnecessary for insertion. A low-rank Gaussian estimated from a few images matches or exceeds activation transplantation across five models and three vision encoders.
- We verify the edit with direct questions, free recall, forced choice, and human judgment, showing that the effect is specific to the inserted concept rather than a generic tendency to affirm.

Additional controls, null results, and failed predictions are reported in Appendix D.

## 2 Related Work

**Concepts as directions.** Steering vectors, refusal directions, and task vectors represent concepts as directions in activation space (Arditi et al., 2024; Li et al., 2025; Huang et al., 2024). The closest work to ours is GCS (Zhao et al., 2025), which models uncertainty over probe-derived concept vectors with a Gaussian but still adds one sampled direction to the representation. Our Gaussian instead models the distribution of token activations, and an intervention substitutes $m$ samples from that distribution. A single draw from a concept-vector Gaussian behaves like the other single-vector operators in our

experiments. Birardi and Paulo (2025) also report that suppression is easier than targeted steering for attribution-graph features in language models, which they attribute to feature coverage. We identify a different mechanism in the vision-language setting.

**Second-order statistics in feature transfer.** Neural style transfer has represented style through Gram matrices since Gatys et al. (2016). AdaIN matches per-channel means and variances (Huang and Belongie, 2017), while WCT transfers the full covariance (Li et al., 2017). We use these operators as controlled probes of object identity in a frozen VLM. This moves the question from image synthesis to language readout and from matching style statistics to synthesizing object-token distributions.

**Editing VLM representations.** ProjectAway removes objects without training by orthogonalizing image features against a text embedding (Jiang et al., 2025). VISTA-SAE (Kravets et al., 2026)[1] performs localized removal and replacement after training a projector for roughly two hundred GPU-hours. Our method is training-free and supports both removal and insertion.

**Class-conditional features, activation patching, and dictionaries.** Class-conditional Gaussian models are standard in anomaly detection (Lee et al., 2018; Defard et al., 2021). Our shared-covariance variant uses the same statistical model generatively. Activation patching is an established interpretability primitive (Vig et al., 2020; Meng et al., 2022; Heimersheim and Nanda, 2024), and Huang et al. (2024) use mean activations constructively. Our mechanism predicts that a tiled mean should be insufficient for insertion. Sparse autoencoders provide another route to visual features, but they require additional training (Joseph et al., 2025; Pach et al., 2025; Leask et al., 2025; Klindt et al., 2023). Archetypal variants constrain learned atoms to the convex hull of the data (Fel et al., 2025), whereas Gaussian samples are not subject to this constraint.

# 3 METHOD

## 3.1 PROBLEM SETTING

We consider a frozen VLM whose vision encoder produces patch activations $h_1, \ldots, h_N \in \mathbb{R}^d$. These activations pass through a projector before they are read by the language model. Let $P$ denote the set of patches covering an object bounding box and let $m = |P|$. An insertion intervention replaces $\{h_p\}_{p \in P}$ with $m$ substitute vectors. Each substitute is scaled to the norm of the activation it replaces. No other part of the forward pass changes and no weight is updated. We call the resulting $m \times d$ matrix the *substituted set* and ask which statistics of this set determine whether the language model reports a target concept.

## 3.2 MOMENT-MATCHED SURROGATES

To isolate the relevant statistics, we construct surrogates of a real donor block $D \in \mathbb{R}^{m \times d}$ taken from an image of the target class. Let $\mu$ be the row mean of $D$, let $d_i$ denote its rows, and define $X = D - \mu$.

The *mean* surrogate tiles $\mu$ over all $m$ patches. Its first moment is exact and its covariance is zero. The *marginals* surrogate independently permutes every feature dimension across rows. It therefore preserves every per-dimension marginal and contains exactly the same $md$ numerical values as $D$, while destroying their cross-dimensional covariance.

For the *Gaussian* surrogate, we draw

$$x_j = \mu + \sum_i c_{ij}(d_i - \mu), \qquad c_{ij} \sim \mathcal{N}(0, 1/m), \tag{1}$$

and recenter the samples so that their empirical mean is $\mu$. The population covariance of the sampling distribution is $X^\top X/m$. A finite set of $m$ samples matches this covariance in expectation, and the per-patch norm matching described above perturbs both moments slightly. No sample is a real activation. Finally, the *four-object* surrogate draws real target-class activations from four images. This preserves class-level marginals without preserving the joint structure of a single object.

[1]We use "VISTA-SAE" to distinguish the SAE-transfer method of Kravets et al. (2026) from the unrelated VISTA method of Li et al. (2025).

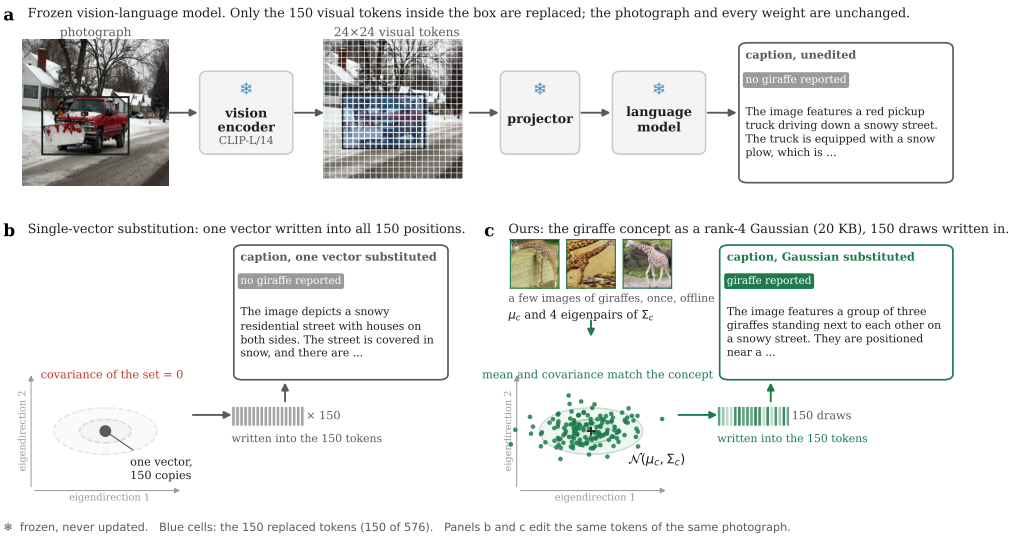


Figure 2: Intervention on one image. The frozen model encodes the photograph into a $24 \times 24$ grid of visual tokens. We replace only the tokens inside the marked box. Repeating one vector across these positions gives a substituted set with zero covariance and does not make the model report the target. Our method estimates a low-rank Gaussian for the target concept from a few images and writes $m$ independent draws into the same positions. The model then reports the target.

### 3.3 Gaussian concept bank

If a concept is adequately described by the mean and covariance of its token distribution, these statistics can be estimated offline and stored once. For each class $c$, we pool the activations inside the bounding boxes of a small number of images. We store the mean $\mu_c$ together with the leading $r$ eigenpairs of the covariance $\Sigma_c$. At edit time, we draw $m$ activations from the resulting rank-$r$ Gaussian and substitute them at the positions in $P$ (Figure 2). The storage cost is $(r + 1)d$ values per concept. For CLIP-L/14 with $r = 4$, this is 20 KB. The procedure requires no training, no segmentation model, and no clustered dictionary.

