What Do VLM Language States Remember? Low-Rank Recovery of Instance-Level Visual Evidence

Anonymous authors
Paper under double-blind review

Abstract

Does a multimodal language model retain, in its contextualized language states, evidence pointing to the specific visual instance a referring expression denotes: the dog on the left, not merely dog? We probe this with a capacity-limited linear decoder, asking whether language states can linearly recover the corresponding spatially localized visual evidence from frozen vision representations. Across Qwen3-VL, LLaVA-OneVision, and Llama-3.2-Vision on RefCOCO, a low-rank linear mapping recovers the instance-level correspondence, reaching 88.4%, 88.6%, and 71.7% Pointing, with most gains saturating near rank 32; rank sweeps and cross-modal geometry localize this structure to a few shared dimensions, while native-attention, CLIP, and statistical-projection baselines perform substantially worse. CKA/CCA, semantic mismatch, anchored concept swapping, and unseen-category transfer further show that recovery follows cross-modal structure already present in frozen representations rather than arbitrary bbox associations. Spatially matched suppression and retention show that the recovered evidence has a location-specific causal effect on outputs in Qwen and LLaVA, distinguishing decodable from functionally used information. Token-level analyses show instance specificity emerges only as the expression is contextualized (bike sharpening into bike on the far left), and across architectures the location and strength of recoverability depend on how visual representations are injected into the language stream. Thus, the models retain, in their language states, a compact low-rank structure of linearly recoverable instance-level visual evidence, beyond what native attention reveals.

1 Introduction

Multimodal large language models (VLMs) can describe images and reason about them, yet how visual information persists in the language stream remains unclear (Zhang et al., 2025b; Miao et al., 2026). Given "the man in a blue hat," does the contextualized state encode only man, or also which person and where? This determines whether language reasoning carries instance-level evidence rather than only category semantics.

We distinguish four levels: successful answering, representational presence, decodability, and causal use. Native attention and saliency describe routing but neither exhaust representational content nor establish use (Darcet et al., 2024). Detectors and referring-expression systems output regions (Minderer et al., 2022; Liu et al., 2024), but do not reveal what a frozen VLM's language states already retain.

We hypothesize that VLM language states preserve compact, spatially specific evidence for a particular instance. We freeze the VLM and use a low-capacity linear Local Grounding Bridge to project contextualized target-phrase and pre-merger/projector patch states into a shared low-rank space. Successful recovery through this deliberately restricted interface exposes cross-modal structure already present in the frozen VLM.

Our experiments follow a progression from representation to behavior. RefCOCO grounding first tests whether language states remember a specific instance. Rank sweeps, CKA/CCA, semantic mismatch, anchored concept swapping, and unseen-category transfer characterize the recovered cross-modal structure. Token-wise analysis asks when instance specificity forms; spatially matched inter-

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.

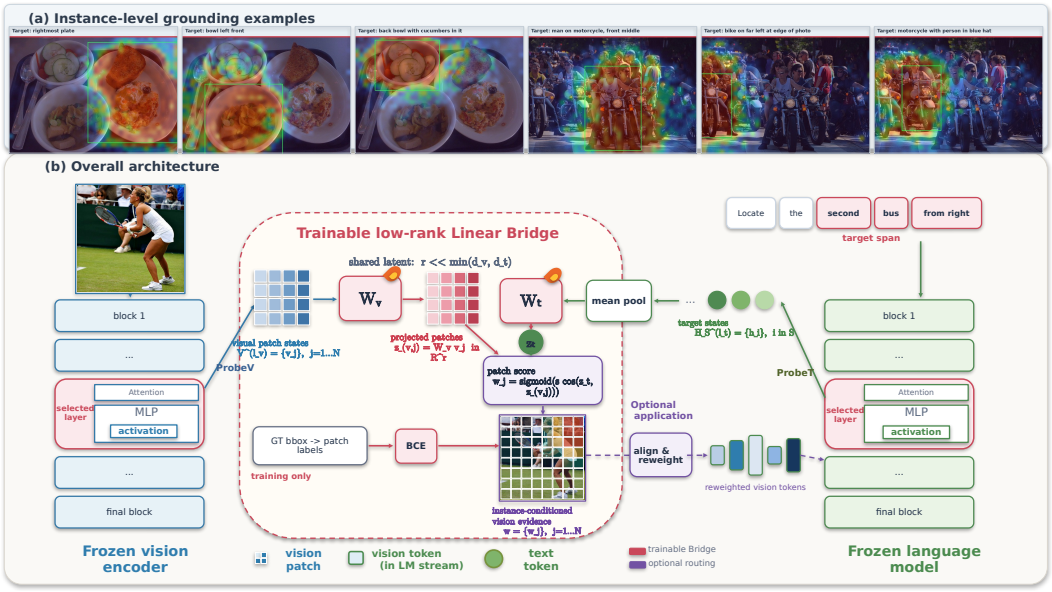

Figure 1: Overview of the core method and architecture. Top: the Bridge retrieves different instances in the same scene according to different target phrases. Bottom: language-entity and visual-patch representations from a frozen VLM are projected into a low-rank space to form local evidence for grounding, causal analysis, and inference modulation.

ventions ask whether the evidence is used; cross-architecture layer scans test how visual injection controls accessibility; and evidence reweighting examines consequences for information flow and local visual reasoning.

Our main contributions are as follows:

- We show across three frozen VLM architectures that contextualized language states retain instance-specific visual evidence recoverable by a deliberately low-capacity linear Bridge.
- Geometry, rank, semantic-perturbation, and transfer analyses show that this evidence occupies a compact cross-modal structure already present in frozen representations; token-wise analysis reveals how it forms in context.
- Spatially matched interventions establish location-specific behavioral use, while architecture and reweighting analyses show how accessibility depends on visual injection and how the latent evidence can alter later information flow.

2 Related Work

Cross-modal representation alignment and local recoverability. Prior work has investigated representation structure in frozen models using linear probes, representation similarity, and subspace analysis, with CKA and SVCCA serving as classical tools for comparisons across layers and networks (Raghu et al., 2017; Kornblith et al., 2019). Recent studies further explore internal vision–language correspondence in multimodal models, including low-complexity mappings from visual representations to language space (Venhoff et al., 2025), shared-concept analysis based on sparse autoencoders (Cornet et al., 2025), and cross-layer stages of image understanding, reasoning, and answer decoding (Yu & Lee, 2025). Rather than global image–text consistency or concept-level shared directions, we examine whether contextualized language entities can recover spatially localized, instance-level visual patch evidence through a low-rank linear mapping.

Explicit visual grounding and referring-expression comprehension. Referring-expression datasets such as RefCOCO(Yu et al., 2016) have advanced the study of contextual relations and instance discrimination. CLIP establishes a cross-modal semantic space through global image–text contrastive

learning, but does not directly constrain correspondence between local patches and referring expressions (Radford et al., 2021). OWL-ViT, GroundingDINO (Minderer et al., 2022; Liu et al., 2024), and methods such as Kosmos-2, Shikra, and Ferret instead learn explicit text-to-region grounding through open-vocabulary detection, location tokens, coordinate generation, or region representations (Peng et al., 2024; Chen et al., 2023; You et al., 2024). These methods train systems that output regions but introduce additional training or inference overhead; we instead freeze the VLM and recover instance-level local correspondence directly from its existing visual and language hidden states. Related studies also use activation transplantation to analyze contextualized token states (Pal et al., 2023; Ghandeharioun et al., 2024), whereas we further examine how referring information is progressively composed within contextualized tokens and forms local visual correspondence.

Location-specific causal explanation and visual evidence routing. Prior work uses activation steering to probe visual concepts and their causal roles in VLMs (Deng et al., 2026); Amnesic Probing further argues that "information is decodable" does not imply "information is used by the model," which requires evidence from counterfactual interventions (Elazar et al., 2021). Another line of work improves fine-grained perception through visual search, cropping, re-encoding, or local enlargement (Wu & Xie, 2024; Zhang et al., 2025a; Chen et al., 2025), but requires additional region selection and visual computation. We recover patch evidence from a frozen VLM, test its location-specific effect through spatially matched interventions, and explore activation-space evidence routing without cropping, thereby connecting local recoverability, causal use, and inference modulation.

3 Low-Rank Cross-Modal Representation Recovery

3.1 Representations and Activation Extraction

Given an image I, a target expression q, and a frozen VLM F, we ask whether the contextualized representation formed by the target expression in the language network can be traced back with fine-grained precision to the spatially ordered patch representations in the vision encoder. Let $\mathcal{T}(q)$ denote the set of tokens corresponding to the target expression after tokenization, and let $\{\boldsymbol{h}_i^{(\ell_t)}\}_{i\in\mathcal{T}(q)}$ denote their activations at language layer ℓ_t. The text representation is obtained by mean pooling:

$$\boldsymbol{t} = \frac{1}{|\mathcal{T}(q)|}\sum_{i\in\mathcal{T}(q)} \boldsymbol{h}_i^{(\ell_t)}. \tag{1}$$

Here, $\boldsymbol{h}_i^{(\ell_t)}$ is a contextualized state from the complete input, rather than a static word representation obtained by encoding the word in isolation. We take the pre-merger or pre-projector patch activations at layer ℓ_v, denoted by $\{\boldsymbol{v}_j^{(\ell_v)}\}_{j=1}^{N}$. These representations correspond to two-dimensional patches, allowing us to intersect the GT bbox with the grid to obtain patch labels $y_j \in \{0, 1\}$.

3.2 Low-Rank Local Grounding Bridge

The Bridge consists of one linear projection for the text tower and one for the vision tower:

$$\boldsymbol{z}_t = \boldsymbol{W}_t\boldsymbol{t}, \qquad \boldsymbol{z}_{v,j} = \boldsymbol{W}_v\boldsymbol{v}_j, \tag{2}$$

$$a_j = s\cdot\frac{\boldsymbol{z}_t^\top\boldsymbol{z}_{v,j}}{\|\boldsymbol{z}_t\|_2\|\boldsymbol{z}_{v,j}\|_2}, \qquad w_j = \sigma(a_j), \tag{3}$$

The shared dimension r is the Bridge rank, s is a learnable logit scale, and w_j is the evidence score for visual patch j. All VLM parameters remain frozen during training, and only the Bridge is optimized with patch-level binary cross entropy. Because visual and language states differ in dimensionality and coordinates, the two projections provide the minimal interface needed for comparison. Their restricted rank limits the probe's capacity to relearn visual semantics, making successful recovery evidence of a linearly accessible correspondence in the original representations.

3.3 Spatially Matched Causal Interventions

To distinguish "decodable information" from "information actually used by the model," we align patch scores with the corresponding vision tokens in the LLM and intervene on their activations $\boldsymbol{h}_j$

Table 1: Formal RefCOCO grounding (3,421 expressions over 1,500 images). IoU uses threshold 0.6; brackets following Pointing contain image-cluster bootstrap 95% confidence intervals.

