GAMA: COMPOSITIONAL OBJECT-CONCEPT BINDING FOR CONTROLLABLE MULTIMODAL EMBEDDINGS

Anonymous authors
Paper under double-blind review

ABSTRACT

Frozen multimodal encoders can recognize objects and attributes while still entangling which concept belongs to which object. This limits counterfactual object-level editing: moving a feature direction for one object may unintentionally perturb another object or the surrounding scene. We propose GAMA, a post-hoc structured intervention interface that reconstructs frozen scene embeddings from concept-position tuples while exposing editable object slots. Its core mechanism combines reusable additive concept semantics with gated low-rank multiplicative conjunction residuals. Across controlled text and visual scenes, GAMA preserves frozen-space geometry while supporting targeted replacement and object-concept decomposition. Continuous-factor recovery and a Grounding-DINO/CLIP pipeline test a practical relaxation from curated tuples to automatically extracted pseudo-tuples. Because GAMA reorganizes only information retained by a frozen encoder, its scope is compositional recombination within given structured concept spaces, not end-to-end perception or unrestricted open-vocabulary recognition of unseen object identities.

1 INTRODUCTION

Frozen multimodal encoders are strong recognition layers, yet controlled diagnostics expose failures under reordered relations, attributes, and captions Thrush et al. (2022); Yuksekgonul et al. (2023); Ma et al. (2023). Recognition is also weaker than editable object-level binding: a scene vector may encode a red cube and a blue sphere without exposing which attribute belongs to which object. Editing only the cube should therefore act on one object factor rather than globally perturbing the sphere or scene context.

We study whether a frozen representation can be reorganized into object-instance factors that preserve its geometry and support local replacement. Prior analyses find attribute misbinding in CLIP-like models even when binding information remains probe-readable Lewis et al. (2024); Gurung et al. (2025); Koishigarina et al. (2026), while pooled text representations can lose compositional detail Kamath et al. (2023). We use counterfactual editing as a representation-control diagnostic, not as causal discovery or causal inference.

We assume each scene is available as concept-position tuples and reconstruct a target from a frozen encoder. This is a deliberate modular decomposition between perception and binding: GAMA is a post-hoc interpretation and control layer, not an end-to-end detector, and an automatic real-image pipeline tests whether external perception modules can populate its interface under noise. A stronger modular frontend can therefore improve the supplied tuples without requiring retraining of the large frozen backbone. Because the encoder remains fixed, GAMA is upper-bounded by information already retained in its representation and cannot create missing perceptual content; its contribution is to reorganize entangled frozen-space geometry into more editable and interpretable slots.

GAMA combines reusable additive concept tables with gated low-rank multiplicative residuals for conjunction-specific structure. Position, context, and global streams preserve scene geometry while bound slots remain accessible for replacement and decomposition. The paper focuses on recombination within given structured concept spaces, not unrestricted open-vocabulary recognition. Graph/CrossAttn variants are exploratory multi-object extensions; Section 4.6 summarizes

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.

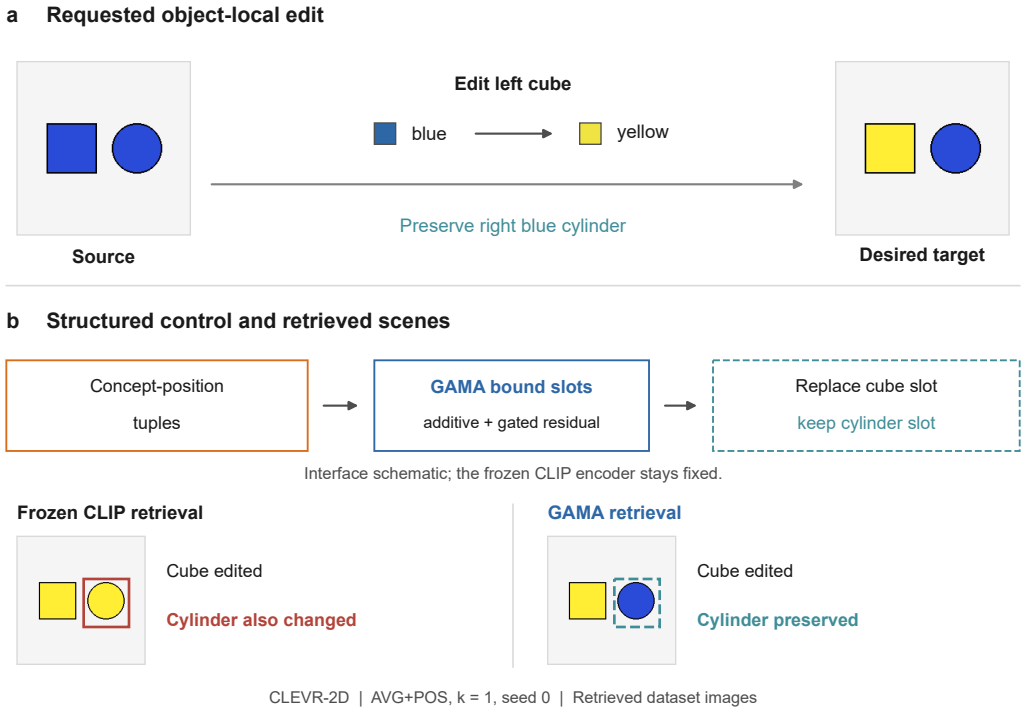

Figure 1: Object-local control in a frozen representation space. (a) Change the left cube from blue to yellow while preserving the right blue cylinder. (b) Concept-position tuples enter the GAMA structured binding interface for local replacement; the interface is schematic, not a visualization of measured activations. In a stored CLEVR-2D example under the AVG+POS replacement protocol ($k = 1$, seed 0), the frozen CLIP intervention changes both objects, whereas GAMA retrieves the intended target and preserves the cylinder. All thumbnails are stored dataset images returned by retrieval, not generated images.

their dense-scene reconstruction–efficiency trade-off, while full architectures and diagnostics are reported in Appendix G.1.