## 4 Experiments and Results

### 4.1 Experimental setup

Unless stated otherwise, we intervene on a frozen LLaVA-1.5-7B (Liu et al., 2023) at the input of its multimodal projector. The visual stream contains $N = 576$ patch activations of dimension $d = 1024$ from a CLIP-L/14-336 vision tower (Radford et al., 2021). We use images from the COCO validation set (Lin et al., 2014) in which the source object occupies 2–30% of the frame. Removal can only be scored when the unedited caption names the source object. Of 400 eligible images, 286 satisfy this criterion. We use the first 150 for condition sweeps and reserve the remaining 136 as held-out recognized images. The 114 images on which the source is not named are later used to test insertion outside the selected population.

We report four quantities. Injection is the proportion of edited captions that name the target concept. We validate this metric against direct questioning in Section 4.5 and against human judgment in Appendix I. Removal is the proportion of edited captions in which the source object is no longer named. Retention is the proportion of content words from the unedited caption, excluding the source and target concepts, that remain in the edited caption. We use it as an automatic proxy for scene fidelity. Full replacement is the proportion of captions that name the target and no longer name the source. Every condition within an experiment uses the same images and target assignments, so comparisons are paired. Gaussian and donor conditions are resampled for each experiment. Their absolute rates therefore vary by up to 0.02 across experiments (seed s.d. 0.016), and each comparison

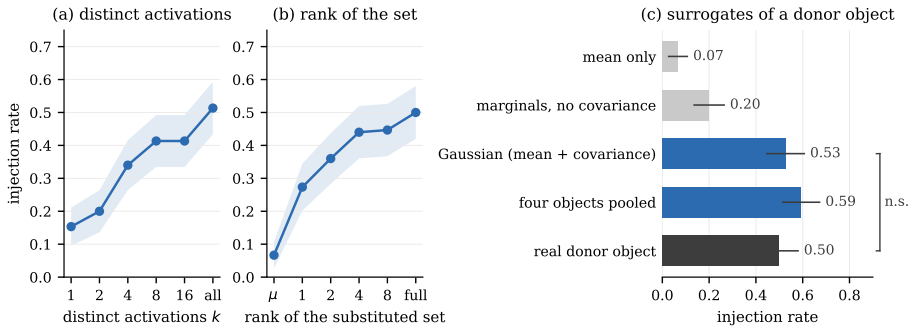


Figure 3: What insertion requires on the same 150 COCO images. (a) Increasing the number of distinct target activations raises injection while holding coverage fixed. (b) A low-rank reconstruction of the donor block shows that most of the effect appears by rank four. (c) The mean and marginal-preserving surrogates fail, whereas a Gaussian with the target first two population moments matches the real donor object. Pooling four objects performs better still. Shaded bands and error bars are 95% binomial intervals.

reports the values from its own run. We use exact McNemar tests and report discordant counts as $a$:$b$, where $a$ is the number of images on which the first condition succeeds and the second fails, and $b$ is the converse. Appendix A gives the complete protocol.

### 4.2 What insertion requires

We first test whether the provenance of a single vector matters. Substituting one vector at every position in $P$ gives a set with mean $v$ and zero covariance. Adding the same $v$ to each token instead preserves the covariance of the source tokens. Neither operation supplies the target-specific second-order structure. Adding a target direction obtained from a clustered visual dictionary gives 13.8% injection. Overwriting the same tokens with that direction gives 11.3%. A medoid target activation gives 22.6%, an arbitrary real target activation gives 1.9%, and a direction sampled from GCS (Zhao et al., 2025) gives 5.3%. The vector source therefore has little effect on the basic failure.

The result changes when the substituted set contains multiple distinct activations. Figure 3(a) repeats $k$ target activations over the same $m$ positions while keeping coverage fixed. Injection rises monotonically from 0.15 for one activation to 0.51 for the full donor set. Figure 3(b) gives the same pattern when the donor block is reconstructed at increasing rank. Four directions already recover most of the effect.

The moment-matched surrogates in Table 1 identify the relevant property. The tiled mean and the marginal-preserving surrogate are both significantly below the real donor object. The Gaussian surrogate is statistically indistinguishable from it (11:15, $p = 0.56$), even though it contains no real activation. Pooling activations from four objects exceeds the donor (10:24, $p = 0.024$). These results support three conclusions. First, the model requires second-order structure rather than the individual numerical values of a real donor block. Second, real target activations are unnecessary at edit time. Third, the joint structure of one particular object is unnecessary. The four-object result is consistent with a better class-level estimate giving stronger insertion. Section 4.4 tests this prediction with statistics estimated offline. The non-significant Gaussian-versus-donor test does not establish equality. With $n = 150$, it excludes differences larger than approximately 0.1.

### 4.3 Comparison with existing operators

If the distributional account is correct, existing operators should rank according to the target second-order structure they transfer, even when they were designed for different purposes. AdaIN (Huang and Belongie, 2017) matches the per-channel mean and variance and therefore transfers the diagonal of the covariance. WCT (Li et al., 2017) whitens source activations and recolors them with the target covariance. The steering operators of Jiang et al. (2025), Li et al. (2025), and Kravets et al. (2026) do

Table 1: Moment-matched surrogates of a real donor object, $n = 150$. Each condition is paired against the donor. Bold and underline mark the best and second-best injection rates.

| surrogate | preserved | inject ↑ | retain ↑ | vs. donor |
|---|---|---|---|---|
| mean | first moment | 0.067 | 0.536 | 66:1, $p = 9 \times 10^{-19}$ |
| marginals | marginals | 0.200 | 0.513 | 49:4, $p = 7 \times 10^{-11}$ |
| Gaussian | first and second moments | <u>0.527</u> | 0.389 | 11:15, $p = 0.56$ |
| four objects | class marginals | **0.593** | 0.368 | 10:24, $p = 0.024$ |
| real donor object | all | 0.500 | 0.378 | – |

Table 2: Cross-dataset comparison of representative insertion operators. COCO uses the 150-image discovery set. LVIS uses 300 images from 50 classes absent from the COCO evaluation, with all settings frozen from COCO.

| operator | COCO inject ↑ | LVIS inject ↑ | LVIS retain ↑ |
|---|---|---|---|
| mean only | 0.020 | 0.030 | 0.548 |
| text direction (Jiang et al., 2025) | 0.320 | 0.283 | 0.546 |
| AdaIN (Huang and Belongie, 2017) | 0.180 | 0.190 | 0.558 |
| WCT, rank 4 (Li et al., 2017) | 0.307 | 0.320 | 0.502 |
| real donor object | 0.500 | 0.483 | 0.386 |
| ours, rank-4 Gaussian | 0.593 | **0.560** | 0.391 |
| ours, rank-8 Gaussian | **0.627** | **0.583** | 0.383 |

not transfer target second-order structure. We restrict every operator to the object patches and tune its strength independently, as described in Appendix B.

Table 2 tests representative operators on both COCO and an independent LVIS benchmark. The LVIS evaluation uses 300 images from 50 classes that do not appear in the COCO evaluation, and every setting is frozen from COCO. The pooled Gaussian variants are strongest on both datasets and exceed a real donor object. WCT consistently exceeds AdaIN. The text-direction baseline is competitive on COCO but drops on LVIS and remains well below the Gaussian. The complete COCO baseline comparison, including GCS, residual steering, VISTA-SAE, WCT rank 16, retention, and paired tests, is reported in Appendix B, Table 6. The source block spans a median of 96 out of 1024 dimensions, and WCT reaches the target covariance only up to a median relative error of 0.345. The pooled Gaussian is not limited by this source geometry.