Method	Pointing	Hit@5	Hit@10	IoU
Qwen Linear-64 Bridge	**88.42 [87.25,89.57]**	**96.20**	**97.14**	**52.07**
LLaVA Linear-64 Bridge	**88.60 [87.46,89.71]**	**95.12**	**96.40**	**52.61**
Llama Linear-64 Bridge	**71.70 [70.07,73.38]**	**85.41**	**89.36**	**39.24**
GroundingDINO-T	59.08 $[57.43, 60.66]$	64.63	68.49	51.23
OWL-ViT	48.18 $[46.47, 49.85]$	54.02	57.21	35.81
CLIP ViT-L/14@336	16.78 $[15.50, 18.10]$	43.70	58.73	13.52

at the input to a designated language layer. The three operations are

$$\text{Suppression:} \quad \boldsymbol{h}'_j = (1 - \gamma w_j)\boldsymbol{h}_j, \tag{4}$$

$$\text{Retention:} \quad \boldsymbol{h}'_j = (f + (1 - f)w_j)\boldsymbol{h}_j, \tag{5}$$

$$\text{Reweight:} \quad \boldsymbol{h}'_j = (1 + \beta w_j)\boldsymbol{h}_j. \tag{6}$$

Suppression tests the necessity of evidence by attenuating high-scoring patches; Retention suppresses the background while preserving high-scoring patches to test their sufficiency; and Reweight amplifies the corresponding vision tokens to increase subsequent attention. The key control is a matched random map obtained by permuting the score vector w across locations within the same image. Bridge and random use the same score set, intervention magnitude, and number of tokens, differing only in score–patch spatial correspondence. Thus, $\mathrm{Acc}_{random} - \mathrm{Acc}_{bridge}$ under Suppression and $\mathrm{Acc}_{bridge} - \mathrm{Acc}_{random}$ under Retention measure location-specific causal effects. The oracle bbox is included only as a reference and is not treated as a strict upper bound because its sparsity is not strictly matched to the Bridge map.

4 Experiments and Results

Experimental setup. We conduct experiments on frozen Qwen3-VL-4B, LLaVA-OneVision-1.5-4B, and Llama-3.2-11B-Vision-Instruct (Bai et al., 2025; An et al., 2025; Meta AI, 2024), using Linear-64 Bridges at T19–V6, T24–V11, and T15–V14, respectively. All three Bridges are jointly trained on RefCOCO, RefGQA, and Visual Genome, split by image ID, with ratios of $0.4/0.3/0.3$. Formal grounding uses 3,421 expressions over 1,500 images from RefCOCO testA/testB, where testA contains only person targets and testB only non-person targets; causal and application experiments use V* and RefCOCO-QA, V*, ADE20K Count, and POPE-random, respectively.

We study `given target phrase` $\rightarrow$ `local visual evidence` to isolate target-routing error. Baselines include Native Attention, CLIP, GroundingDINO-T, OWL-ViT, PCA+CCA, and Random Procrustes. Grounding reports Pointing, Hit@k, and patch IoU ($\tau = 0.6$) with image-cluster bootstrap; V* causal results use paired bootstrap and McNemar tests. Applications use fixed Bridge-centered configurations: Qwen L11 ($\beta = .9$) and LLaVA L9 ($\beta = .7$), without formal-set retuning. Appendix A.1 gives the full protocol.

4.1 Instance-Level Visual Evidence in Language States

In Table 1, the Linear-64 Bridge reaches 88.42%, 88.60%, and 71.70% Pointing on Qwen, LLaVA, and Llama, respectively. Narrow confidence intervals and consistent trends on testA/testB show that the result is not driven by a few images or a single person split. Llama's lower score may reflect a representational interface mismatch between a single vision layer and periodic cross-attention fusion; Section 4.5 and Appendix A.3 provide the fuller architectural analysis.

Pointing tests the maximum-response location, Hit@k measures top-k target coverage, and IoU evaluates regional overlap. GroundingDINO has relatively high IoU, but Appendix Figure 12 shows simultaneous responses to multiple similar or same-category objects, preventing reliable unique reference. Llama's lower Pointing but 89.36% Hit@10 instead suggests evidence distributed across candidate patches. CLIP reaches only 16.78% Pointing, while Table 10 provides the native-attention control. Thus, language states retain spatial detail that distinguishes instances and provides information complementary to native attention and global alignment spaces.

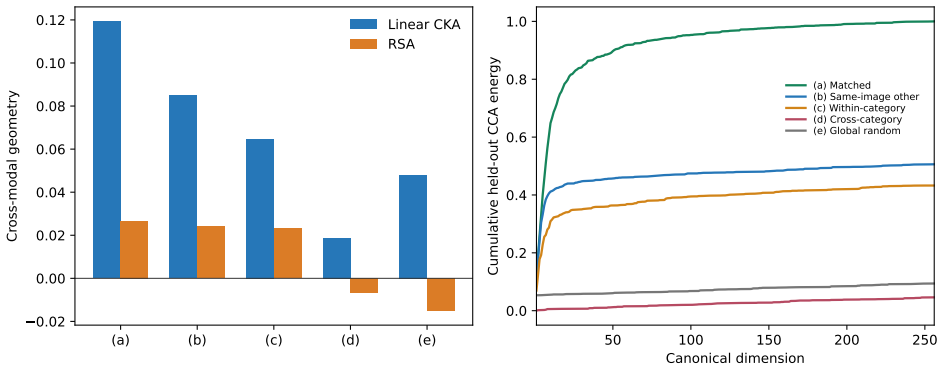

Figure 2: Cross-modal geometry of frozen representations. Left: Linear CKA and RSA. Right: the image-disjoint held-out CCA spectrum normalized by the total matched energy. (a) matched target, (b) same-image other instance, (c) within-category permutation, (d) cross-category permutation, and (e) global random span–region.

Table 2: Local person–dog concept swapping (%). Balanced Pointing uses 25 expressions per category; directional columns report exclusive GT-box hits in co-occurrence images; the bold second row is the anchored experiment.

Training condition	Person–dog pairing	Balanced Pointing	P→P	P→D	D→D	D→P
Anchored	Correct	90.0	81.6	0.0	70.0	0.0
Anchored	**Swapped**	**88.0**	**81.6**	**5.3**	**70.0**	**0.0**
Pair-only	Correct	74.0	63.2	7.9	80.0	0.0
Pair-only	Swapped	30.0	21.1	15.8	10.0	60.0

4.2 Cross-Modal Structure and Supervision Dependence

To separate native vision–language structure from bbox-induced associations, we combine CKA/CCA (Kornblith et al., 2019; Raghu et al., 2017), semantic mismatch, anchored concept swapping, category-held-out transfer, and supervision-size controls. In Figure 2, matched pairs have the highest CKA and held-out CCA energy; same-image other instances and within-category permutations retain 50.60% and 43.26%, but cross-category and global-random conditions retain only 4.61% and 9.40%. Thus, although the spaces do not coincide globally, samples denoting the same instance share more consistent relative geometry.

Table 2 shows that, with other categories as anchors, swapping person and dog reduces Pointing only from 90.0% to 88.0%, and both phrases still primarily localize their original categories. Without anchors, the same swap reduces Pointing from 74.0% to 30.0%, with 60.0% of dog phrases localizing person. Incorrect labels can distort an unconstrained mapping but do not override geometry jointly anchored by other categories. Moreover, unseen-category Pointing remains 84.91% after excluding one quarter of the categories, ruling out category-template memorization. Bridge grounding therefore recovers cross-category linear semantic structure rather than arbitrary supervised associations; Appendix A.2 gives full results.

The low-rank result reflects learned cross-modal recovery rather than dimensionality alone: with the same activations, samples, and 32-dimensional space, PCA+CCA and Random Procrustes achieve AP values of 23.40% and 19.20%, whereas Linear Bridge reaches 74.03% (Appendix Table 9). Figure 3 shows that Qwen, LLaVA, and Llama already reach three-source macro-average AUROCs of 88.61%, 88.78%, and 88.04% at rank 8; most gains occur from rank 8 to 32, after which performance saturates. Consistently, in the held-out CCA spectrum of correctly paired instances Figure 2, the first 8 and 32 directions already account for approximately 56% and 85% of the cross-modal correlation energy, showing geometrically that the recoverable correspondence is concentrated in a small number of shared directions. Complete layerwise results are provided in Appendix Figure 10.

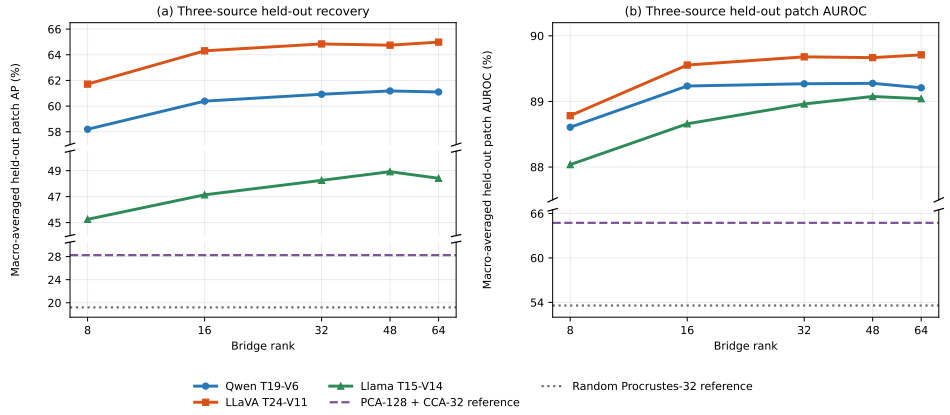

Figure 3: Low-rank recovery curves. Macro-average AP (left) and AUROC (right) for different Bridge ranks on image-disjoint RefCOCO, RefGQA, and VG validation sets. Horizontal dashed lines denote the PCA+CCA and Random Procrustes baselines.

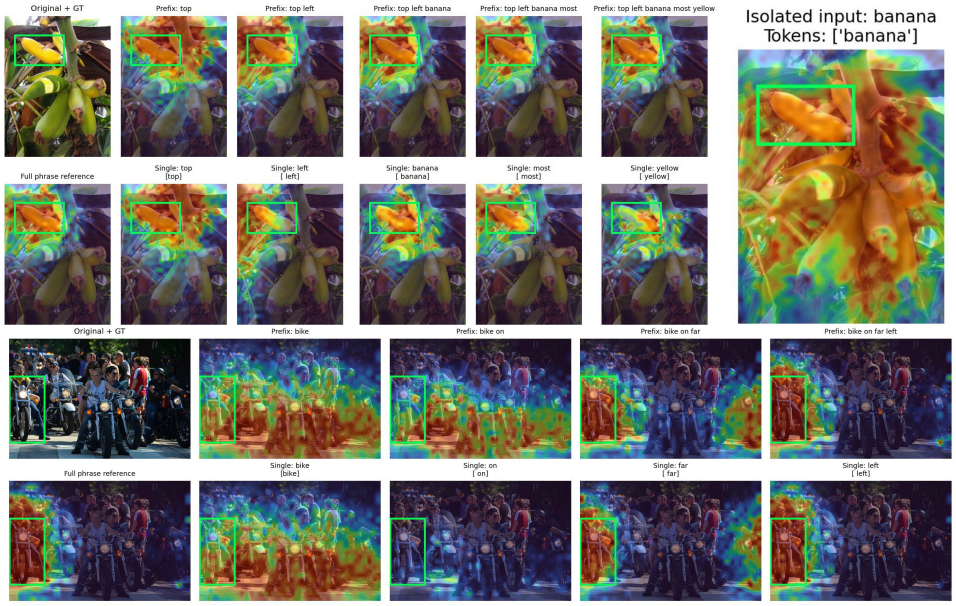

Figure 4: Contextual formation of referring representations. Rows show incremental-prefix pooling and contextualized single-token representations; the far-right panel uses banana in isolation.