Our contributions are fourfold: (1) a post-hoc structured interface for editable binding in frozen multimodal spaces; (2) additive semantics with gated low-rank conjunction residuals for editable slots; (3) controlled evidence on frozen-space reconstruction, targeted replacement, and object-concept decomposition; and (4) scope analyses for continuous factors, automatic pseudo-tuples, multi-object interaction, and unseen-object boundaries.

2 Related Work

Compositional generalization and recombination diagnostics. Systematic generalization asks whether familiar parts recombine in unfamiliar configurations Lake & Baroni (2018). CLEVR and modular networks provide controlled settings Johnson et al. (2017); Andreas et al. (2016), while Winoground, ARO, CREPE, VL-CheckList, and SugarCrepe diagnose failures under order, attributes, relations, and hard negatives Thrush et al. (2022); Yuksekgonul et al. (2023); Ma et al. (2023); Zhao et al. (2022); Hsieh et al. (2023). Related work also studies pooled-text information loss and structured contrastive learning Kamath et al. (2023); Singh et al. (2023). We instead ask whether a frozen scene embedding can be factorized and edited post hoc.

Object-centric and concept-structured representations. FactorVAE separates latent factors Kim & Mnih (2018); MONet, IODINE, and Slot Attention represent scenes with object-like components Burgess et al. (2019); Greff et al. (2019); Locatello et al. (2020). Concept Activation Vectors and Concept Bottleneck Models expose concepts Kim et al. (2018); Koh et al. (2020), but

readability does not ensure correct object attachment. CLIP, DINOv2, and SigLIP provide strong frozen spaces Radford et al. (2021); Oquab et al. (2024); Zhai et al. (2023); object-centric contrastive and region-level multimodal methods learn related structure Assouel et al. (2025); Fu et al. (2026), whereas GAMA binds a fixed representation post hoc.

Object-concept binding and post-hoc intervention. Controlled studies distinguish recognition from reliable object-attribute binding in CLIP-like models Lewis et al. (2024); Gurung et al. (2025); Koishigarina et al. (2026). The binding-complexity protocol of Uselis et al. (2026) motivates our decomposition, replacement, subtraction/permutation, and sum-multiplicative diagnostics. GAMA keeps the encoder fixed and learns editable slots aligned to its scene geometry.

Open-vocabulary grounding and automatic structured inputs. Open-vocabulary detectors such as Grounding DINO map prompts to regions Liu et al. (2024). Together with frozen crop scoring, they can populate noisy labels, attributes, and position bins for structured models. We use grounding only as an automatic input provider and evaluate the resulting pseudo-tuples under perception uncertainty; open-vocabulary perception itself is not the target of GAMA.

3 METHOD

GAMA is a structured surrogate over a frozen representation rather than a perception model. It reconstructs a frozen target while exposing editable scene parts: additive tables carry reusable concepts, gated low-rank products correct conjunctions, and context/global branches preserve scene geometry without replacing the object slots. The method can only reorganize information retained by the target encoder; it does not infer perceptual content absent from that representation.

3.1 THEORETICAL MOTIVATION: REUSE VERSUS CONJUNCTION CAPACITY

We express the tension between reusable factors and combination-specific geometry as

$$f(c_1, \dots, c_M) = \sum_m A_m(c_m) + \Delta(c_1, \dots, c_M).$$

A pure additive model sets $\Delta = 0$ and underfits nonlinear conjunctions; a full lookup fits combinations but weakens reuse. GAMA uses a low-rank multiplicative residual as a constrained correction and gates it dimensionwise when conjunction-specific capacity is needed.

3.2 PROBLEM FORMULATION

Let a dataset contain N scenes. Scene i has K ordered object slots and M concept dimensions. We write the symbolic scene as

$$s_i = \{x_{i,1}, \dots, x_{i,K}\}, \qquad x_{i,k} = (c_{i,k}^{(1)}, \dots, c_{i,k}^{(M)}, p_{i,k}), \tag{1}$$

where $c_{i,k}^{(m)} \in \{1, \dots, V_m\}$ is a concept value and $p_{i,k}$ is an optional position or slot index. Object identity, when available, is one concept dimension; order is retained because visual embeddings are layout-sensitive.

A frozen encoder E maps the original text or image observation y_i corresponding to symbolic scene s_i into a target embedding

$$z_i = E(y_i) \in \mathbb{R}^d. \tag{2}$$

The encoder parameters remain fixed. We learn

$$\hat{z}_i = f_\theta(s_i), \tag{3}$$

whose prediction remains close to z_i while exposing object and concept components. Thus GAMA is a structured surrogate for the frozen space, not a tuple classifier; its decomposition is an inductive bias rather than a claim about the encoder's internal architecture.