### 4.4 Gaussian concept bank

We next evaluate a bank estimated for 73 classes from a mean of 503 patches per class. Table 3 and Figure 4(a) show how injection changes with rank. The mean alone gives 0.020, rank two gives 0.547, and rank four gives 0.593. Rank four exceeds a real donor object (28:14, $p = 0.044$). Performance is approximately flat from rank four to rank 32. It declines at rank 64 when the bank is estimated from more images, which is consistent with the tail of the covariance estimate containing noise. We therefore use rank four as the compact CLIP-L/14 representation. Appendix J gives qualitative examples.

We further study how many images are needed to estimate each concept while keeping the Gaussian rank fixed at four. With four estimation images, the rank-4 bank already reaches 0.560 injection, compared with 0.593 using twelve images, while retention remains nearly unchanged. This shows that most of the benefit is obtained from only a few examples, without increasing the 20 KB representation size. Appendix F, Figure 8 and Table 8, give the full results.

The result is stable beyond the discovery set. On the 136 held-out images whose source is recognized, rank four gives 0.603 injection. On the 114 images whose source is not named by the unedited model, it gives 0.570. Across all 400 eligible COCO images, it gives 0.589. The donor object remains between 0.482 and 0.500 over these splits (Appendix F, Table 7). Insertion therefore does not depend on the model first naming the object that is replaced. Table 2 also reports the independent LVIS evaluation. With every setting frozen from COCO, the rank-4 Gaussian reaches 0.560 compared with 0.483 for the donor object (44:21, $p = 5.9 \times 10^{-3}$), and the same operator ordering is preserved. Appendix G, Table 9, breaks the LVIS result down by class frequency. The mean-versus-distribution gap also persists at other intervention points, across precisions and serving implementations, and with different projector architectures (Appendix H, Tables 10–13).

Table 3: Injection against storage cost, $n = 150$, with each representation paired against a real donor object.

| representation | storage ↓ | inject ↑ | full replacement ↑ | vs. donor |
|---|---|---|---|---|
| mean only | 4 KB | 0.020 | 0.013 | 0:72, $p = 4 \times 10^{-22}$ |
| mean, rank 2 | 12 KB | 0.547 | 0.400 | 26:19, $p = 0.37$ |
| mean, rank 4 | 20 KB | 0.593 | 0.433 | 28:14, $p = 0.044$ |
| mean, rank 8 | 36 KB | **0.627** | 0.420 | 29:10, $p = 3.4 \times 10^{-3}$ |
| mean, rank 16 | 68 KB | 0.593 | **0.453** | 27:13, $p = 0.039$ |
| real donor object | one image | 0.500 | 0.387 | – |

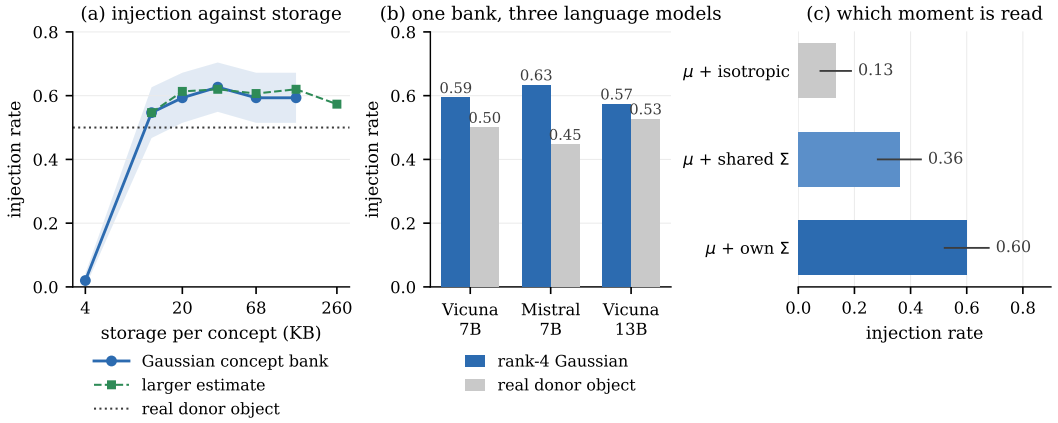


Figure 4: Gaussian concept bank. (a) Injection against storage per concept. The rank-4 representation uses 20 KB and exceeds a real donor object. The dashed curve repeats the sweep with statistics estimated from more images. (b) One CLIP-space bank transfers to three language models that share the vision encoder. (c) A shared covariance makes the mean readable, while the concept-specific covariance gives the strongest response and carries additional identity information.

We then test transfer across language models. Vicuna-7B, Mistral-7B in the BakLLaVA configuration, and Vicuna-13B share the same CLIP feature space but use different projectors and language models. Their CLIP features are numerically identical up to a relative difference of $1.2 \times 10^{-2}$ and a mean cosine similarity of 0.9999. A single rank-4 bank estimated under Vicuna-7B transfers to the other two models without refitting, as shown in Table 4 and Figure 4(b).

The mechanism also extends to different vision encoders, although the required rank changes. LLaVA-OneVision uses SigLIP-so400m-384 with 729 tokens of dimension 1152. The mean alone gives 0.047 and injection rises monotonically with rank. Rank four does not outperform the donor object (12:19, $p = 0.28$), while rank eight reaches parity (20:15, $p = 0.50$). The compression point is therefore a property of the representation rather than a fixed constant of the method. On Qwen2.5-VL-7B, the mean alone gives 0.033 and rank four gives 0.300. Ranks eight and sixteen exceed the donor-object rate of 0.347 (23:10, $p = 0.035$ and 26:8, $p = 2.9 \times 10^{-3}$). The effect is therefore not specific to the LLaVA family.

Finally, we vary the first and second moments independently to determine which one carries concept identity. A mean with isotropic noise of matched total variance gives 0.133. The same mean with a pooled within-class covariance shared across all 73 concepts gives 0.360. Using the concept's own covariance raises injection to 0.600. Thus, covariance first makes the mean readable and then contributes concept-specific information. When the mean and covariance come from different concepts, the covariance concept is named more often than the mean concept (0.353 versus 0.227, 44:25, $p = 0.030$). Reversing the assignment gives the same direction but is not significant (37:28, $p = 0.32$). Subtracting each class's rate in the unedited captions changes these differences by less than 0.001. A compact bank that shares one covariance across all classes, estimated with the leading 64 eigenpairs of the pooled within-class covariance, gives 0.440 while storing only the class mean at 4 KB per concept. This is indistinguishable from a real donor object (33:24, $p = 0.29$), but remains below the per-class covariance (31:7, $p = 1.2 \times 10^{-4}$).