4.3 Contextual Formation of Referring Representations

We compare the complete phrase, incremental prefixes, and single-token activations from the complete input on RefCOCO expressions. All heatmaps share the raw Bridge probability scale, avoiding visually incomparable results caused by per-image min–max normalization.

Figure 4 shows how referring representations become contextualized. In "top left banana most yellow," the individual banana token from the complete input inherits its left context and responds mainly to the upper-left target; when banana is forwarded in isolation outside the original sentence, its response expands to all bananas in the image. In "bike on far left," both the incremental prefix and single-token analyses initially show category-level localization: bike responds to multiple bicycle instances. After far is added, the response shifts toward distant bicycles, and after left it rapidly converges on the target instance. The individual left token in the complete context also produces localization close to that of the full phrase, whereas the insufficiently contextualized preposition on

Table 3: Spatially matched causal interventions on V*. Primary effects are Random−Bridge under suppression and Bridge−Random under retention. Qwen/LLaVA/Llama intervene at L11/L9/L13.

Model	Operation	Baseline	Bridge	Random	Primary difference
Qwen	Suppress, $\gamma = .7$	86.97	71.85	87.39	**15.55**
Qwen	Retain, $f = .2$	86.97	84.03	60.50	**23.53**
LLaVA	Suppress, $\gamma = .7$	64.29	55.88	65.13	**9.24**
LLaVA	Retain, $f = .2$	64.29	66.39	48.32	**18.07**
Llama	Suppress, $\gamma = .7$	51.68	49.16	52.52	**3.36**
Llama	Retain, $f = .2$	51.68	52.94	53.78	-0.84

shows no clear local response. This indicates that single-token activations encode not only isolated lexical meaning but also preceding semantics, forming compositional representations capable of distinguishing instances once key qualifying information appears. Quantitative results further support this progression from category-level response to instance-level localization: as key qualifiers enter the context, agreement between the prefix and complete-phrase maps rises rapidly. Complete Pointing, IoU, and Pearson results are provided in Appendix Table 11.

After the third or fourth word enters context, correlations between prefix and complete-phrase maps reach approximately 0.90–0.96, showing that this convergence extends beyond the visualized cases.

4.4 Location-Specific Causal Effects of Bridge Evidence

Suppression attenuates high-scoring patches to test their necessity, whereas Retention suppresses the background while preserving those patches to test their relative sufficiency. Bridge and matched random share the same scores, token count, and total intervention strength, differing only in spatial assignment. Random−Bridge under Suppression and Bridge−Random under Retention therefore isolate the behavioral effect of where an intervention is applied. In Table 3, the Qwen/LLaVA differences are 15.55/9.24 pp for Suppression and 23.53/18.07 pp for Retention. Paired bootstrap and McNemar tests support location specificity, showing that Bridge-ranked locations contribute to model behavior beyond indiscriminate scaling. Llama's single-layer effects are less stable: after an intervention at one cross-attention layer, later periodic cross-attention layers can reread the unmodified visual state and compensate for or override the change. This distributed retrieval explains the weak single-layer effect, while RefCOCO grounding independently establishes Llama's cross-modal consistency (Figure 11); Appendix A.8 reports the complete scans.

4.5 Layerwise Effects of Visual-Injection Architectures

We prescreen language layers with a token-wise object probe and vision layers with target–background patch separation, then evaluate sparse Bridges between candidate layers. For target patches B and background $\bar{B}$, separation is $\mathbb{E}_{i,j\in B}\cos(\mathbf{v}_i, \mathbf{v}_j) - \mathbb{E}_{i\in B, j\in\bar{B}}\cos(\mathbf{v}_i, \mathbf{v}_j)$. These single-modality scores only select candidates; the Bridge grid and external grounding determine the final pairs.

Figure 5 characterizes layer selectivity from three complementary views: the language-side object probe measures entity decodability, vision-side separation measures target–background discriminability, and the Bridge matrix tests linear compatibility between particular layer pairs. Their broad agreement indicates that high-value regions in the matrix reflect how information evolves within both modalities rather than isolated layer-pair fluctuations.

The architectures nevertheless produce distinct patterns. LLaVA inserts vision tokens through a single merger; its language and vision curves rise and fall smoothly, and the Bridge matrix changes continuously with an early-to-middle high-value region. Qwen's multi-stage DeepStack injection produces repeated fluctuations. V17 is the final injection stage, where the corresponding matrix column drops sharply, indicating that newly injected features reorganize the existing representation and its linear compatibility with language states. Llama instead rereads a separate visual state through periodic cross-attention: text-layer accessibility varies more smoothly and forms a broad middle-stage plateau around cross-attention layers such as T13/T18 rather than a sharp peak. Grounding establishes the presence of instance-level evidence, while these scans determine where it remains

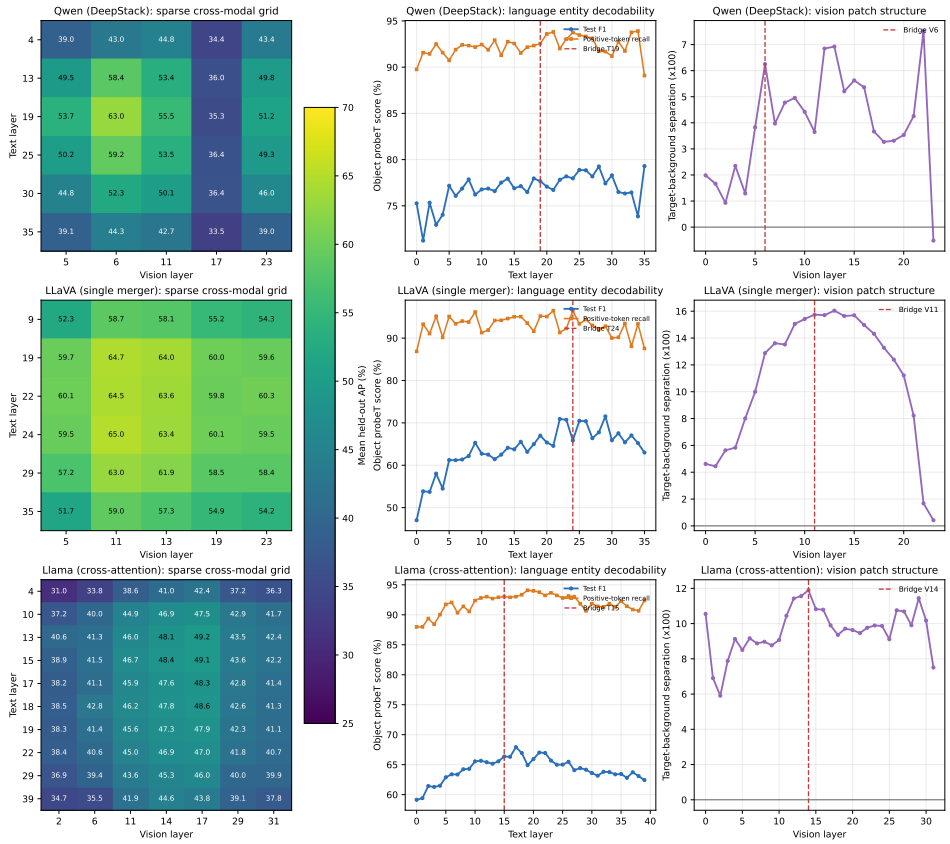

Figure 5: Accessibility across three visual-injection architectures. From left: sparse Bridge layer matrices, language-layer object probes, and vision-layer target–background separation; dashed lines mark selected layers.



Figure 6: Bridge-guided correction on RefCOCO-QA Small. Columns show the input/baseline, pre-reweight attention, Bridge evidence, attention change, and corrected output (Qwen L11, $\beta = .9$).

accessible and how it is distributed across depth under different visual-injection mechanisms. Appendix A.3 provides complete layerwise results.

4.6 Evidence Propagation and Downstream Applications

To test whether the recovered local evidence affects internal information flow, we reweight vision tokens at the input to a language layer according to Bridge scores and trace subsequent target-to-vision native attention. Changes are most pronounced in the first 1–5 layers after intervention: after injection at L13, Qwen's L16 Pointing improves by 8.0 pp; after injection at L9, LLaVA's L11 Pointing improves by 4.03 pp. This shows that Bridge reweighting not only changes the magnitude of vision tokens but also concentrates subsequent attention more strongly on the target region.

Table 4: Application results under fixed configurations. RefCOCO-QA contains 600 reviewed questions; ADE20K Count is a CV-Bench subset; POPE uses the random split.

Model	Dataset	Baseline	Bridge-guided	Δ
Qwen	RefCOCO-QA	77.67	79.33	**+1.67**
Qwen	small challenge	70.37	74.07	**+3.70**
Qwen	V*	86.97	89.08	**+2.10**
Qwen	ADE20K Count	56.43	59.65	**+3.22**
Qwen	POPE-random	91.37	91.90	**+0.53**
LLaVA	RefCOCO-QA	73.00	75.00	**+2.00**
LLaVA	small challenge	72.22	75.00	**+2.78**
LLaVA	V*	64.29	64.71	**+0.42**
LLaVA	ADE20K Count	57.60	59.06	**+1.46**
LLaVA	POPE-random	90.53	90.70	**+0.17**

Figure 6 shows how this internal change further affects model output. In this RefCOCO-QA Small example, the baseline incorrectly answers that the shirt color of the person referred to by "woman right glasses" is white. With the final frozen Qwen configuration applying centered reweighting at L11, the increment in L15 attention concentrates on the target person, and the answer changes from the incorrect option C to the correct option A (gray). This case connects Bridge evidence, changes in the model's native information flow, and the final answer correction; additional attribute, spatial-relation, and V* cases are shown in Appendix Figure 13.

To test whether this case generalizes, we evaluate the fixed configurations above on local visual tasks; Appendix A.8 details their selection. Table 4 shows measurable gains from simple scalar reweighting. Qwen/LLaVA improve by 1.67/2.00 pp on RefCOCO-QA and 3.70/2.78 pp on its Small subset; Qwen further gains 2.10 pp on V* and 3.22 pp on ADE20K Count. Uniform/random controls and corrected-versus-corrupted counts rule out indiscriminate scaling as the explanation. These results connect the recovered representational structure to downstream information flow and task behavior under fixed configurations.

5 Conclusion and Discussion

Conclusion. Using a deliberately low-capacity linear Bridge, we find that frozen VLM language states retain a compact trace of the local visual evidence for a specific instance. Grounding establishes that this trace exists across three architectures; CKA/CCA and rank saturation show that it concentrates in a few cross-modal directions; and semantic mismatch, anchored concept swapping, and category-held-out transfer show that recovery follows structure already present in frozen representations rather than arbitrary bbox associations.