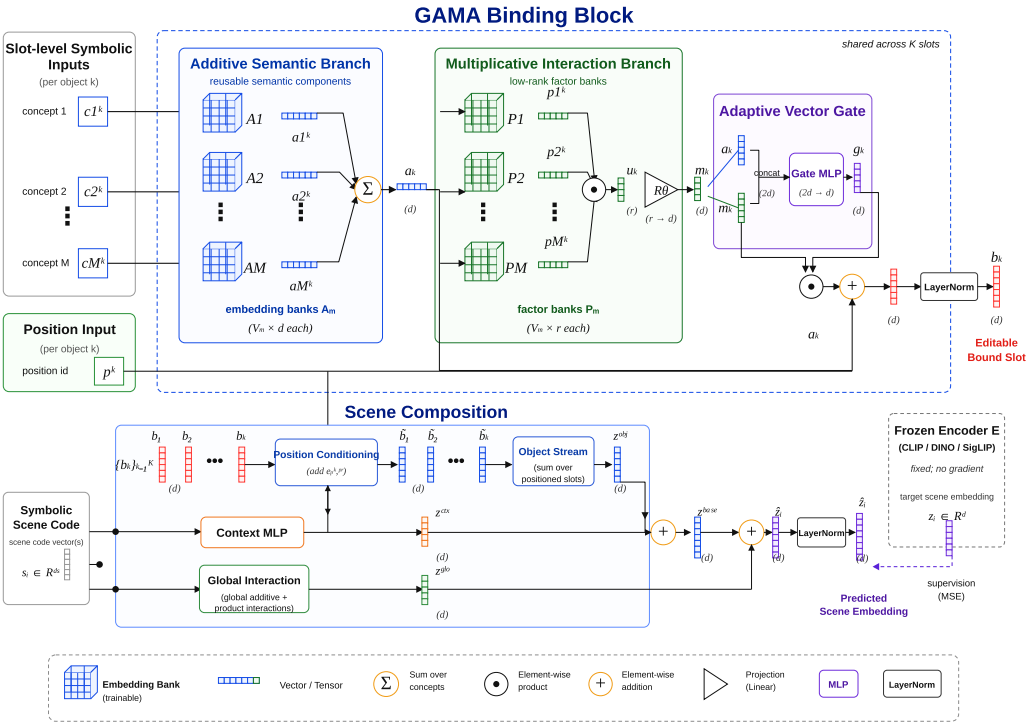

Figure 2: GAMA architecture at the module level. The frozen encoder supplies only the target embedding and receives no gradient. A shared slot-binding core maps concept tokens to editable bound slots through additive concept tables, low-rank multiplicative residuals, and a learned vector gate. Position conditioning is applied afterward before the slots enter the object stream. The object stream is combined with context and global interaction streams, and the resulting representation is layer-normalized and aligned to the frozen target geometry.

3.3 Reusable Semantics and Conjunction Residuals

For reusable semantics, each concept dimension m has an additive table

$$A_m \in \mathbb{R}^{V_m \times d}. \tag{4}$$

The additive representation of slot k in scene i is

$$a_{i,k} = \sum_{m=1}^{M} A_m(c_{i,k}^{(m)}). \tag{5}$$

This shares each concept direction across objects and scenes.

For conjunction residuals, GAMA stores a low-rank factor table

$$P_m \in \mathbb{R}^{V_m \times r}, \qquad r \ll d. \tag{6}$$

The factors for an object slot are multiplied coordinatewise,

$$u_{i,k} = \prod_{m=1}^{M} P_m(c_{i,k}^{(m)}), \tag{7}$$

The coordinatewise product is projected into the frozen embedding dimension,

$$m_{i,k} = R_\theta(u_{i,k}) \in \mathbb{R}^d. \tag{8}$$

This captures deviations such as “red cube” from red plus cube without a dense lookup per combination.

3.4 Adaptive Bound Object Slots

We combine additive and residual branches with a learned vector gate,

$$g_{i,k} = \sigma(G_\theta([a_{i,k}; m_{i,k}])) \,, \tag{9}$$

then define the bound object slot

$$b_{i,k} = \mathrm{LN}(a_{i,k} + g_{i,k} \odot m_{i,k}) \,, \tag{10}$$

where $[\cdot;\cdot]$ is concatenation and $\odot$ is elementwise multiplication. The gate suppresses unnecessary residual dimensions while retaining reusable additive semantics.

3.5 Scene Composition

Object slots become scene components after position conditioning. For each slot or position we learn

$$\pi_p \in \mathbb{R}^d, \tag{11}$$

then form

$$\tilde{b}_{i,k} = b_{i,k} + \pi_{p_{i,k}}. \tag{12}$$

We report position-independent AVG and position-aware AVG+POS components.

The object contribution to the scene is the sum of slot embeddings,

$$h_i^{\mathrm{obj}} = \sum_{k=1}^{K} \tilde{b}_{i,k}. \tag{13}$$

This sum supports slot subtraction/addition; a small context network captures residual scene regularities,

$$h_i^{\mathrm{ctx}} = C_\theta(\mathrm{vec}(s_i)). \tag{14}$$

For cross-object structure, we construct a global product factor

$$q_i = \prod_{k=1}^{K} \prod_{m=1}^{M} \Big(1 + \eta \tanh(P_m^{\mathrm{glob}}(c_{i,k}^{(m)}))\Big) - 1, \tag{15}$$

where η stabilizes the product, together with a global additive term

$$r_i^{\mathrm{add}} = \sum_{k=1}^{K} \sum_{m=1}^{M} A_m^{\mathrm{glob}}(c_{i,k}^{(m)}). \tag{16}$$