Table 4: Transfer across models. Values are injection rates for the stored Gaussian and the donor-object baseline of each model. Dashes denote ranks not evaluated in the fixed-rank transfer experiment.

| | Vicuna-7B | Mistral-7B | Vicuna-13B | OneVision | Qwen2.5-VL |
|---|---|---|---|---|---|
| vision encoder | CLIP-L/14-336 | | | SigLIP-384 | Qwen ViT |
| mean only | 0.020 | 0.020 | 0.033 | 0.047 | 0.033 |
| rank-4 Gaussian | 0.593 | **0.633** | **0.573** | 0.300 | 0.300 |
| rank-8 Gaussian | **0.627** | – | – | 0.380 | 0.433 |
| rank-16 Gaussian | 0.593 | – | – | **0.387** | **0.467** |
| real donor object | 0.500 | 0.447 | 0.527 | 0.347 | 0.347 |

Table 5: Direct questioning on the same 150 images and rank-4 Gaussian edit. The target is the inserted concept and the control is an absent, unrelated concept.

| | yes/no | inventory | forced choice |
|---|---|---|---|
| target affirmed, unedited image ↓ | 0.020 | 0.013 | 0.020 |
| target affirmed, edited image ↑ | 0.880 | 0.647 | 0.800 |
| control affirmed, edited image ↓ | 0.053 | 0.033 | 0.027 |
| source affirmed, edited image | 0.500 | 0.460 | 0.533 |
| target vs. control, edited | 122:0, $p = 3.8 \times 10^{-37}$ | 94:2, $p = 1.2 \times 10^{-25}$ | 117:1, $p = 7.2 \times 10^{-34}$ |

### 4.5 DIRECT QUESTIONING

A caption-level string match could reflect lexical emission rather than a broader change in the model's answers. We therefore ask closed questions about object presence. Before editing, the model answers "no" for the absent target in 98.0% of cases and "yes" for the source object in 98.7%. After substituting the rank-4 Gaussian, it answers "yes" for the target in 88.0%, corresponding to a change on 129 of 150 images. Its affirmation of the source drops to 50.0%.

We next test specificity. Under the same intervention, we ask whether a third, unrelated concept is present. The model affirms this control in 5.3% of edited images, compared with 2.7% before editing. The target is affirmed while the control is rejected on 122 images, and the reverse never occurs ($p = 3.8 \times 10^{-37}$). We then repeat the test with two prompts that do not presuppose the target's presence. Table 5 reports all three prompt conditions. Free recall still names the target in 0.647 of edited images while naming the unrelated control in only 0.033. Forced choice selects the target in 0.800 of cases and the control in 0.027. The effect is therefore robust to prompt format and remains specific to the inserted concept. Appendix E gives the exact prompt text and the concept-level susceptibility analysis.

The direct questions also show that caption matching is conservative for this condition. The caption omits the target while the model affirms it on 49 images. The reverse occurs on only six ($p = 1.8 \times 10^{-9}$). Concept susceptibility is not uniform. Discrete objects of intermediate size are affirmed reliably, while thin implements and large furniture are less susceptible. Appendix E reports the concept-level profile.

### 4.6 REMOVAL

Removal uses a separate clustered dictionary of approximately $K = 230$ visual regions. We build it from a held-out activation stream using a cosine threshold. For a source object, we identify the region whose centroid is most over-represented inside $P$ relative to the rest of the image, and project that region direction out of the object tokens with strength $\beta$. At the strength used in the fidelity sweep, this removes the object from 39.3% of captions. A random direction at matched norm removes it in 0.7%. The model is therefore insensitive to generic perturbations of the same magnitude.

Figure 5(a) shows why removal cannot be judged by removal rate alone. Projecting out the global mean direction removes the object in 86.0% of cases but preserves only 14.1% of the remaining caption. The object's region direction gives the best balance among the four operators. This rank-one

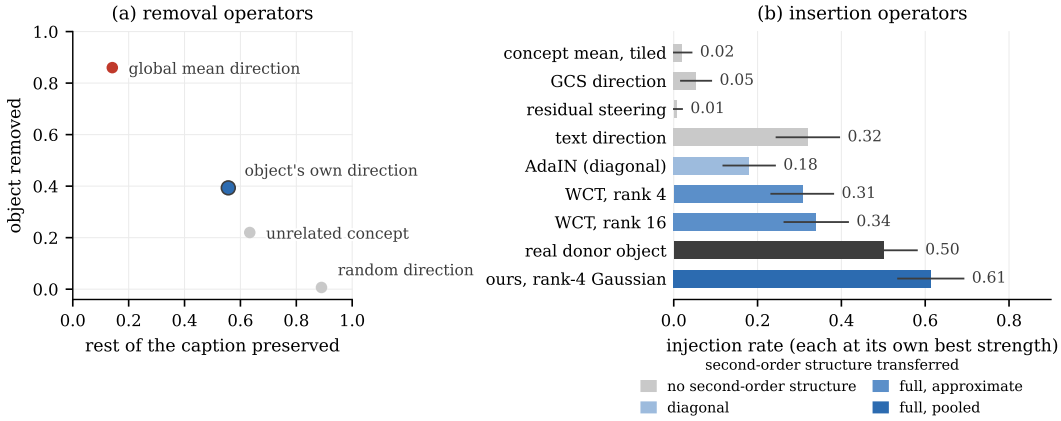


Figure 5: Removal and insertion are not symmetric. (a) Removal operators are plotted by the fraction of captions in which the source disappears against the fraction of the remaining caption that is preserved. A single object direction can remove the source while retaining most of the scene description. (b) Insertion operators on the same images rank with the target second-order structure they transfer. Operators that transfer no target covariance perform worst, AdaIN transfers only the diagonal, WCT approximates full covariance, and the pooled Gaussian performs best.

removal result is consistent with linear concept-erasure theory. Belrose et al. (2023) show that linear guarding can be removed by equalizing class-conditional means and eliminating feature-label cross-covariance. Avitan et al. (2026) further show that within-class covariance asymmetry remains after mean equalization and can still support nonlinear recovery. Our insertion results provide a complementary causal observation: the model acts on the second-order structure that a first-moment intervention leaves behind.

At its tuned strength $\beta = 20$, our operator removes the source in 44.7% of cases. A tuned text-derived direction reaches 54.7% (29:14, $p = 0.031$). Our operator is more stable as the strength increases. Over $\beta \in [1, 160]$, its retention never falls below 0.47, whereas the text-direction baseline drops from 0.414 at $\beta = 20$ to 0.050 at $\beta = 40$ (Appendix C).

### 4.7 Insertion without annotation

The experiments above use the ground-truth box to select intervention positions. Insertion is much less dependent on exact location. A centered square with the same area as the object gives 0.520 injection, compared with 0.587 for the annotated box (23:13, $p = 0.13$). A fixed $8 \times 8$ centered square, which uses no scene information, gives 0.460. A randomly placed square of matched area gives 0.473. Insertion is therefore sensitive to how much of the visual stream is replaced but comparatively insensitive to the precise location. Full replacement behaves differently. It falls from 0.420 with the annotated box to 0.173 with a centered square because removing the source does require its location. Replacing all 576 tokens yields 0.680 injection but only 0.019 retention, at which point the model describes almost exclusively the inserted concept.

## 5 Conclusion and Limitations

We have shown that removal and insertion behave differently in frozen VLM visual-token streams. A rank-one modification can suppress an object, whereas insertion requires target-specific second-order structure. A real donor is unnecessary. On CLIP-L/14, its distributional estimate is stored at rank four in 20 KB per concept. SigLIP and Qwen require modestly higher rank, so the compression point depends on the representation. The mechanism replicates on an independent long-tailed benchmark, at multiple visual-to-language intervention points, and across precisions, serving implementations, and projector architectures. Experiments are limited to object concepts and white-box access. Within this scope, low-dimensional target-specific second-order structure supports training-free counterfactual editing, and the visual-to-language activation path should be treated as a security-sensitive interface.