The trace is neither static nor uniformly accessible. Token-wise analysis shows category information narrowing into a unique visual reference as qualifiers accumulate, while layer scans tie accessibility to visual-injection architecture. Spatially matched suppression/retention further establishes location-specific behavioral use in Qwen and LLaVA, and reweighting shows that the same latent evidence can redirect later attention and sometimes correct answers. Together, the experiments distinguish four stages—existence, compact structure, contextual formation, and behavioral use—and make the internal life cycle of instance-level visual evidence measurable.

Discussion and limitations. Building on these findings, two directions remain open. First, visual question answering still requires manually specified target tokens; the system cannot automatically identify entities in full questions or route corresponding visual evidence. Future work could combine token-level probing with automatic span selection. Second, we verify low-rank linear recoverability only at instance and local-patch levels; whether abstract semantics such as events, values, compositional concepts, and reasoning states exhibit similar shared geometry remains open.

AI use statement

Generative AI tools were used in this work to assist with brainstorming, generating and modifying code, summarizing relevant literature, and drafting selected parts of the manuscript. Generative AI was also used to generate candidate questions and answers for the RefCOCO-QA evaluation set; all samples were manually reviewed and filtered by the authors. All AI-generated or AI-modified code was inspected and tested by the authors. The reported results were obtained from independently executed experiments and verified against saved predictions and evaluation records. Relevant literature and bibliographic information were manually checked against the original publications. AI was not used to formulate or prove formal mathematical claims or to originate the research methodology or conceptual framework; the remaining required disclosure categories are not applicable to this work. The authors reviewed all AI-assisted content and take full responsibility for the final text, claims, code, data, figures, and experimental conclusions presented in this paper.

Reproducibility statement

Appendix A.1 details the model and Bridge configurations, activation extraction and token–patch alignment procedures, training hyperparameters, data splits, random seeds, intervention settings, and statistical testing protocols. The fixed evaluation sets, data-split hashes, model checkpoints, and per-sample results are provided in the supplementary materials. The complete implementation, including activation hooks, patch restoration, Bridge training, and evaluation pipelines, is available through the anonymous code repository: `https://anonymous.4open.science/r/bridge-review-artifact-a7f3-7428`.

References

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Chunsheng Wu, Huajie Tan, Chunyuan Li, Jing Yang, Jie Yu, Xiyao Wang, Bin Qin, Yumeng Wang, Zizhen Yan, Ziyong Feng, Ziwei Liu, Bo Li, and Jiankang Deng. LLaVA-OneVision-1.5: Fully Open Framework for Democratized Multimodal Training. arXiv preprint arXiv:2509.23661, 2025. URL https://arxiv.org/abs/2509.23661.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL Technical Report. arXiv preprint arXiv:2511.21631, 2025. URL https://arxiv.org/abs/2511.21631.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing Multimodal LLM's Referential Dialogue Magic. arXiv preprint arXiv:2306.15195, 2023. URL https://arxiv.org/abs/2306.15195.

Shiqi Chen, Tongyao Zhu, Ruochen Zhou, Jinghan Zhang, Siyang Gao, Juan Carlos Niebles, Mor Geva, Junxian He, Jiajun Wu, and Manling Li. Why is spatial reasoning hard for VLMs? an attention mechanism perspective on focus areas. In Proceedings of the 42nd International Conference on Machine Learning, volume 267 of Proceedings of Machine Learning Research, pp. 9910–9932. PMLR, 2025. URL https://proceedings.mlr.press/v267/chen25cr.html.

Clément Cornet, Romaric Besançon, and Hervé Le Borgne. Explaining How Visual, Textual and Multimodal Encoders Share Concepts. arXiv preprint arXiv:2507.18512, 2025. URL https://arxiv.org/abs/2507.18512.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In The Twelfth International Conference on Learning Representations, 2024. URL https://openreview.net/forum?id=2dnO3LLiJ1.

Zehao Deng, Tianjie Ju, Zheng Wu, Liangbo He, Jun Lan, Huijia Zhu, Weiqiang Wang, and Zhuosheng Zhang. Causal Probing for Internal Visual Representations in Multimodal Large Language Models. arXiv preprint arXiv:2605.05593, 2026. URL https://arxiv.org/abs/2605.05593.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. Transactions of the Association for Computational Linguistics, 9:160–175, 2021. doi: 10.1162/tacl_a_00359. URL https://aclanthology.org/2021.tacl-1.10/.

Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. Patchscopes: A unifying framework for inspecting hidden representations of language models. In Proceedings of the 41st International Conference on Machine Learning, volume 235 of Proceedings of Machine Learning Research, pp. 15466–15490. PMLR, 2024. URL https://proceedings.mlr.press/v235/ghandeharioun24a.html.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In Proceedings of the 36th International Conference on Machine Learning, volume 97 of Proceedings of Machine Learning Research, pp. 3519–3529. PMLR, 2019. URL https://proceedings.mlr.press/v97/kornblith19a.html.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In Computer Vision – ECCV 2024, pp. 38–55. Springer, 2024.

Meta AI. Llama 3.2: Revolutionizing Edge AI and Vision with Open, Customizable Models. Meta AI Blog, 2024. URL https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/. Accessed 2026-09-04.

Yanting Miao, Yutao Sun, Dexin Wang, Mengyu Zhou, Pascal Poupart, Lei Lv, Qi Zhao, Li Wang, Hao Li, Xiaoxi Jiang, and Guanjun Jiang. Fill the GAP: A Granular Alignment Paradigm for Visual Reasoning in Multimodal Large Language Models. arXiv preprint arXiv:2605.12374, 2026. URL https://arxiv.org/abs/2605.12374.

Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, Xiao Wang, Xiaohua Zhai, Thomas Kipf, and Neil Houlsby. Simple open-vocabulary object detection with vision transformers. In Computer Vision – ECCV 2022, pp. 728–755. Springer, 2022. doi: 10.1007/978-3-031-20077-9_42.

Koyena Pal, Jiuding Sun, Andrew Yuan, Byron Wallace, and David Bau. Future lens: Anticipating subsequent tokens from a single hidden state. In Jing Jiang, David Reitter, and Shumin Deng (eds.), Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL), pp. 548–560, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.conll-1.37. URL https://aclanthology.org/2023.conll-1.37/.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. In The Twelfth International Conference on Learning Representations, 2024.

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 Proceedings of the 38th International Conference on Machine Learning, volume 139 of Proceedings of Machine Learning Research, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/hash/dc6a7e655d7e5840e66733e9ee67cc69-Abstract.html.

Constantin Venhoff, Ashkan Khakzar, Sonia Joseph, Philip Torr, and Neel Nanda. How visual representations map to language feature space in multimodal LLMs. In 4th Explainable AI for Computer Vision Workshop at CVPR, 2025. URL https://ora.ox.ac.uk/objects/uuid:ba02b60f-1d30-4a28-98ce-425c7d1f0395.

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal LLMs. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13084–13094, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Wu_V_Guided_Visual_Search_as_a_Core_Mechanism_in_Multimodal_CVPR_2024_paper.html.

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. In The Twelfth International Conference on Learning Representations, 2024. URL https://openreview.net/forum?id=2msbbX3ydD.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling context in referring expressions. In Computer Vision – ECCV 2016, pp. 69–85. Springer, 2016. doi: 10.1007/978-3-319-46475-6_5.

Zhuoran Yu and Yong Jae Lee. How multimodal LLMs solve image tasks: A lens on visual grounding, task reasoning, and answer decoding. In Conference on Language Modeling, 2025. URL https://openreview.net/forum?id=kjNJYWvfPA.

Jiarui Zhang, Mahyar Khayatkhoei, Prateek Chhikara, and Filip Ilievski. MLLMs know where to look: Training-free perception of small visual details with multimodal LLMs. In The Thirteenth International Conference on Learning Representations, 2025a. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/aaa0ac4253da75faf9b0dc0dda062612-Abstract-Conference.html.

Xinjie Zhang, Jintao Guo, Shanshan Zhao, Minghao Fu, Lunhao Duan, Guo-Hua Wang, Qing-Guo Chen, Zhao Xu, Weihua Luo, and Kaifu Zhang. Unified multimodal understanding and generation models: Advances, challenges, and opportunities. arXiv preprint arXiv:2505.02567, 2025b. URL https://arxiv.org/abs/2505.02567.

A Supplementary Experiments and Implementation Details

A.1 Complete Experimental Protocol and Implementation Details

Models, data, and fixed configurations. The three VLMs remain frozen throughout. Bridge layer combinations are first selected using internal validation sets and sparse layer scans, and are then evaluated on an external RefCOCO test set. Each training source is split by image ID into a 95% training set and a 5% internal validation set to prevent leakage between expressions from the same image. The formal RefCOCO set retains all expressions from each selected image and uses image-cluster bootstrap. RefCOCO-QA contains 600 manually reviewed four-choice questions, including 266 attribute, 220 position, and 114 count questions; the standard cohort contains 482 questions, and the small challenge contains 118 targets occupying 1%–5% of the image area. Intervention layers for the application configurations are determined by matched suppression/retention scans on V*, while the strength and centered mode are selected from development-stage cross-task sensitivity experiments; no further tuning is performed on the formal datasets.

Target phrases, comparison methods, and metrics. RefCOCO uses the official referring expression, V* uses the official target object, RefCOCO-QA preserves the referring expression verbatim in each question, and POPE parses the object phrase from a fixed question template. Comparison methods include direct internal model readouts, specialized open-vocabulary localizers, statistical projections with matched shared dimensionality, and Linear Bridges at different layers and ranks. Pointing Accuracy is defined by whether the center of the highest-scoring patch falls inside the GT bbox, while Hit@k indicates whether at least one of the top-k patches hits the bbox; region performance is measured by patch IoU after binarization at $\tau = 0.6$. RefCOCO reports 95% CIs from 5,000 image-cluster bootstrap samples, V* reports paired bootstrap and McNemar tests, and RefCOCO-QA additionally reports per-question paired results for the ordinary Bridge, Bridge-centered, matched uniform, and spatially shuffled random conditions. Development-set, predefined-subset, and formal full-set results are labeled separately.

For Qwen, visual features are spatially ordered MLP activations before the Patch Merger; for LLaVA, visual features are spatial patch activations with the CLS token removed before the projector. On the language side, all models use the MLP activations corresponding to the target span in the complete multimodal input. Before training, each bbox is rasterized into a patch mask according to the model's native patch ordering. When Bridge visual evidence is used to intervene on LLM vision tokens, each Qwen merger token corresponds to a genuine 2×2 patch block. Bridge scores are aggregated only within that block using mean or max; no 1×4 reordering or interpolation is used to create additional spatial resolution. The primary Bridges use rank 64, patch-level BCE, 64 entities/pairs per batch, and fixed source sampling; the VLMs remain frozen throughout.

Table 5: Training configurations of the Linear-64 Bridges used in the main text. All activations are extracted offline from frozen VLMs, and all three models use the same data mixture, sampling strategy, loss, and optimization schedule.