Let

$$h_i^{\mathrm{base}} = h_i^{\mathrm{obj}} + h_i^{\mathrm{ctx}}. \tag{17}$$

The global product branch is then

$$h_i^{\mathrm{glob}} = r_i^{\mathrm{add}} + \gamma_i \odot R_\theta^{\mathrm{glob}}(q_i), \tag{18}$$

with gate

$$\gamma_i = \sigma\Big(H_\theta([h_i^{\mathrm{base}}; r_i^{\mathrm{add}}; R_\theta^{\mathrm{glob}}(q_i)])\Big) \,. \tag{19}$$

The predicted scene embedding is

$$\hat{z}_i = \mathrm{LN}\Big(h_i^{\mathrm{base}} + h_i^{\mathrm{glob}}\Big) \,. \tag{20}$$

The object stream supplies editable slots, the context stream absorbs scene residuals, and the global branch adds cross-object capacity. This is a functional inductive bias, not an orthogonality claim.

3.6 Exploratory Multi-Object Extensions

We additionally explore explicit Graph/CrossAttn interaction in harder multi-object settings. Section 4.6 summarizes the dense K=6 reconstruction–efficiency trade-off; full architectures, equations, efficiency measurements, and extended diagnostics appear in Appendix G.1. These variants remain exploratory, and their retrieval gains are not uniform.

3.7 Automatic Structured Inputs for Real Images

For real images, Grounding DINO proposes regions Liu et al. (2024), CLIP crop scoring provides coarse attributes, and box centers define position bins. These pseudo-tuples are a practical relaxation of fully curated symbolic input and populate the same structured interface. The experiment tests whether GAMA can consume automatically extracted factors under perception noise; it does not eliminate the structured-input assumption or constitute end-to-end open-world grounding (Appendix J).

3.8 Training Objective

Training aligns GAMA to the frozen encoder while keeping intermediate factors readable:

$$\mathcal{L} = \mathcal{L}_{\text{geom}} + \mathcal{L}_{\text{read}} + \mathcal{L}_{\text{bind}}, \tag{21}$$

$\mathcal{L}_{\text{geom}}$ aligns coordinate, angular, and batch-discriminative geometry; $\mathcal{L}_{\text{read}}$ preserves object/scene concept readability; and $\mathcal{L}_{\text{bind}}$ encourages consistent bound slots. Exact terms and weights are in Appendix B. A parameter-matched ablation shows metric-specific trade-offs not explained solely by trainable capacity (Appendix Table A.27).

In that three-seed CLEVR-2D ablation, all variants have 2.27M trainable parameters. Removing InfoNCE lowers test MSE from 0.301 to 0.266 but also lowers mean concept-probe accuracy from 0.518 to 0.437; removing scene-concept supervision yields 0.276 MSE and 0.523 concept accuracy. Thus the terms trade reconstruction fit against factor readability rather than each improving every metric.

4 Experiments

4.1 Setup

Reconstruction and decomposition test frozen-space fidelity and factorization; counterfactual replacement tests local editability; appendix subtraction, permutation, and stress tests probe localization and robustness. We evaluate synthetic text scenes, CLEVR, CLEVR-2D, and PUG-SPAR-33x5, a fixed 33-object by 5-texture subset of the public PUG:SPAR release Bordes et al. (2023). Unless stated otherwise, the target is frozen CLIP-B/32 Radford et al. (2021); DINOv2 Oquab et al. (2024) provides visual-encoder context. Compared methods use the same images, split, feature pipeline, and diagnostics, so differences are within-suite rather than external benchmark claims.

4.2 Does GAMA Preserve Frozen-Space Geometry?

Table 1: Main reconstruction, retrieval, and probing. R^2 is AVG / AVG+POS; retrieval and probing use AVG+POS. Delta rows are absolute differences.

Dataset	Model	R^2	Retrieval	Probing
Text	CLIP	0.900 / 0.924	0.970	0.990
Text	GAMA-Text/v13	0.927 / 0.937	0.970	1.000
Text	Delta	+0.027 / +0.013	+0.000	+0.010
CLEVR-2D	CLIP	0.75 / 0.77	0.82	0.99
CLEVR-2D	DINOv2	0.78 / 0.79	0.68	0.97
CLEVR-2D	GAMA	0.94 / 0.86	0.99	1.00
CLEVR-2D	Delta vs CLIP	+0.19 / +0.09	+0.17	+0.01
CLEVR-2D	Delta vs DINOv2	+0.16 / +0.07	+0.31	+0.03
CLEVR	CLIP	0.78 / 0.83	0.94	0.96
CLEVR	DINOv2	0.85 / 0.88	0.86	0.96
CLEVR	GAMA	0.95 / 0.88	1.00	1.00
CLEVR	Delta vs CLIP	+0.17 / +0.05	+0.06	+0.04
CLEVR	Delta vs DINOv2	+0.10 / +0.00	+0.14	+0.04

Table 1 shows that GAMA remains close to frozen geometry on text, CLEVR, and CLEVR-2D while preserving or improving probe accuracy; the matched comparison is CLIP versus GAMA under the same protocol.

Table 2: Object replacement under the position-aware AVG+POS protocol. Frozen visual embeddings are position-sensitive under local editing, so AVG+POS is the standard visual editing setting; position-free AVG variants are treated as diagnostics.