## Reproducibility Statement

Section 3 specifies the intervention and Section 4 describes the evaluation protocol and scoring criteria. Appendix A gives the calibration procedure, sample sizes, target assignment, and surrogate construction. Conditions within each experiment use the same images and fixed target assignments and are compared with exact paired tests. Code for the experimental harness, all intervention conditions, and the analysis scripts will be released upon acceptance.

## Ethics Statement

This work studies inference-time manipulation of a frozen VLM by modifying its internal visual activations without changing the image or model weights. The intervention requires access to activations passed between the vision encoder and the language model. It is therefore a white-box attack on the inference pipeline rather than an image-space attack. A relevant threat model is a compromised serving component or an untrusted adapter that can modify this activation stream. An adversary limited to supplying input images cannot perform the intervention studied here. Once internal access is available, the procedure requires only stored concept statistics and no gradient computation. Protecting the integrity of activations exchanged between model components is therefore a natural mitigation. We report this behavior to support evaluation of systems that expose or modify intermediate representations. All model experiments use public models and the public COCO and LVIS datasets. Two annotators evaluated model-generated captions (Appendix I). No personal, sensitive, or private participant data were collected.

## AI Use Statement

AI assistants were used for wording refinement, LaTeX troubleshooting, and debugging some Bash scripts. They were not used to generate experimental results, formulate the method, or make final scholarly judgments. The authors take full responsibility for all claims, citations, data, analyses, and submitted material.

## References

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems*, 2024.

Matan Avitan, Yoav Goldberg, and Yanai Elazar. MANCE: Manifold aware concept erasure. arXiv:2607.03973, 2026.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. LEACE: Perfect linear concept erasure in closed form. In *Advances in Neural Information Processing Systems*, 2023.

Giuseppe Birardi and Gonçalo Paulo. Automated attribution graph interpretation via probe prompting. arXiv:2511.07002, 2025.

Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. PaDiM: A patch distribution modeling framework for anomaly detection and localization. In *ICPR International Workshops*, 2021.

Thomas Fel, Ekdeep Singh Lubana, Jacob S. Prince, Matthew Kowal, Victor Boutin, Isabel Papadimitriou, Binxu Wang, Martin Wattenberg, Demba Ba, and Talia Konkle. Archetypal SAE: Adaptive and stable dictionary learning for concept extraction in large vision models. In *International Conference on Machine Learning*, 2025.

Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

Agrim Gupta, Piotr Dollár, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

Stefan Heimersheim and Neel Nanda. How to use and interpret activation patching. arXiv:2404.15255, 2024.

Brandon Huang, Chancharik Mitra, Assaf Arbelle, Leonid Karlinsky, Trevor Darrell, and Roei Herzig. Multimodal task vectors enable many-shot multimodal in-context learning. arXiv:2406.15334, 2024.

Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *IEEE International Conference on Computer Vision*, 2017.

Nick Jiang, Anish Kachinthaya, Suzanne Petryk, and Yossi Gandelsman. Interpreting and editing vision-language representations to mitigate hallucinations. In *International Conference on Learning Representations*, 2025.

Sonia Joseph, Praneet Suresh, Lorenz Hufe, Edward Stevinson, Robert Graham, Yash Vadi, Danilo Bzdok, Sebastian Lapuschkin, Lee Sharkey, and Blake Aaron Richards. Prisma: An open source toolkit for mechanistic interpretability in vision and video. arXiv:2504.19475, 2025.

David Klindt, Sophia Sanborn, Francisco Acosta, Frédéric Poitevin, and Nina Miolane. Identifying interpretable visual features in artificial and biological neural systems. arXiv:2310.11431, 2023.

A. Kravets, D. Li, C. Li, D. Chen, and V. P. Namboodiri. Interpretability transfer from language to vision via sparse autoencoders. In *International Conference on Machine Learning*, 2026.

Patrick Leask, Bart Bussmann, Michael Pearce, Joseph Bloom, Curt Tigges, Noura Al Moubayed, Lee Sharkey, and Neel Nanda. Sparse autoencoders do not find canonical units of analysis. In *International Conference on Learning Representations*, 2025.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, 2018.

Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature transforms. In *Advances in Neural Information Processing Systems*, 2017.

Zhuowei Li, Haizhou Shi, Yunhe Gao, Di Liu, Zhenting Wang, Yuxiao Chen, Ting Liu, Long Zhao, Hao Wang, and Dimitris N. Metaxas. The hidden life of tokens: Reducing hallucination of large vision-language models via visual information steering. arXiv:2502.03628, 2025.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, 2014.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, 2023.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, 2022.

Mateusz Pach, Shyamgopal Karthik, Quentin Bouniot, Serge Belongie, and Zeynep Akata. Sparse autoencoders learn monosemantic features in vision-language models. In *Advances in Neural Information Processing Systems*, 2025.

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. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

T. Schaumlöffel, M. G. Vilas, and G. Roig. Mechanisms of object localization in vision-language models. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2026.

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. In *Advances in Neural Information Processing Systems*, 2020.

Haiyan Zhao, Heng Zhao, Bo Shen, Ali Payani, Fan Yang, and Mengnan Du. Beyond single concept vector: Modeling concept subspace in LLMs with Gaussian distribution. In *International Conference on Learning Representations*, 2025.

## A Protocol Details

A forward pre-hook on the multimodal projector captures the $576 \times 1024$ visual-token matrix and replaces the selected rows. Each substitute is scaled to the norm of the activation it replaces.

We construct the COCO evaluation set from images in which a single instance of a source class occupies between 2% and 30% of the frame. This gives 400 images over 73 classes. An image is scorable for removal when the unedited caption contains the class name or a listed synonym, which holds for 286 images. The condition sweeps use the first 150 scorable images in a fixed order.

For an image $i$ with source class $c$, the target is selected as the $i$-th element, modulo the list length, from the sorted list of classes other than $c$ that are not annotated as present in that image. The same assignment is used for every condition and experiment.

Injection and removal are scored with word-boundary regular-expression matches against the lowercased class name and its synonyms. Retention is the Jaccard-style proportion of content words from the unedited caption that remain in the edited caption after stop words and both class names are excluded.

For each concept, we pool activations inside the bounding boxes of up to twelve images and store the mean and the leading 32 covariance eigenpairs. The pooled within-class covariance is the sum of per-class scatter matrices divided by the total degrees of freedom, and we store its leading 64 eigenpairs. The surrogate constructions are given in Section 3.2. Gaussian mixing coefficients are drawn once per image from a seed derived from the image identifier, making the surrogate deterministic across conditions.

All experiments run on a single NVIDIA RTX 5000 Ada or RTX A5000 GPU. Three models use four-bit precision. LLaVA-OneVision uses half precision because its MLP width of 4304 is not aligned with the block size of the four-bit kernel. Appendix H verifies that this precision difference does not change the mechanism.