Configuration	Qwen	LLaVA-OneVision	Llama-3.2-Vision
Text/vision layers	T19 / V6	T24 / V11	T15 / V14
Input dimensions (text/vision)	2560 / 1024	2560 / 1024	4096 / 1280
Shared rank	64		
Trainable parameters	229,378 (about 0.23M)	229,378 (about 0.23M)	344,066 (about 0.34M)
Data-source ratios	RefCOCO / RefGQA / VG = 0.4 / 0.3 / 0.3		
Training length	80 epochs $\times$ 240 steps		
Batch and patch sampling	64 entities; 8 patches per entity; 50% positive and 50% negative		
Small/tiny target sampling	0.50 / 0.15; area thresholds 5% / 1%		
Objective	patch-level binary cross entropy		
Optimizer and learning rate	AdamW; 10^{-3}, then 5×10^{-4} after epoch 45		
Held-out split and random seed	image-disjoint 5%; seed 0		

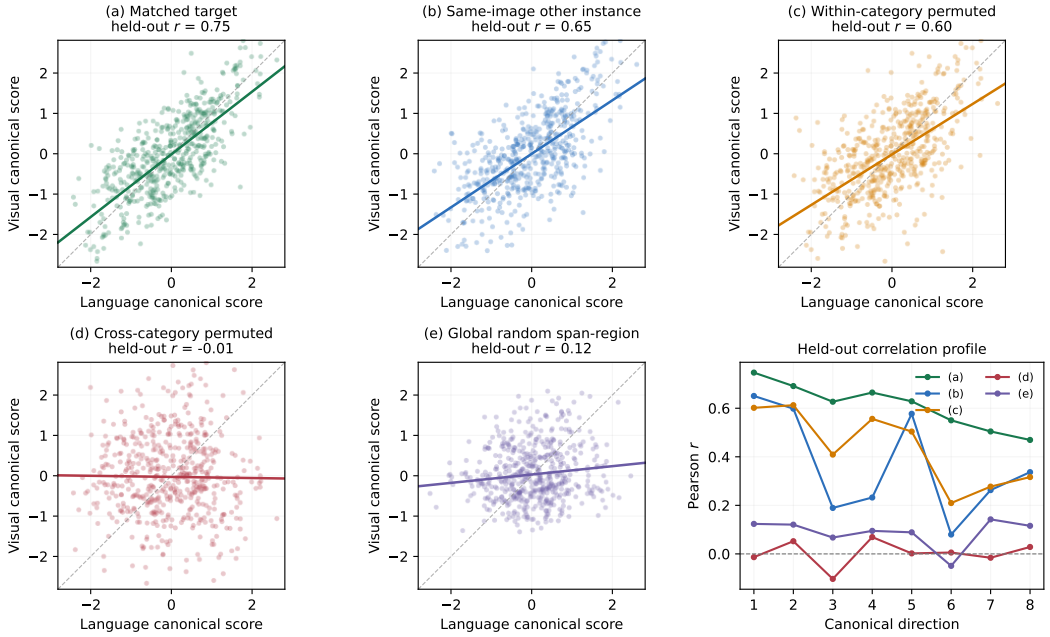

Figure 7: Held-out cross-modal pairs in a shared CCA basis. CCA is fitted only on matched training images; its projections are then fixed and evaluated under five image-disjoint pairing conditions. In (a–e), each point's horizontal and vertical coordinates are the scores of one language–vision representation pair along the first canonical direction. The gray dashed line is $y = x$, and the colored solid line is the linear-regression trend. The lower-right panel summarizes held-out Pearson correlations along the first eight canonical directions.

A.2 Native Cross-Modal Structure and Supervision-Dependence Controls

Research question and evidence chain. Because the Bridge is trained with bbox supervision, grounding accuracy alone cannot distinguish whether supervision reads out cross-modal structure already present in frozen representations or creates correspondence between otherwise unrelated spaces. We therefore design four complementary experiments: (1) directly measure the geometry of frozen representations without Bridge training; (2) preserve the bboxes and training procedure while disrupting only text–instance semantic pairing; (3) isolate supervision by category and evaluate on categories never observed during training; and (4) vary the amount of supervision to examine the sample efficiency of the low-rank mapping. The first test measures native structure, the second tests the need for correct semantic pairing, the third rules out category lookup, and the fourth characterizes the sample efficiency of supervised calibration.

Principles of CKA, RSA, and CCA. Let the centered language representations and target-patch pooled representations for the same set of entities be $X \in \mathbb{R}^{n \times d_t}$ and $Y \in \mathbb{R}^{n \times d_v}$. Linear CKA (Kornblith et al., 2019) does not fit a Bridge, but instead compares the overall relational structure of the two spaces over samples:

$$\mathrm{CKA}(X, Y) = \frac{\|X^\top Y\|_F^2}{\|X^\top X\|_F \, \|Y^\top Y\|_F}. \tag{7}$$

It is invariant to orthogonal transformations and isotropic scaling on either side, making it suitable for comparing visual and language representations with different dimensions and coordinate systems. RSA separately computes pairwise cosine-similarity matrices among entities and then calculates the Spearman correlation between their upper-triangular entries, testing whether the ordering of instance distances in language space is preserved in vision space.

CCA further finds linear directions a_k, b_k that maximize the correlation between Xa_k and Yb_k on the training images (Raghu et al., 2017):

$$\rho_k = \max_{a_k, b_k} \operatorname{corr}(Xa_k, Yb_k), \qquad E(K) = \frac{\sum_{k=1}^{K} \rho_k^2}{\sum_k (\rho_k^{\text{matched}})^2}. \tag{8}$$

All CCA directions are fitted only on training images, and ρ_k is evaluated on image-disjoint images. The five curves are normalized by the total held-out correlation energy of the matched condition rather than independently. Thus, $E(K)$ reflects both the number and strength of shared linear directions and cannot be made to reach one by separately rescaling each condition.

Representation-geometry controls. We sample at most 1,000 entities from each of the three data sources and restrict the common cohort to samples with a distinct annotated instance in the same image and valid random donors, yielding 2,614 entities, 2,085 training samples, and 529 image-disjoint evaluation samples. The text side is the mean activation of target tokens at T19, and the vision side is the mean activation of GT-positive patches at V6. We compare five inputs: (a) the correct text–target pair; (b) another instance in the same image; (c) a cross-instance permutation within the same category; (d) a cross-category permutation; and (e) a length-matched contiguous token span and random region drawn separately from unrelated images. From (a) to (e), instance, category, and image-context consistency are progressively disrupted, allowing us to determine whether correlation arises from precise semantic pairing, shared category information, or generic statistical structure.

Figure 7 uses canonical scores only as coordinates along shared correlation directions, not as grounding confidence. Correct instance pairs show the clearest diagonal trend along the first direction ($r = .75$). Another instance from the same image and within-category permutations retain some scene or category structure ($r = .65/.60$), whereas cross-category and global-random conditions fall close to zero ($r = -.01/.12$). Across the first eight directions, the matched condition consistently retains held-out correlations of .47–.75, showing that the correspondence is not driven by a single accidental direction. Because all controls use the projection learned from the matched condition without refitting, this hierarchy directly reflects the strength of cross-modal structure retained by correct-instance, category-sharing, and unrelated pairs in the native representations.

Supervision-dependence controls. The semantic-mismatch experiment holds fixed the number of training entities, bbox labels, Bridge capacity, optimization steps, and visual inputs, while pairing every bbox with a text entity representation from the wrong image and category. If supervision could create arbitrary grounding, this model should still achieve high localization performance. The category-held-out experiment trains on approximately three quarters of the categories and evaluates on 17 categories entirely absent from training, ruling out an explanation based on memorizing "category name–visual template" associations. The supervision-size experiment uses fixed subsets and an exposure-matched training schedule as a supplementary observation of whether the structure can already be read out with limited supervision. Complete results follow.

Correct pairing achieves the highest CKA (0.119) and CCA energy (100%) in the geometry analysis without Bridge training, whereas the cross-category and global-random conditions retain only 4.61% and 9.40%. Semantic mismatch further reduces Pointing by 58.23 pp, with an image-cluster bootstrap 95% CI of $[-60.08, -56.39]$ pp; category-held-out training reduces it by only 2.16 pp, with an interval of $[-5.38, 1.08]$ pp. These results respectively show that bbox supervision cannot forcibly recover high-quality correspondence under arbitrary semantic pairing and that Bridge localization cannot be reduced to memorization of observed categories. With 1%, 10%, and 100% of the training entities, the three-source validation macro-average AP values are 46.29%, 57.40%, and 63.15%. This sample efficiency is consistent with the interpretation that supervision calibrates existing structure and supports the conclusion that semantic consistency is linearly recoverable at low rank. In particular, the CCA curves further reveal the source of this low-rank property: before any Bridge is trained, the first 8 shared linear directions of the frozen representations already account for approximately 56% of the held-out cross-modal correlation energy, and the first 32 account for approximately 85%. This agrees with the Bridge being effective at low rank and exhibiting diminishing gains beyond rank 32, indicating that the recoverable cross-modal correspondence is concentrated in a small number of shared directions.

Table 6: Cross-modal geometry and supervision-dependence controls. CCA is the held-out correlation energy relative to the total matched energy; Pointing is evaluated on formal RefCOCO or its held-out-category subset (%).

Analysis	Condition	Frozen-representation geometry			Pointing
		Linear CKA	RSA	CCA energy	
Geometry	Matched target	0.119	0.027	100.00	–
Geometry	Same-image other instance	0.085	0.024	50.60	–
Geometry	Within-category permuted	0.065	0.023	43.26	–
Geometry	Cross-category permuted	0.019	−0.007	4.61	–
Geometry	Global random span–region	0.048	−0.015	9.40	–
Semantic pairing	Correct text–instance pairing	–	–	–	89.56
Semantic pairing	Cross-image, cross-category text permutation	–	–	–	31.34
Unseen categories	Full-category supervision reference	–	–	–	87.07
Unseen categories	17 test categories removed	–	–	–	84.91



Figure 8: Representative heatmaps from the semantically anchored person–dog swapped Bridge. During training, person and dog language activations are intentionally paired with visual targets of the opposite category, while all other categories retain correct pairings. Green boxes mark the target GT for the current phrase, magenta boxes mark opposite-category GT, and yellow boxes mark the top-scoring patch; all heatmaps use a shared raw-probability scale of $[0, 1]$. All top-scoring patches lie within regions where the GT boxes of the two categories overlap while remaining localized to the original target category, visually illustrating how correctly paired categories anchor the shared representational geometry.

Local concept swapping and semantic anchoring. Because global semantic mismatch disrupts every category, we further test whether a small number of incorrect pairings can rewrite the instance-level correspondence of the shared Bridge. Under the Qwen T19–V6 Linear-64 setting, we retain all available dog entities within the training split of each of RefCOCO, RefGQA, and VG, and sample an equal number of person entities, yielding 582 training entities for each category in total. In the swapped condition, person language activations are paired with visual activations and bboxes of dog targets from different images, and vice versa; the visual activations, bbox labels, and correct pairings of all other categories remain unchanged. The reference condition uses exactly the same training entities and budget, but restores the correct person–dog pairings. Swapping is applied only to the training set; the image-disjoint validation set always retains correct pairings.