Dataset	Model	Probing	Retrieval
Text	CLIP	0.99	0.99
Text	GAMA-Text/v13	1.00	0.99
Text	Delta	+0.01	+0.00
CLEVR	CLIP	0.98	1.00
CLEVR	DINOv2	0.98	0.95
CLEVR	GAMA	1.00	1.00
CLEVR	Delta vs CLIP	+0.02	+0.00
CLEVR	Delta vs DINOv2	+0.02	+0.05
CLEVR-2D	CLIP	0.98	0.99
CLEVR-2D	DINOv2	0.98	0.97
CLEVR-2D	GAMA	1.00	1.00
CLEVR-2D	Delta vs CLIP	+0.02	+0.01
CLEVR-2D	Delta vs DINOv2	+0.02	+0.03

Table 3: Decomposition scores. Values are AVG / AVG+POS; Delta is GAMA minus CLIP.

Model	Scenes from objects	Objects from concepts	Scenes from concepts
CLIP-B/32 (txt)	0.900 / 0.924	0.914 / 0.909	0.823 / 0.840
GAMA-Text/v13	0.927 / 0.937	0.906 / 0.905	0.840 / 0.848
Delta	+0.027 / +0.013	-0.008 / -0.004	+0.017 / +0.008
CLIP-B/32 (img)	0.75 / 0.84	0.35 / 0.83	0.30 / 0.71
GAMA (visual)	0.959 / 0.968	0.936 / 0.936	0.897 / 0.907
Delta	+0.209 / +0.128	+0.586 / +0.106	+0.597 / +0.197

4.3 Can GAMA Edit One Object Without Disturbing the Scene?

Object replacement changes one slot while preserving the remaining scene. Because frozen visual embeddings are position-sensitive, AVG+POS is the standard visual protocol; position-free AVG remains an appendix diagnostic.

GAMA improves intervention probing and reaches 1.00 retrieval on CLEVR and CLEVR-2D under AVG+POS, supporting position as part of the edited visual object instance.

4.4 Does GAMA Recover More Explicit Object–Concept Structure?

Decomposition asks whether object and concept factors linearly explain the frozen scene space; subtraction/permutation and the matched sum-multiplicative control are reported in the appendix.

Table 3 shows that GAMA makes visual concept-to-scene structure more linearly available and improves reconstruction from object factors. Compared with the matched original complexity baseline, it also improves concept accuracy and lowers MSE across the reported datasets (Appendix Table A.29). Exploratory Graph/CrossAttn results and efficiency trade-offs are in Appendix G.1.

4.5 Continuous Factors and Automatic Pseudo-Tuples

The true-continuous CLEVR-2D variant recovers hue, size, and position with mean $R^2 = 0.965$ at train fraction 0.5 (Appendix Table A.10). On 2000 HF COCO images and 1000 automatic pseudo-counterfactual edits, GAMA obtains 0.810 edited-object and 0.815 preserved-object match; exact full-scene Retrieval@1/5 is 0.413/0.557, whereas tuple-exact Retrieval@1 is 0.619. A gallery-identifiability audit finds that 400 of the 1000 targets share their complete object–color–position tuple with another gallery image. On these collision cases, exact Retrieval@1 is 0.155 but tuple-exact Retrieval@1 is 0.670, showing that exact image-ID recovery additionally tests instance and context information absent from the input tuple. Exact Retrieval@1 is still only 0.585 on the 600 uniquely identified targets, so collisions do not explain all errors (Appendix Table A.18). Among 100 randomly sampled exact-retrieval failures, manual audit assigns 26% to frontend perception, 43% to background/context entanglement, and 31% to residual GAMA binding failure. These failure-conditioned proportions are descriptive rather than a causal decomposition; together with GT-box retraining, they indicate mixed perception, context, and binding limitations rather than a single frontend cause (Appendix Tables A.31 and A.19).

Adapted object-centric comparison. Table 4 brings the fixed-split ObjEmbed diagnostic into the main text. The same 1400/200/400 split and 140 strict test edits are used, with ObjEmbed features extracted from the same detected boxes. The adapter reconstructs frozen targets more accurately than GAMA, while its deployable text-direction edit changes the requested object less often. Target-slot transplantation is an oracle diagnostic, not a deployable comparison.

Table 4: Real-image adapted-baseline diagnostic on one fixed split. Reconstruction and editing are different operations; the target-slot row uses oracle information. Full metrics and protocol details are in Appendix Table A.20.

Method / operation	Recon. MSE	Edit Ret@1	Edited obj.
GAMA local tuple edit	0.163	0.093	0.483
ObjEmbed-Adapter reconstruction	0.106	–	–
ObjEmbed-Adapter text direction	–	0.031	0.124
ObjEmbed target-slot transplant (oracle)	–	0.950	0.950

Frontend and background checks. Training and evaluating on COCO ground-truth boxes improves reconstruction Ret@1 from 0.437 to 0.498 and edited/preserved-object match from $0.810/0.815$ to $0.847/0.878$, but exact edit Ret@1 is almost unchanged (0.413 versus 0.414). This is not a paired localization-only ablation: the retained pools contain 2000 versus 1868 images and the edits are regenerated. Separately, masking background in a paired 200-image diagnostic reduces cosine distance from 0.531 to 0.431 but barely changes Ret@1 (0.158 to 0.167). Neither result supports assigning the real-image retrieval gap solely to the detector or frozen encoder (Appendix J).