## B Baseline Tuning

Every baseline has a strength parameter, so we tune each method independently rather than evaluating it at the strength used for our operator. We report the best injection rate subject to retention of at least 0.35. Text-direction insertion is evaluated at $\alpha \in \{0.25, 0.5, 1, 2, 4\}$ and reported at $\alpha = 1$. Text-direction removal is evaluated at $\beta \in \{1, 2, 5, 10, 20, 40\}$ and reported at $\beta = 20$. Residual-stream steering is evaluated at the same $\alpha$ values. No setting both inserts and retains, so we report $\alpha = 0.25$. Our removal operator is reported at $\beta = 20$.

These baselines re-implement the published intervention primitives and restrict them to the object patches. They are not the complete released systems. VISTA-SAE (Kravets et al., 2026) is the only baseline that requires training. We use the authors' released projector for LLaVA-1.5-7B and apply its replacement operator at the object patches using the SAE feature of the target. We sweep $\{0.25, 0.5, 1, 2, 4\}$ and report strength 1.0. VISTA-SAE reaches 0.273 injection at 0.482 retention. This exceeds the other single-vector operators but remains well below the Gaussian (53:2, $p = 8.6 \times 10^{-14}$), consistent with the account in Section 4.2. Table 6 gives the complete COCO comparison underlying the representative cross-dataset results in Table 2.

## C Removal Strength

Figure 6 shows the complete strength sweep for our removal operator and the tuned text-direction baseline. The object-removal rate saturates near $\beta = 20$, whereas retention degrades more gradually for our operator.

## D Additional Controls and Negative Results

We first test whether the spatial arrangement of substitute activations matters. Permuting them within the object changes the outcome on only 17 of 286 images (17:14, $p = 0.72$). The same

Table 6: Complete COCO operator comparison on the 150-image discovery set. Each method is evaluated at its best strength and paired against our operator. † indicates a method that trains a projector.

| operator | target second-order structure | inject ↑ | retain ↑ | vs. ours |
|---|---|---|---|---|
| concept mean, tiled | none | 0.020 | 0.557 | 89:0, $p = 3 \times 10^{-27}$ |
| GCS direction (Zhao et al., 2025) | none | 0.053 | 0.508 | 85:1, $p = 2 \times 10^{-24}$ |
| residual steering (Li et al., 2025) | none | 0.007 | 0.798 | 74:0, $p = 1 \times 10^{-22}$ |
| text direction (Jiang et al., 2025) | none | 0.320 | 0.535 | 39:12, $p = 2 \times 10^{-4}$ |
| VISTA-SAE† (Kravets et al., 2026) | none | 0.273 | 0.482 | 53:2, $p = 8.6 \times 10^{-14}$ |
| AdaIN (Huang and Belongie, 2017) | diagonal | 0.180 | 0.570 | 65:0, $p = 5 \times 10^{-20}$ |
| WCT, rank 4 (Li et al., 2017) | full, approximate | 0.307 | 0.490 | 47:1, $p = 3 \times 10^{-13}$ |
| WCT, rank 16 (Li et al., 2017) | full, approximate | 0.340 | 0.479 | 42:1, $p = 1 \times 10^{-11}$ |
| real donor object | full, one sample | 0.500 | 0.378 | 28:11, $p = 9.5 \times 10^{-3}$ |
| ours, rank-4 Gaussian | full, pooled | **0.613** | 0.376 | – |

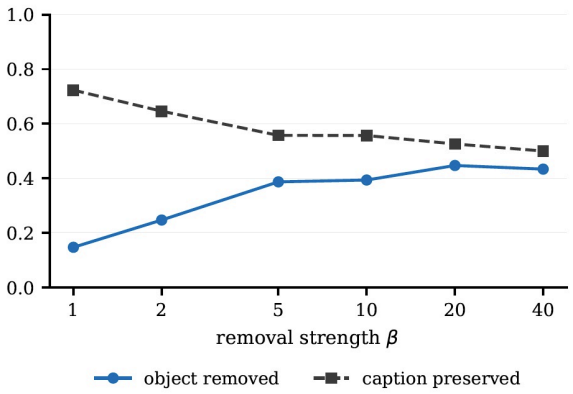


Figure 6: Removal as a function of intervention strength $\beta$. The solid line is the fraction of captions in which the object disappears, and the dashed line is the fraction of the remaining caption content that is preserved. Removal saturates near $\beta = 20$, while retention degrades gradually. The tuned text-direction baseline reaches a higher peak removal rate, but its retention falls from 0.41 at $\beta = 20$ to 0.05 at $\beta = 40$.

null result holds for Mistral-7B (11:16, $p = 0.44$) and Vicuna-13B (9:15, $p = 0.31$). On the same three models, the full donor set strongly outperforms its tiled mean, with discordant counts of 124:1 ($p = 6 \times 10^{-36}$), 97:1 ($p = 6 \times 10^{-28}$), and 121:2 ($p = 1 \times 10^{-33}$), respectively, for $n \geq 259$. This agrees with the observation of Schaumlöffel et al. (2026) that token arrangement inside an object boundary has little effect on reading out a bounding box. We observe the same behavior when writing one in.

The clustered dictionary used for removal partitions the visual stream into $K$ regions. Across three complete rebuilds, $K$ ranges from 227 to 245. Operators that do not use this dictionary have an injection standard deviation of 0.000 across rebuilds. Operators that do use it have standard deviations of 0.016 and 0.017.

Coverage and activation diversity are both necessary. At 25% object coverage, increasing diversity changes injection by 0.067. At full coverage, the same change is 0.373. Injection is approximately linear in coverage, and the absolute box size is only weakly correlated with injection (0.175). Figure 7 shows the full grid.

Across four donor images per class, injection is $0.512 \pm 0.030$. At least one donor succeeds on 78.0% of images, while all four succeed on 22.0%.

We also tested two diagnostics that did not support useful conclusions. First, we asked for the color of an inserted object in an attempt to distinguish a perceptual change from lexical emission. All 132 valid edited responses named a color. However, all 150 responses also named a color when asked about a control concept that the model had explicitly denied was present. The model accepts the presupposition of the question, so this probe does not discriminate the mechanisms of interest. Second, we predicted that the rank at which injection saturates would correlate with the participation-ratio dimension of the donor covariance spectrum. The per-image correlation is $\rho = -0.071$ ($p = 0.51$, $n = 86$). We also predicted a coverage threshold, but a logistic fit does not yield an interpretable midpoint.

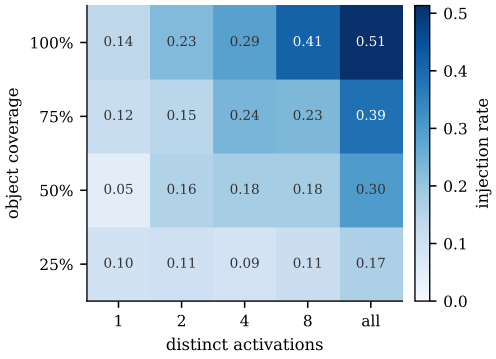


Figure 7: Injection as a function of the proportion of object tokens replaced and the number of distinct donor activations spread over those positions, $n = 150$. Neither coverage nor diversity is sufficient on its own. Increasing diversity changes injection by only 0.07 at 25% coverage but by 0.37 at full coverage.

## E Direct Questioning: Prompts and Susceptibility

Section 4.5 uses the yes/no question *Is there a {target} in this image? Answer yes or no.* We repeat the experiment on the same 150 images and edits with two additional prompts. The inventory prompt, *List every object you can see in this image, separated by commas. Do not describe the scene*, never names the target. A target in the answer must therefore come from the model. We score the target in the list as injection and the unrelated control concept as a specificity error. The forced-choice prompt asks, *Which of these is in the image: a {source}, a {target}, both, or neither? Answer with one word.* This prompt offers both concepts together with a "neither" option and therefore measures removal, insertion, and replacement without presupposing that either object is present.