To distinguish the contribution of other categories, we train both anchored and pair-only groups: the former retains all other correctly paired categories as semantic anchors, while the latter retains only balanced person and dog entities. Each group contains a correctly paired and a swapped Bridge. We first compare balanced Pointing on 25 person expressions and all 25 dog expressions from formal RefCOCO. We then evaluate 48 expressions (38 person and 10 dog) from 24 images containing both categories. We measure the proportion for which the top-scoring patch lies exclusively within the target-category or opposite-category GT box. A patch inside overlapping boxes, or outside both, is not assigned an unambiguous category direction.

Table 2 shows that, with other categories retained, swapping reduces balanced Pointing only from 90.0% to 88.0%. In the co-occurrence evaluation, 81.6% of person phrases and 70.0% of dog phrases still point exclusively to their original categories despite the incorrect supervision; only 5.3% of person phrases point to dog, while no dog phrase points to person. After removing the anchors provided by other categories, the same swap reduces balanced Pointing from 74.0% to 30.0%, and 60.0% of dog phrases point exclusively to person. Person phrases more often fall outside both category boxes rather than exhibiting a symmetric concept swap. Thus, incorrect supervision can change the low-rank mapping, but when other correctly paired categories jointly constrain it, they anchor the geometry of the vision–language representations and prevent local mismatches from causing a systematic reversal. This provides stronger evidence that Bridge training does not forcibly establish correspondence between unrelated representation spaces; rather, it performs low-rank linear recovery of native cross-modal structure on the basis of linearly separable vision–language representational geometry.

A.3 Layerwise Patterns and Effects of Visual-Injection Architectures

The three columns of Figure 5 report held-out AP for sparse Bridge layer combinations, language-side object probeT, and vision-side target–background separation, respectively. Object probeT measures whether entity information in frozen language states can be read linearly, while vision separation measures the similarity difference between patches within the GT target and target–background patches; the two are independent unimodal screening quantities. The following analysis first describes the observed trends and then discusses their relationship to visual-injection architectures, without interpreting any individual curve itself as causal evidence.

LLaVA: language information forms progressively, while visual structure follows a single peak. In the second column, LLaVA's probeT F1 rises overall from 47.0% at L0, reaches approximately 70% several times from L22 to L29, peaks at 71.5% at L29, and then falls to 63.0% at L35. This trend is an "overall rise followed by a final-layer decline" with inter-layer fluctuations, rather than a strictly smooth inverted U. Vision separation in the third column exhibits a clearer single-peak structure: it rises gradually from 4.62 at V0 to 16.04 at V13, then declines continuously to 0.42 at V23. Consistent with these two relatively smooth marginal curves, the matrix in the first column has no isolated spike, but instead forms a continuous and relatively uniform high-value region across T19–T29 and V11–V13. The best combination is T24–V11 (64.99%), which is still not obtained by directly pairing the probeT peak at L29 with the separation peak at V13. This pattern is consistent with LLaVA's architecture, in which visual tokens enter a unified decoder through a single merger/projector and are then progressively integrated by subsequent layers.

Qwen: the language side forms a broad plateau, while the vision side undergoes multi-stage reorganization. Qwen's probeT curve does not show a clear rise followed by a decline: F1 falls from 75.3% at T0 to 71.2% at T1, recovers to 75.3% at T2, and briefly falls again to 72.9% at T3. Most subsequent layers remain at approximately 76–79%; only T34 locally drops to 73.9%, followed by a rebound to 79.3% at T35. Its main characteristic is therefore a high-value plateau spanning most depths, accompanied by local fluctuations in a few layers. The third column shows multiple local peaks near V6, V12–V13, and V22, followed by a sharp drop from 7.52 at V22 to -0.52 at V23, indicating that the target–background structure of visual patches undergoes repeated strengthening, weakening, and remixing with depth. In the first column, high values are concentrated mainly at T13–T25 and V5/V6/V11, with T19–V6 attaining the highest held-out AP of 63.03%. By contrast, the entire V17 column ranges only from 33.53% to 36.45%, a cliff-like decline relative to adjacent candidate visual stages. This occurs because layer 17 is the final additional injection layer of Qwen3-VL-4B's DeepStack mechanism; newly injected visual information substantially reorganizes the existing cross-modal representational structure, supplementing but also disrupting the

current representation to some extent. This matrix pattern echoes the multi-stage fluctuations on the vision side and also shows that a visual layer with higher separation does not automatically yield stronger cross-modal recovery. The high decodability across language layers is consistent with the architectural property that DeepStack continually supplements visual features at multiple depths. However, the nonuniform Bridge grid shows that multi-stage injection does not make all language–vision pairings equivalent; specific language states remain strongly linearly compatible with only some visual stages.

Llama: the language side first rises and then falls, while the vision side has multiple local peaks. Llama's probeT F1 rises gradually from 59.1% at L0 to 67.9% at L17, and then declines overall to 62.4% at L39, showing the clearest rise-then-fall trend among the three models. Vision separation falls from 10.55 at V0 and then rises again, reaching 11.89 at V14 and a secondary peak of 11.44 at V29 before declining to 7.50 at V31, showing that local visual structure is also reorganized at multiple stages. Variation across text layers in the first column is likewise relatively smooth and forms a broad high-value plateau across T13–T22 and V11–V17; periodic cross-attention layers T13/T18 both lie within this plateau. The final checkpoint for T13–V17 and the fully trained T15–V17 reach 49.18% and 49.15%, respectively, but T15–V14 is more stable in external RefCOCO grounding and is therefore adopted as the final configuration. This result is consistent with the architecture in which periodic cross-attention repeatedly reads visual states, while also showing clear matching differences between the language states formed by later reads and different visual layers.

Figure 10 further presents the complete similarity structure on the vision side. The three solid curves denote the average cosine similarities within the target, between target and background, and within the background, respectively; the dashed curve is the difference between within-target and target–background similarity. Qwen's three solid curves fluctuate largely in synchrony, but the distances between them repeatedly expand and contract with depth, showing that the multiple separation peaks are not caused merely by a global shift in the similarity scale. LLaVA consistently maintains higher within-target similarity than target–background similarity from V0 to V20; at V21–V23, all three absolute similarities rise rapidly while the separation gap instead collapses, indicating that final-layer patches become globally homogenized rather than forming clearer target boundaries. Llama's three absolute similarities decline overall in the middle and later layers with local fluctuations, while a nonzero gap remains between within-target and target–background similarity. The joint changes in the solid and dashed curves show that vision layers alter both the global aggregation of patch representations and the local geometry of the target relative to the background; therefore, the separation peak alone remains insufficient to identify the vision layer best suited for cross-modal readout.

Neither the unilateral peak of probeT nor that of vision separation can directly determine the optimal Bridge. DeepStack's multi-stage injection and the periodic reads of cross-attention cause language states at different depths to retain or re-encode visual information from different stages, thereby changing their linear compatibility with each layer of the vision tower; this is consistent with the clearly nonuniform AP grid in the first column. Accordingly, we use probeT and vision separation only to prescreen candidate layers, and then determine Qwen T19–V6, LLaVA T24–V11, and Llama T15–V14 using the actual Bridge grid and external grounding. The interpretation that Llama's single-layer causal intervention is readily compensated by subsequent visual reads comes from the matched intervention in Section 4.3, rather than being inferred from the screening curves in Figure 5 alone.

A.4 Complete Layer and Rank Ablations

Figure 5 in the main text presents all complete training results under a single shared color scale. The sparse layer grids for Qwen, LLaVA, and Llama all exhibit pronounced layer selectivity; the final configurations are T19–V6, T24–V11, and T15–V14, respectively. Rank curves are shown in Figure 3 in the main text. The macro-average AP values for Qwen at Rank-8/16/32/48/64 are 58.19/60.39/60.92/61.18/61.10; those for LLaVA are 61.71/64.31/64.84/64.75/64.99; and those for Llama are 45.25/47.14/48.25/48.92/48.41.

In additional ablations, we test different Bridge architectures and loss functions. Fixed/Learned LayerNorm, MLP-64/256, and BCE+margin do not alter the main conclusion that a purely linear low-rank structure is recoverable and saturates at approximately rank 32. Margin loss instead

Table 7: Macro-average three-source image-disjoint held-out validation AP and AUROC (%) at different Bridge ranks. Rank 8 already retains substantial capability, and performance largely saturates beyond rank 32.

Model	Metric	Rank-8	Rank-16	Rank-32	Rank-48	Rank-64
Qwen T19–V6	AP	58.19	60.39	60.92	**61.18**	61.10
	AUROC	88.61	89.23	89.27	**89.28**	89.21
LLaVA T24–V11	AP	61.71	64.31	64.84	64.75	**64.99**
	AUROC	88.78	89.55	89.68	89.67	**89.71**
Llama T15–V14	AP	45.25	47.14	48.25	**48.92**	48.41
	AUROC	88.04	88.66	88.96	**89.08**	89.04

reduces overall AP and is therefore not used for the primary models. Furthermore, introducing LayerNorm, nonlinear activations, and higher-capacity MLP structures does not further improve grounding capability, indicating that performance does not depend on a complex decoder relearning visual semantics. A low-rank linear mapping can already read local semantic information present in frozen language and visual representations and recover linearly decodable cross-modal consistency between them through simple supervised alignment. It should be noted that RefCOCO substantially outperforms the other two datasets in Table 8, and its held-out AP gap relative to RefGQA/VG does not arise entirely from differences in representational recoverability. RefCOCO referring expressions are designed to distinguish target instances in an image and usually contain sufficient attribute or spatial qualifiers. By contrast, RefGQA and VG are constructed from scene-graph object and relation annotations; many phrases provide only a category name even when the image contains multiple visually plausible instances. These two sources also contain incomplete or inconsistently granular bounding boxes, such as a box covering only a small corner of a large grassy region or a single piece in a plate of chopped carrots labeled as "carrot." In such cases, assigning a high score to another same-category instance or to the complete semantic region is still counted as an error by the patch-level single-box GT, systematically depressing AP. Figure 9 presents four representative cases. Its four columns focus on: (a) only one of two visually similar benches being annotated; (b) the plural referent "customers" corresponding to multiple people while the GT boxes only one; (c) only one of several similarly dressed boys being boxed; and (d) only one small piece in a plate of carrot pieces being labeled as "carrot." These examples show that a prediction falling on another plausible instance or the complete semantic region can still be counted as an error by a single-box metric.

We nevertheless retain RefGQA and VG during training and limit each of their sampling ratios to 0.3. Although these samples should not be interpreted as fully unambiguous instance-level evaluations, they provide richer scenes, object relations, and linguistic variation than RefCOCO. They can therefore serve as noisy multi-source augmentation, reducing Bridge overfitting to the expression templates and annotation conventions of a single dataset and improving robustness to incomplete references and annotation perturbations. To prevent this noise from obscuring the conclusion, we report results separately for the three image-disjoint held-out sources and conduct the large-scale formal instance-localization evaluation on RefCOCO, where the references are clearer.

Table 8 shows that rank 8 is not effective merely by chance on a single data source, but retains measurable local recovery capability on RefCOCO, RefGQA, and VG alike; gains across all three sources generally slow beyond rank 32. Table 9 further shows that this low-rank recoverability cannot be obtained automatically by PCA/CCA or random rotation with the same output dimensionality. The key remains extracting and combining useful dimensions from the original encodings to obtain cross-modal directions; in this work, such useful extraction is achieved through limited local supervised learning.