4.6 Dense Multi-Object Stress Test: An Efficiency–Interaction Trade-Off

Table 5 summarizes the fixed K=6 real-image pseudo-tuple split (700/100/200 scenes, three seeds). GraphCrossAttn-GAMA reduces test MSE from 0.191 to 0.151, including relative to a nearly parameter-matched wide base, but does not improve Ret@1. Forward-only batch-1 latency rises from 2.700 to 3.338 ms on an RTX 4090; all rows exclude frozen-encoder cost. Explicit interaction is therefore an optional reconstruction–efficiency trade-off, not a universal replacement for the independent-slot base or evidence of better dense-scene editing.

Table 5: Dense real-image K=6 reconstruction versus forward-only latency. Retrieval is exact image-ID recovery; full metrics and measurement settings are in Appendix Tables A.8 and A.30.

Variant	Params (M)	Test MSE ↓	Ret@1 ↑	Batch-1 ms ↓
Base GAMA	2.47	0.191	0.167	2.700
Wide Base GAMA	6.68	0.186	0.163	2.724
Graph-GAMA	5.63	0.164	0.162	3.272
GraphCrossAttn-GAMA	6.68	0.151	0.155	3.338

Capacity control in the dense setting. The wide independent-slot base and GraphCrossAttn-GAMA have nearly equal trainable capacity (6.68M) and use the same split, tuples, and frozen targets. Widening the base changes MSE from 0.191 to 0.186, whereas GraphCrossAttn reaches 0.151; its cosine distance is 0.377 versus the base's 0.438 (Appendix Table A.8). Graph-only and CrossAttn-only variants reach 0.164 and 0.179 MSE. Capacity alone therefore does not explain the reconstruction difference, although the comparison does not isolate a single interaction component or remove noise from automatically extracted tuples.

Reconstruction is not edit retrieval. On the same test partition, base / GraphCrossAttn exact Ret@1 is $0.167/0.155$, despite lower MSE; Ret@5 is $0.400/0.413$ (Appendix Table A.8). Crowd and street MSE also decrease, but these descriptive subsets are not separate benchmarks, and the street subset contains only ten scenes. Full-image targets include instance and background information absent from the structured slots. Lower reconstruction error therefore does not establish improved local editing or exact image-ID retrieval. The latency figures likewise exclude detection and the frozen encoder (Appendix Table A.30); they measure the added cost of the post-hoc module, not end-to-end deployment.

5 DISCUSSION

Controlled experiments show that GAMA exposes editable factors while approximating frozen geometry. Automatic pseudo-tuples preserve local object matching better than exact image-ID retrieval; tuple collisions and residual errors on unique targets show that neither metric alone captures binding. As a post-hoc method, GAMA depends on structured input and on information retained by the frozen encoder. Held-out object and dense-scene analyses delimit its recombination and interaction claims.

The held-out object-name audit makes that boundary concrete: with text-semantic anchors, scene all-objects@K is 0.021 for a held-out GAMA seed, versus 0.039 for the frozen CLIP target; the observed identity-level semantic-distance association is inconclusive (Appendix Table A.23). Recombining learned factors should not be conflated with discovering unseen object identities.

What the objective ablation establishes. The parameter-matched CLEVR-2D ablation holds rank, hidden width, and trainable parameters fixed across objective variants (Appendix Table A.27). Removing the contrastive term lowers test MSE from 0.301 to 0.266 but also reduces mean concept-probe accuracy from 0.518 to 0.437. Removing scene-concept supervision lowers MSE to 0.276 while leaving concept readability nearly unchanged (0.523). Thus the full objective should not be read as optimizing every metric independently: reconstruction fidelity and accessibility of object concepts can move in different directions. The objective is selected for the joint aim of frozen-space reconstruction and controllable factorization, not because every auxiliary term improves MSE in isolation. This distinction also matters when comparing models that reconstruct the same target well but implement edits through different operations.

What real-image retrieval measures. The automatic-input experiment asks a structured query to identify a specific image in a gallery, even though the query records only two detected objects, their coarse colors, and position bins. In the fixed 1000-edit audit, 400 targets have at least one gallery competitor with the same complete ordered tuple; their exact-image Ret@1 is 0.155, compared with tuple-exact Ret@1 of 0.670 (Appendix Table A.18). The gap is an identifiability limit of that query-to-image evaluation, not a corrected estimate of edit success. Conversely, exact Ret@1 is only 0.585 among the 600 unique-tuple targets, so underdetermination cannot account for every failure. The paired mask diagnostic changes frozen-target cosine distance more than retrieval, and the GT-box pipeline improves local matching without materially improving exact edit retrieval. Taken together, these tests separate several plausible bottlenecks but do not isolate one causal source. The manually assigned failure categories are conditional on failed cases, not population-level success rates.

What dense-scene interaction establishes. The three-object synthetic variants change both interaction and the base binder/readout; checkpoint selection also reuses the final strict-unseen subset (Appendix G.1). They diagnose an independent-slot boundary, not a controlled graph-only improvement. Dense K=6 variants instead add interaction to base slots on a fixed split and compare with a nearly parameter-matched wide base. Their reconstruction gains, without retrieval gains, support a measured reconstruction–efficiency trade-off rather than improved dense-scene editing. Forward-only latency excludes detection and the frozen encoder, so it cannot be read as end-to-end latency.

6 CONCLUSION

GAMA is a post-hoc structured intervention interface over a fixed representation space. It reorganizes recoverable structure into editable slots while preserving frozen geometry, but cannot create missing information. Its scope is controlled compositional recombination within known structured concept spaces, not zero-shot discovery of previously unseen object identities, unrestricted open-vocabulary recognition, end-to-end perception, or exact real-image scene generation.