Table 5 in Section 4.5 compares the three prompts. Free recall is lower than yes/no recognition (0.647 versus 0.880), as expected, but remains far above the unrelated control (0.033). Forced choice also strongly favors the target, with 0.800 target selection and 0.027 control selection. The effect therefore persists without presupposing the target's presence.

Susceptibility varies by concept. Airplane, apple, backpack, banana, baseball bat, and bear are affirmed on every image on which they are inserted. The least susceptible concepts are fork (0.33), chair (0.00), dining table (0.00), and spoon (0.00). Thin implements occupy few patches, while large furniture can have token statistics that overlap with the background. Both explanations are consistent with the coverage experiment, but we do not test them separately.

## F Replication on Held-Out and Out-of-Selection Images

The 400 images satisfying the size criterion are split according to whether the unedited model names the source. Removal can only be scored on the 286 recognized images, so the mechanistic experiments use the first 150 for discovery and the remaining 136 as a held-out set. Insertion can be scored on any image because the target is absent by construction. We therefore also run the final operator on the

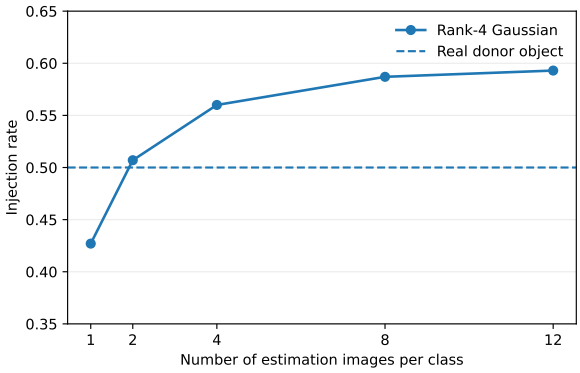


Figure 8: Sample efficiency of the rank-4 concept bank. Injection improves rapidly as the number of images used to estimate each concept increases, reaching 0.560 with four images and 0.593 with twelve. The dashed line shows the real-donor baseline at 0.500. Retention remains nearly constant over the sweep, changing from 0.385 to 0.376.

114 images whose source is not named, using the same target-assignment procedure. Table 7 reports the three splits and their union.

Table 7: Injection of the rank-4 Gaussian on the discovery set, held-out recognized images, images whose source is not named by the unedited model, and their union.

| split | $n$ | mean only ↑ | rank-4 Gaussian ↑ | real donor object ↑ |
|---|---|---|---|---|
| discovery, source recognized | 150 | 0.020 | 0.593 | 0.500 |
| held out, source recognized | 136 | 0.022 | 0.603 | 0.493 |
| source not recognized | 114 | 0.026 | 0.570 | 0.482 |
| all eligible images | 400 | 0.023 | 0.589 | 0.493 |

We also vary the number of images used to estimate each concept while keeping the Gaussian rank fixed at four. Table 8 reports this sample-efficiency experiment. The stored representation remains 20 KB per concept for every setting.

Table 8: Sample efficiency of the rank-4 Gaussian concept bank on the 150-image discovery set. Storage is unchanged as the number of estimation images increases.

| estimation images / class | rank | storage / concept | inject ↑ | retain ↑ |
|---|---|---|---|---|
| 1 | 4 | 20 KB | 0.427 | 0.385 |
| 2 | 4 | 20 KB | 0.507 | 0.381 |
| 4 | 4 | 20 KB | 0.560 | 0.379 |
| 8 | 4 | 20 KB | 0.587 | 0.377 |
| 12 | 4 | 20 KB | **0.593** | 0.376 |
| real donor object | – | one image | 0.500 | 0.378 |

## G Independent Benchmark: LVIS

We evaluate the mechanism on LVIS-v1 (Gupta et al., 2019) to move beyond the 73-class COCO vocabulary. LVIS provides a broader, more fine-grained, and long-tailed set of object categories. We use 300 validation images from 50 classes that do not appear in the COCO evaluation, with six images per class. Concept statistics are estimated independently from up to twelve LVIS training images per class. Targets are absent according to the LVIS annotations. Every setting, including the rank, is frozen from COCO and no hyperparameter is tuned on LVIS.

Table 2 in Section 4.4 compares the representative operators on COCO and LVIS. The ordering from Section 4.3 replicates, and the rank-4 Gaussian gives 0.560 injection compared with 0.483 for a donor

object (44:21, $p = 5.9 \times 10^{-3}$). Injection declines from frequent to rare classes, as shown in Table 9. This indicates that estimate quality still matters in the long tail. Nevertheless, the pooled Gaussian remains above a real donor object in every frequency stratum, and the mean-versus-covariance effect is not specific to COCO.

Table 9: LVIS injection by frequency stratum, with 100 images in each stratum.

| stratum | $n$ | mean only | real donor object | rank-4 Gaussian |
|---|---|---|---|---|
| frequent | 100 | 0.030 | 0.500 | 0.580 |
| common | 100 | 0.030 | 0.490 | 0.570 |
| rare | 100 | 0.030 | 0.460 | 0.530 |
| all | 300 | 0.030 | 0.483 | 0.560 |

## H Robustness of the Distributional Mechanism

The main experiment intervenes at the projector input of LLaVA-1.5-7B, uses four-bit precision, runs through the Hugging Face eager implementation, and uses the model's two-layer MLP projector. We vary these choices one at a time on the same 150 images and target assignments. Table 10 summarizes the results, and Tables 11–13 give the detailed comparisons. In every setting, the tiled mean remains between 0.02 and 0.04 while the Gaussian remains between 0.51 and 0.60. Absolute rates change, but the gap between a mean and a distribution persists.

We first move the intervention point. We intervene at an earlier vision-tower layer, at the projector output after conversion to the language-model embedding space, and in the language-model residual stream after the first transformer block. For each location, we estimate $\mu_c$ and $\Sigma_c$ directly in that representation so the test is not confounded by a coordinate mismatch. At rank eight, the Gaussian exceeds the donor object at all four intervention points. The paired counts from earliest to latest are 29:13 ($p = 0.020$), 29:10 ($p = 3 \times 10^{-3}$), 25:10 ($p = 0.017$), and 22:8 ($p = 0.016$). The absolute rate declines later in the pathway, but the need for a distribution is not specific to the encoder-projector boundary.

We next vary numerical precision and serving implementation. Table 12 reports these comparisons. With the same weights, half and four-bit precision yield 0.600 and 0.593 injection. The Gaussian outcome differs on only nine of 150 images (4:5, $p = 1.0$). Across Hugging Face eager attention, FlashAttention-2, and an independently instrumented vLLM serving path, binary outcomes agree on at least 96% of images and no pairwise difference is significant ($p > 0.5$).

Finally, we vary the projector while holding the vision encoder, language model, and training procedure fixed. Linear, two-layer MLP, and gated MLP projectors all preserve the same qualitative gap. The mean gives 0.02–0.03 injection and the Gaussian gives 0.56–0.59. The second-order requirement therefore originates upstream of the projector architecture.