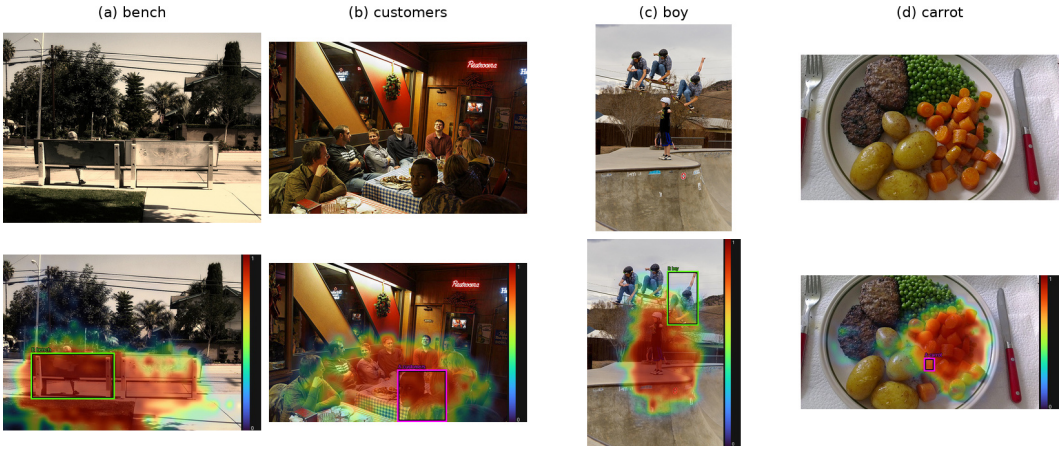

Figure 9: Representative ambiguous references or annotations in RefGQA/VG. The top row shows the original images; the bottom row shows Bridge heatmaps for the specified single entity and its sole scene-graph GT box.

Table 8: Patch AP (%) of low-rank Bridges on each held-out data source. The macro-average corresponds to Figure 3 in the main text. Bold indicates the configuration with the highest macro-average AP for each model.

Model	Rank	RefCOCO	RefGQA	VG	Macro-average
Qwen T19–V6	8	70.94	55.20	48.45	58.19
	16	72.48	58.22	50.46	60.39
	32	73.11	58.29	51.36	60.92
	48	**73.25**	**58.90**	**51.39**	**61.18**
	64	73.27	58.15	51.88	61.10
LLaVA T24–V11	8	73.52	60.45	51.17	61.71
	16	73.65	64.98	54.29	64.31
	32	74.24	66.88	53.41	64.84
	48	73.97	67.55	52.72	64.75
	64	**74.01**	**67.85**	**53.10**	**64.99**
Llama T15–V14	8	48.08	47.32	40.36	45.25
	16	49.14	49.83	42.45	47.14
	32	49.91	51.34	43.51	48.25
	48	**50.11**	**52.42**	**44.22**	**48.92**
	64	49.52	51.97	43.74	48.41

Table 9: Rank-32 controls on the same Qwen RefCOCO activation cohort (%). All methods have the same final shared dimensionality.

Method	AP	AUROC	Pointing
Linear Bridge-32 (ours)	**74.03**	**92.57**	**88.70**
PCA-128 + CCA-32	28.25	64.74	55.60
PCA-32 + CCA-32	23.40	59.90	39.80
Random Procrustes-32	19.20	53.58	29.40
Random unaligned-32	18.01	51.59	21.10

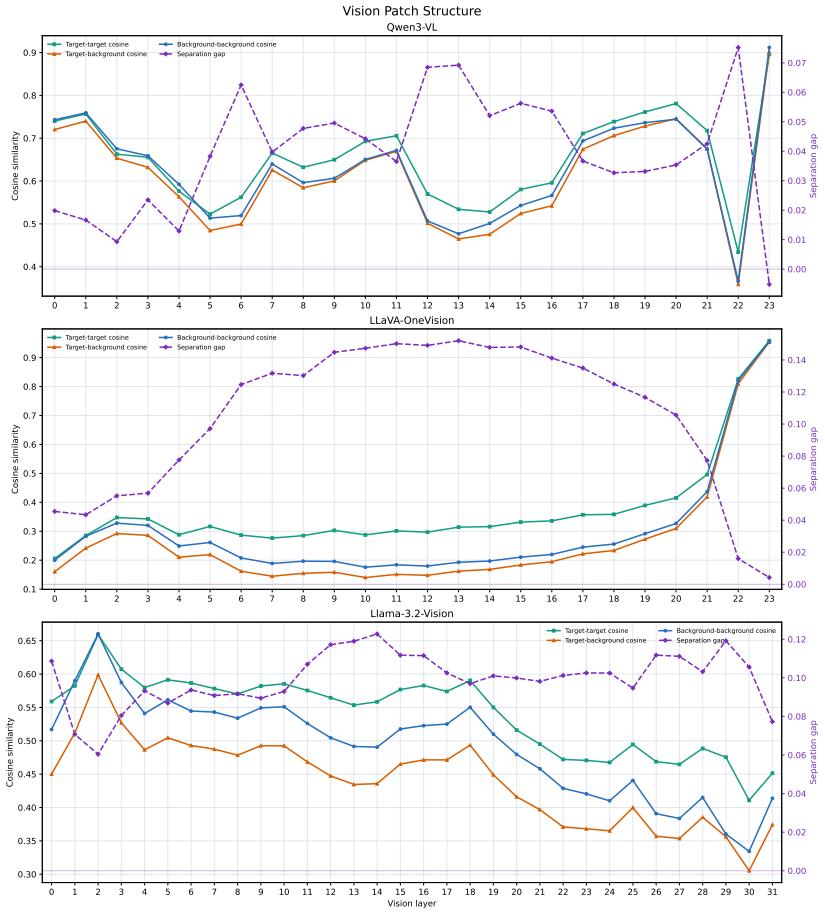

Figure 10: Visual patch structure in the three models. Curves are macro-averaged over the three data sources, and target regions are defined by GT bounding boxes.

A.5 Expanded Grounding Evaluation

Table 10: Grounding Pointing on a RefCOCO subset (150 expressions). All methods use exactly the same expressions and GT bounding boxes. Predicted boxes from GroundingDINO/OWL-ViT are mapped to the same Qwen patch grid for evaluation.

Method	Pointing
Qwen Linear-64 Bridge	**84.67**
Qwen Native Attention	52.00
LLaVA Linear-64 Bridge	**86.00**
LLaVA Native Attention	53.33
LLAMA Linear-64 Bridge	**60.67**
GroundingDINO-T	59.33
OWL-ViT	43.33
CLIP patch–text similarity	12.00

The formal RefCOCO set is fixed by image ID. Pointing on testA/testB is 90.68/86.01 for Qwen, 90.90/86.13 for LLaVA, and 73.50/69.78 for Llama. On an early controlled subset used for preliminary validation, a VG-only Bridge trained without any RefCOCO supervision achieves 56.67/54.67 Pointing at epochs 50/70, substantially above naive CLIP's 12.00 on the same set but below joint three-source training; this is a dataset-transfer zero-shot experiment. GroundingDINO-T and OWL-ViT achieve 59.08 and 48.18 Pointing, respectively, on the formal 3,421-expression set and serve as specialized open-vocabulary localization references. The formal CLIP baseline uses cosine similarity between final-layer contextual patch tokens and the global text vector from OpenAI CLIP ViT-L/14@336. Center cropping is disabled during preprocessing to preserve the coordinate correspondence between the complete image and the GT bbox. On the 3,421 expressions, it achieves 16.78% Pointing, 43.70% Hit@5, 58.73% Hit@10, and 13.52% patch IoU. Scale analysis shows that small objects are substantially more difficult.

A.6 Complete Quantitative Results for Referring Representations

Table 11: Evolution of referring-expression prefixes on Qwen. After the third or fourth word, localization is usually already close to that of the complete phrase.

Prefix length	1	2	3	4	5
Pointing	53.06	63.27	85.71	89.13	87.10
IoU@0.6	34.25	36.89	45.17	49.06	48.48
Pearson with full map	0.622	0.732	0.897	0.961	0.963

A.7 Attention Propagation and Answer-Correction Cases

Figure 13 supplements three complete cases from Bridge evidence and attention changes to answer correction, covering a spatial relation in RefCOCO-QA Small, attribute recognition in RefCOCO-QA, and relative-position judgment in V*. All Qwen cases use the L11, $\beta = 0.9$, Bridge-centered configuration frozen in the main text and read attention at L15.

A.8 Complete Causal and Application Tables

The causal experiments comprise two stages that scan intervention strength and routing layer. We first scan the suppression strength γ and retention floor f at Qwen L13, with results reported in Tables 12 and 13, respectively. We then scan routing layers around the primary coefficients used in the main text; the complete Qwen and LLaVA suppression and retention results are reported in Tables 14 and 15, respectively. The oracle bbox is a reference rather than a strict upper bound because its sparsity is not matched to the Bridge map.