AI USE STATEMENT

An LLM assistant supported experimental design and baseline adaptation; drafted data-preparation, training, and evaluation scripts; analyzed logs and quantitative and qualitative results; and edited prose, LaTeX tables, and figures. The 625 Synthesized-Product625 images were generated with an OpenAI image-generation tool through Codex, not the Gemini system used in prior work. Reported measurements come from executed experiments, not LLM estimates. The authors checked AI-assisted code, analyses, and manuscript against available results and logs and take responsibility for the final content.

REPRODUCIBILITY STATEMENT

Unless noted, targets use CLIP-B/32; only labeled DINOv2 rows use DINOv2-B/14. The main GAMA train fraction is 0.3. Text v13 uses $\lambda_{\text{con}} = 30$ and $\lambda_{\text{scene}} = 1$; the selected PUG-SPAR-33x5 full-coverage model uses seed2244 and $\lambda_{\text{con}} = 15$. PUG-SPAR-33x5 selects 4356 rows from public PUG:SPAR labels and images with custom splits; CLIP, DINOv2, SigLIP, and GAMA use the same images and metadata. Selected filenames and seven source label fields were checked against the public release. Synthesized-Product625 is a separate Codex-tool-generated suite with fixed images, category-pattern metadata, and splits; label-derived prompt templates are not records of the original generation calls.

Diagnostics include reconstruction and decomposition R^2, object-edit retrieval/probing, subtraction/permutation, the original sum-multiplicative baseline, pseudo-tuple edits, frontend localization, and held-out-composition retrieval. Probe-dependent metrics require compatible probes; DINOv2 probing is therefore omitted in Table A.12. Reported multi-seed settings use fixed seeds and include mean and standard deviation.

ETHICS STATEMENT

The experiments evaluate representations on existing or synthesized images, not decisions about people. Real images may contain people or sensitive context; automatic detection and attribute extraction can err. Pseudo-tuple results should not be treated as reliable open-world perception or used for high-stakes decisions.

REFERENCES

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 39–48, 2016. URL https://openaccess.thecvf.com/content_cvpr_2016/html/Andreas_Neural_Module_Networks_CVPR_2016_paper.html.

Rim Assouel, Pietro Astolfi, Florian Bordes, Michal Drozdzal, and Adriana Romero-Soriano. Object-centric binding in contrastive language-image pretraining. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://papers.nips.cc/paper_files/paper/2025/hash/be6047b64b8e38bc8af2d2c6aec9c4ae-Abstract-Conference.html.

Florian Bordes, Shashank Shekhar, Mark Ibrahim, Diane Bouchacourt, Pascal Vincent, and Ari S. Morcos. PUG: Photorealistic and semantically controllable synthetic data for representation learning. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/8d352fd0f07fde4a74f9476603b3773b-Abstract-Datasets_and_Benchmarks.html.

Christopher P. Burgess, Loic Matthey, Nicholas Watters, Rishabh Kabra, Irina Higgins, Matthew Botvinick, and Alexander Lerchner. MONet: Unsupervised scene decomposition and representation. *arXiv preprint arXiv:1901.11390*, 2019. URL https://arxiv.org/abs/1901.11390.

Shenghao Fu, Yukun Su, Fengyun Rao, Jing Lyu, Xiaohua Xie, and Wei-Shi Zheng. ObjEmbed: Towards universal multimodal object embeddings. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026. URL https://arxiv.org/abs/2602.01753.

Klaus Greff, Raphael Lopez Kaufman, Rishabh Kabra, Nick Watters, Christopher Burgess, Daniel Zoran, Loic Matthey, Matthew Botvinick, and Alexander Lerchner. Multi-object representation learning with iterative variational inference. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2424–2433, 2019. URL https://proceedings.mlr.press/v97/greff19a.html.

Bijay Gurung, David T. Hoffmann, and Thomas Brox. Common data properties limit object-attribute binding in CLIP. In *Pattern Recognition: 47th DAGM German Conference, GCPR 2025, Proceedings*, pp. 353–368. Springer Nature Switzerland, 2025. URL https://link.springer.com/book/10.1007/978-3-032-12840-9.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. SugarCrepe: Fixing hackable benchmarks for vision-language compositionality. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/63461de0b4cb760fc498e85b18a7fe81-Abstract-Datasets_and_Benchmarks.html.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2901–2910, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Johnson_CLEVR_A_Diagnostic_CVPR_2017_paper.html.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. Text encoders bottleneck compositionality in contrastive vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 4933–4944. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.301. URL https://aclanthology.org/2023.emnlp-main.301/.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2668–2677, 2018. URL https://proceedings.mlr.press/v80/kim18d.html.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2649–2658, 2018. URL https://proceedings.mlr.press/v80/kim18b.html.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5338–5348, 2020. URL https://proceedings.mlr.press/v119/koh20a.html.

Darina Koishigarina, Arnas Uselis, and Seong Joon Oh. CLIP behaves like a bag-of-words model cross-modally but not uni-modally. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/pdf/6f986c1d24d867e83cdcd459cdb8d8904d36e799.pdf.

Brenden M. Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2873–2882, 2018. URL https://proceedings.mlr.press/v80/lake18a.html.