Table 10: Robustness summary on $n = 150$ images.

| setting | rank-4 Gaussian ↑ | mean only ↑ |
|---|---|---|
| projector input (main setting) | 0.593 | 0.020 |
| earlier vision-tower layer | 0.553 | 0.027 |
| projector output | 0.547 | 0.033 |
| language-model residual, block 1 | 0.507 | 0.040 |
| half precision | 0.600 | 0.020 |
| four-bit precision | 0.593 | 0.020 |
| FlashAttention-2 | 0.587 | 0.020 |
| vLLM serving path | 0.580 | 0.027 |
| linear projector | 0.560 | 0.027 |
| gated MLP projector | 0.573 | 0.027 |

Table 11: Injection at each intervention point, $n = 150$, with concept statistics estimated directly in each representation.

| intervention point | mean only | rank-4 Gaussian | rank-8 Gaussian | real donor object |
|---|---|---|---|---|
| earlier vision-tower layer | 0.027 | 0.553 | 0.580 | 0.480 |
| projector input (main setting) | 0.020 | 0.593 | 0.627 | 0.500 |
| projector output | 0.033 | 0.547 | 0.580 | 0.460 |
| language-model residual, block 1 | 0.040 | 0.507 | 0.553 | 0.433 |

Table 12: Precision and serving implementation for the same LLaVA-1.5-7B weights, $n = 150$.

| setting | mean only | rank-4 Gaussian | real donor object | retention |
|---|---|---|---|---|
| half precision, eager | 0.020 | 0.600 | 0.500 | 0.381 |
| four-bit, eager (main) | 0.020 | 0.593 | 0.493 | 0.376 |
| FlashAttention-2 | 0.020 | 0.587 | 0.493 | – |
| vLLM serving path | 0.027 | 0.580 | 0.493 | – |

## I Human Validation of the Caption Metric

Two annotators who were blind to the condition independently evaluated the 150 edited captions from the rank-4 Gaussian condition. For each caption, they judged whether it described the target concept as present in the scene, which corresponds to injection. Inter-annotator agreement is 99.3% with Cohen's $\kappa = 0.99$. The annotators resolved their disagreement by consensus before comparison with the automatic metric.

Metric and human judgment agree on 148 of 150 captions (98.7%, $\kappa = 0.97$). Table 14 gives the corresponding confusion matrix. Both disagreements are false negatives of the automatic metric. It never credits a caption that humans judge as not describing the target, but it misses two descriptions that avoid the literal class name: a baseball player holding a bat for *baseball bat*, and a glass of wine for *wine glass*. The automatic injection metric is therefore conservative, consistent with the direct-question comparison in Section 4.5.

## J Qualitative Examples

Figure 9 compares the operators from Table 2 on representative images. The target concept is absent from every photograph. A successful insertion therefore requires the edited caption to report a concept that is not in the input image. Removal can succeed with a single direction. Among the insertion methods shown, the Gaussian is the only operator that succeeds on all four examples while integrating the target with the surrounding scene.

Table 13: Projector architecture with vision encoder, language model, and training procedure fixed, $n = 150$.

| projector | mean only | rank-4 Gaussian | rank-8 Gaussian | real donor object |
|---|---|---|---|---|
| linear | 0.027 | 0.560 | 0.587 | 0.480 |
| two-layer MLP (main) | 0.020 | 0.593 | 0.627 | 0.500 |
| gated MLP | 0.027 | 0.573 | 0.600 | 0.487 |

Table 14: Automatic injection score against human judgment on 150 edited captions. Agreement is 98.7% with Cohen's $\kappa = 0.97$.

| | human: target present | human: target absent |
|---|---|---|
| metric: target named | 89 | 0 |
| metric: target not named | 2 | 59 |

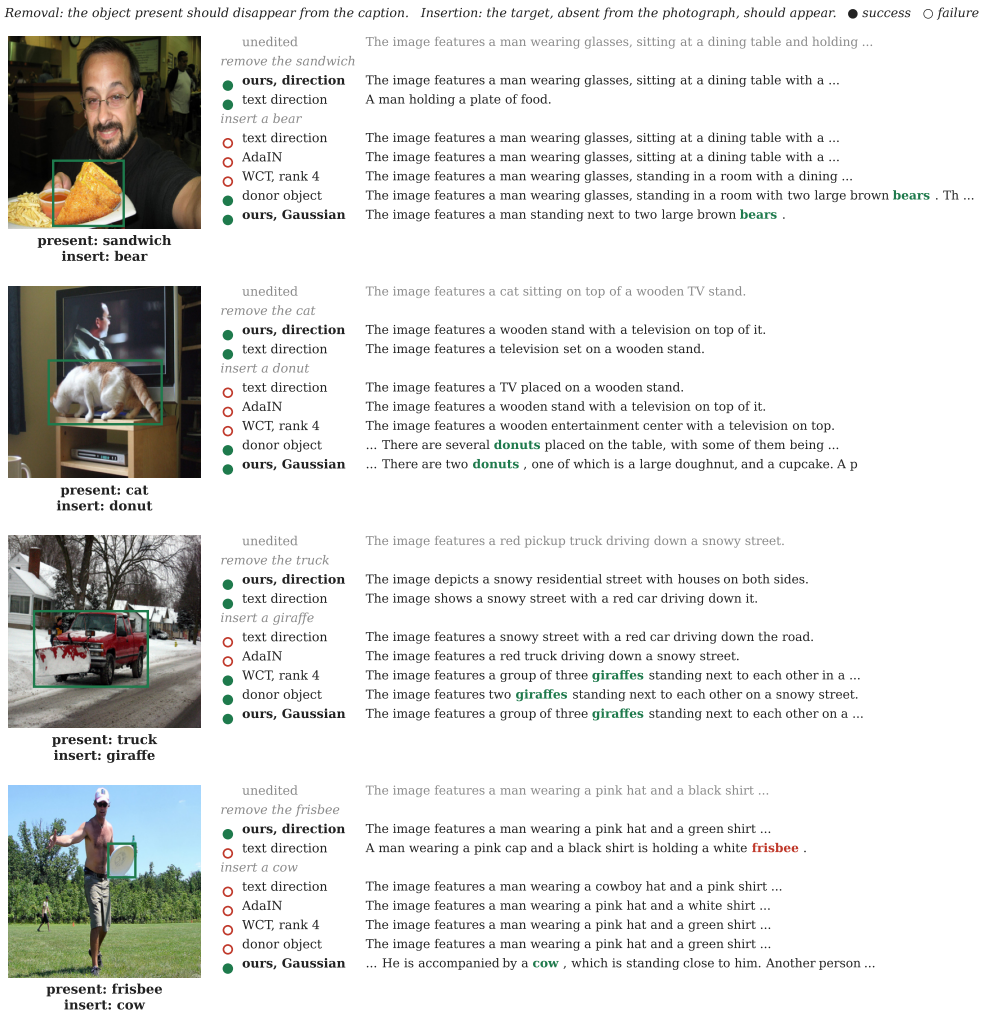


Figure 9: Qualitative comparison of removal and insertion on the same images. Each row shows one COCO image, the source object, and an absent target concept. Filled markers denote success. The removal block compares the object's clustered-dictionary direction with a text-derived direction. The insertion block compares a text direction, AdaIN, WCT, a real donor object, and the rank-4 Gaussian. The Gaussian succeeds on all four examples and integrates the inserted target with the scene.