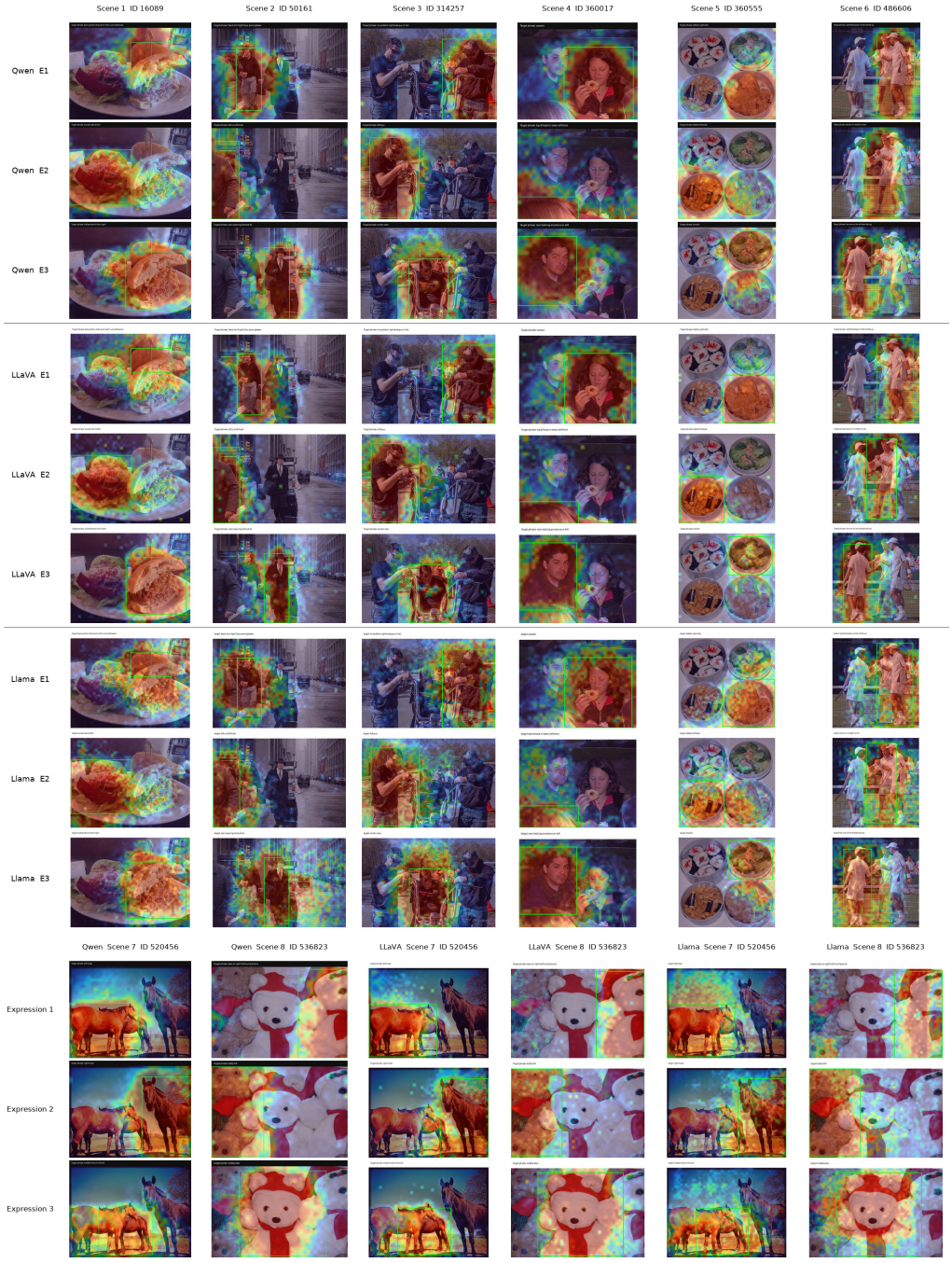

Figure 11: Instance-level grounding by three VLMs. The upper panel presents the first six shared scenes in a 9×6 layout, with Qwen, LLaVA, and Llama each occupying three consecutive rows. The lower panel arranges the remaining two scenes in a 3×6 layout, with each model occupying two consecutive columns. All models use the same original images and referring expressions, enabling comparison of instance-conditioned localization by the low-rank Bridge across different vision–language fusion architectures.



Figure 12: Qualitative reference for localization baselines. From top to bottom: GroundingDINO-T, OWL-ViT, and CLIP patch–text similarity. Results for each model are arranged as three expression rows by five shared-scene columns; the three images in each column correspond to three referring expressions for the same image. Specialized open-vocabulary localizers can find the relevant category, but more often produce broad or incorrect regions when attributes and relative positions are needed to distinguish same-category instances. CLIP, trained for global alignment, has almost no local localization capability.

Table 12: Development-stage scan of suppression strength γ at Qwen L13 (V* accuracy, %). Random uses the same set of score values in each row and only permutes their patch locations; Gap is Random−Bridge.

γ	Baseline	Bridge	Random	Oracle	Gap
.3	86.97	85.29	86.55	84.87	1.26
.5	86.97	83.19	86.97	83.61	3.78
.7	86.97	76.89	86.97	77.73	10.08
.9	86.97	67.65	86.13	72.69	18.49

Table 13: Development-stage scan of retention floor f at Qwen L13 (V* accuracy, %). Lower f more strongly suppresses the low-scoring background; Gap is Bridge−Random. This table shows coefficient sensitivity, while the primary causal result in the main text uses L11, selected by a subsequent layer scan.

f	Baseline	Bridge	Random	Oracle	Gap
.1	86.97	64.71	50.00	58.40	14.71
.2	86.97	73.11	57.14	66.39	15.97
.3	86.97	81.09	68.49	80.25	12.61
.4	86.97	85.29	78.15	84.03	7.14
.5	86.97	85.71	81.09	85.29	4.62
.6	86.97	86.55	82.35	86.55	4.20

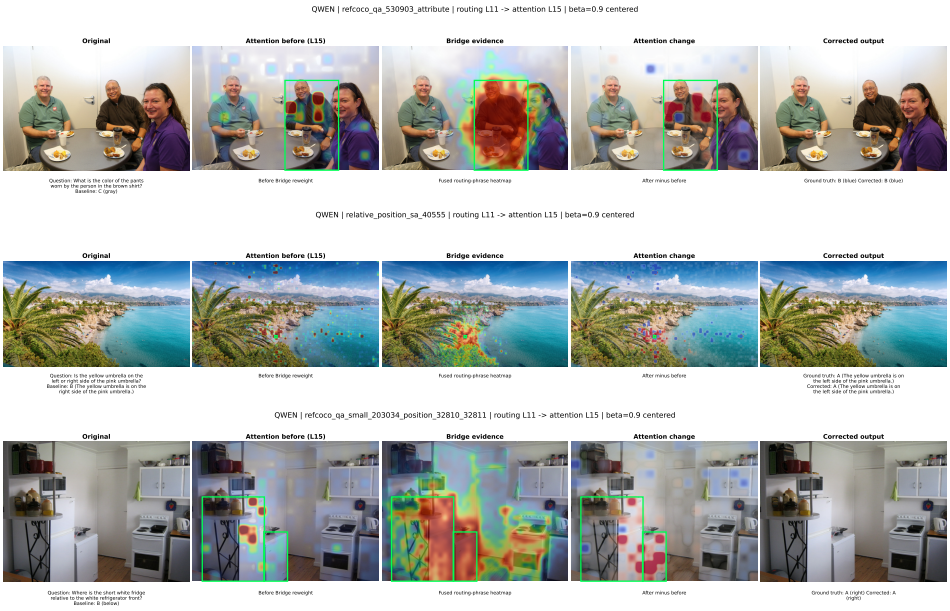

Figure 13: Additional answer-correction cases from Bridge-guided reweighting. From top to bottom: (1) the pants color of the person in a brown shirt in RefCOCO-QA, corrected from C to B (blue); (2) the position of the yellow umbrella relative to the pink umbrella in V*, corrected from B (right) to A (left); and (3) the relative position of two refrigerators in RefCOCO-QA Small, corrected from B (below) to A (right). As in Figure 6, the five columns in each row connect the original input and baseline, attention before reweighting, Bridge evidence, the attention change, and the corrected output.

Qwen L11 produces the largest location difference of 15.55 pp under $\gamma = .7$ suppression and achieves a 23.53 pp Bridge–Random difference under retention with $f = .2$. The main text consistently reports L9 for LLaVA: the suppression and retention differences are 9.24 pp and 18.07 pp, respectively. L11's suppression point estimate is 9.66 pp, only 0.42 pp higher, and its confidence interval overlaps strongly with that of L9; it is therefore retained as a layer-scan robustness result. The current floor scan reports $.1/.2/.3/.4/.5/.6$ and verifies the mathematical equivalence of retention with $f = .1$ and low-score suppression with $\gamma = .9$ on the Bridge map; the random controls for the two paths differ because each uses an independently sampled spatial permutation.

Table 14: Complete intervention-layer scan for V* suppression ($\gamma = .7$, %). Gap is Random−Bridge; a larger positive value indicates that disrupting high-scoring Bridge locations is more harmful than disrupting random locations at the same strength. Bold marks the causal layers consistently reported in the main text. LLaVA L11's highest point estimate exceeds L9 by only 0.42 pp, and their confidence intervals strongly overlap.

Model/layer	Baseline	Bridge	Random	Oracle	Gap
Qwen L9	86.97	77.31	87.39	78.57	10.08
Qwen L11	86.97	71.85	87.39	74.37	15.55
Qwen L13	86.97	76.89	86.97	77.73	10.08
LLaVA L9	64.29	55.88	65.13	60.50	9.24
LLaVA L11	64.29	53.78	63.45	59.66	9.66
LLaVA L13	64.29	58.82	63.45	60.50	4.62

Table 15: Complete primary-layer and floor scan for V* retention (%). Gap is Bridge−Random; a larger positive value indicates that high-scoring Bridge locations preserve more task capability under strong background suppression.

Model/layer	f	Baseline	Bridge	Random	Oracle	Gap
Qwen L9	.2	86.97	82.35	61.34	83.19	21.01
Qwen L9	.4	86.97	87.39	83.19	88.66	4.20
Qwen L11	.2	86.97	84.03	60.50	82.35	23.53
Qwen L11	.4	86.97	88.24	79.83	86.97	8.40
Qwen L13	.2	86.97	73.11	57.14	66.39	15.97
Qwen L13	.4	86.97	85.29	78.15	84.03	7.14
LLaVA L9	.2	64.29	66.39	48.32	61.76	18.07
LLaVA L9	.4	64.29	63.45	57.56	63.03	5.88
LLaVA L11	.2	64.29	64.71	47.48	60.08	17.23
LLaVA L11	.4	64.29	62.18	57.14	61.76	5.04
LLaVA L13	.2	64.29	63.45	49.16	60.08	14.29
LLaVA L13	.4	64.29	63.87	62.18	65.13	1.68

The Llama scan covers eight periodic cross-attention layers: L3/L8/L13/L18/L23/L28/L33/L38. The Random–Bridge suppression difference does not exceed 1.68 pp except for 3.36 pp at L13, while the Bridge–Random retention difference fluctuates only within $[-1.26, 0.84]$ pp, failing to form a location effect as stable as those observed on Qwen/LLaVA.

For applications, model-level configurations are fixed rather than tuned separately for each dataset. On the 600 RefCOCO-QA questions, baseline/Bridge/Centered/uniform/random scores are 77.67/78.17/79.33/77.33/77.83 for Qwen and 73.00/75.00/75.00/74.17/74.17 for LLaVA. Qwen Centered yields 11 wins and 1 loss relative to the baseline (exact paired $p = .0063$); LLaVA Centered yields 14 wins and 2 losses ($p = .0042$).

Table 16: Development-stage sensitivity to application mode and strength at a fixed routing layer (accuracy, %). V* and ADE20K Count are predefined sets; POPE-1k and GQA-dev are used only for configuration selection and do not replace the formal results in the main text. Bold marks the final frozen model-level configuration column, selected as a cross-task compromise rather than as the optimum for each dataset.

Model	Dataset	Baseline	$\beta = .7$		$\beta = .9$	
			Bridge	Centered	Bridge	Centered
Qwen	V*	86.97	87.82	88.66	87.82	**89.08**
Qwen	POPE-1k	90.40	90.50	90.80	91.40	**90.80**
Qwen	GQA-dev	78.90	79.70	79.20	79.10	**79.10**
Qwen	ADE20K Count	56.43	59.36	59.36	58.77	**59.65**
LLaVA	V*	64.29	64.71	**64.71**	64.71	64.29
LLaVA	POPE-1k	91.80	91.50	**91.60**	90.90	91.30
LLaVA	GQA-dev	84.00	84.40	**84.70**	83.50	84.50
LLaVA	ADE20K Count	57.60	57.02	**59.06**	57.89	58.19

Table 17: Complete matched controls on the 600 RefCOCO-QA questions (accuracy, %). Bold indicates the best result for each model.

Model	Baseline	Bridge	Bridge-centered	Uniform	Random
Qwen	77.67	78.17	**79.33**	77.33	77.83
LLaVA	73.00	**75.00**	**75.00**	74.17	74.17