Martha Lewis, Nihal V. Nayak, Peilin Yu, Jack Merullo, Qinan Yu, Stephen H. Bach, and Ellie Pavlick. Does CLIP bind concepts? probing compositionality in large image models. In *Findings of the Association for Computational Linguistics: EACL 2024*, pp. 1487–1500, 2024. doi: 10.18653/v1/2024.findings-eacl.101. URL https://aclanthology.org/2024.findings-eacl.101/.

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 *Proceedings of the European Conference on Computer Vision*, 2024. URL https://arxiv.org/abs/2303.05499.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. In *Advances in Neural Information Processing Systems*, volume 33, pp. 11525–11538, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/8511df98c02ab60aea1b2356c013bc0f-Abstract.html.

Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. CREPE: Can vision-language foundation models reason compositionally? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10910–10921, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Ma_CREPE_Can_Vision-Language_Foundation_Models_Reason_Compositionally_CVPR_2023_paper.html.

Maxime Oquab, Timothee Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=a68SUt6zFt.

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, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Harman Singh, Pengchuan Zhang, Qifan Wang, Mengjiao Wang, Wenhan Xiong, Jingfei Du, and Yu Chen. Coarse-to-fine contrastive learning in image-text-graph space for improved vision-language compositionality. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 869–893. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.56. URL https://aclanthology.org/2023.emnlp-main.56/.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5238–5248, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Thrush_Winoground_Probing_Vision_and_Language_Models_for_Visio-Linguistic_Compositionality_CVPR_2022_paper.html.

Arnas Uselis, Darina Koishigarina, and Seong Joon Oh. How can embedding models bind concepts? In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026. URL https://arxiv.org/abs/2605.31503.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.01936.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language-image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11975–11986, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Zhai_Sigmoid_Loss_for_Language_Image_Pre-Training_ICCV_2023_paper.html.

Tiancheng Zhao, Tianqi Zhang, Mingwei Zhu, Haozhan Shen, Kyusong Lee, Xiaopeng Lu, and Jianwei Yin. An explainable toolbox for evaluating pre-trained vision-language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 30–37. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-demos.4. URL https://aclanthology.org/2022.emnlp-demos.4/.

A Full Architecture Overview

Figure A.1 expands the main-text architecture in Figure 2. It shows the frozen-target path, the full per-slot binding pipeline, position conditioning, and the object, context, and global scene-composition streams. Full training objectives are defined in Appendix B, while evaluation operators are described in Appendix C.

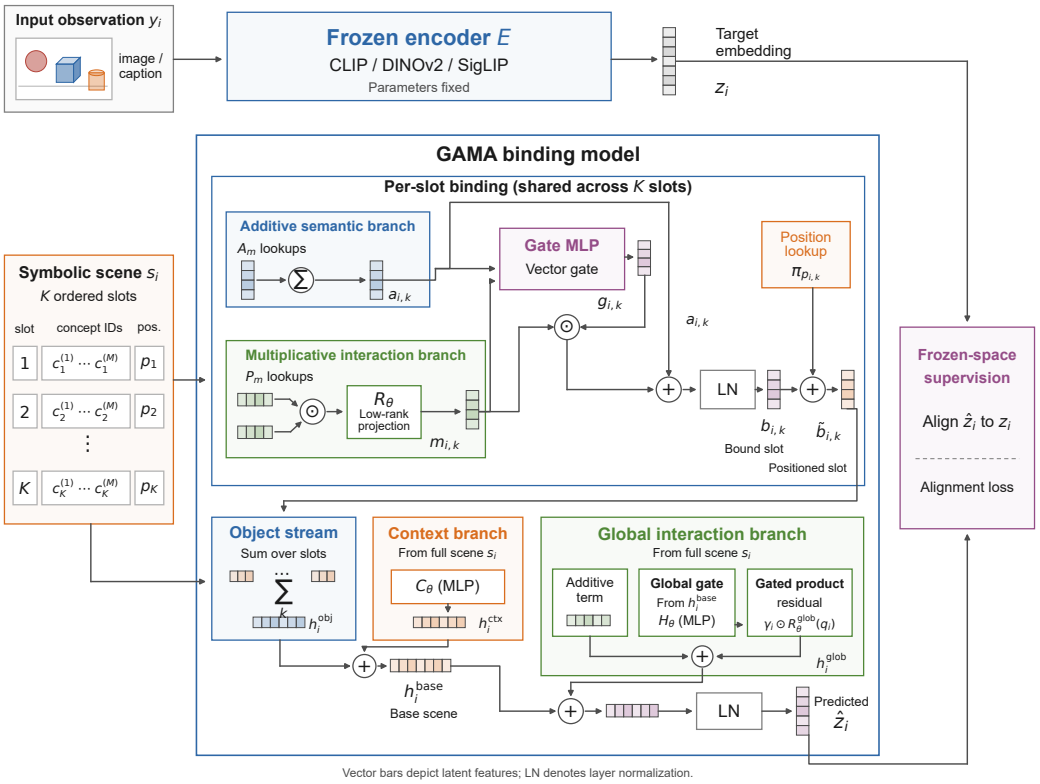

Figure A.1: Expanded GAMA architecture. Structured per-slot binding and scene composition predict a frozen target embedding; objectives and evaluation operators are defined in Appendices B and C.

B Full Training Objective Definitions

The main text gives the combined objective. This appendix records the individual losses used in that objective. Mean-squared reconstruction and cosine distance preserve the frozen target coordinates and angular retrieval geometry: The grouped notation in Eq. 21 expands to the implemented objective

$$\mathcal{L}_{\text{geom}} = \mathcal{L}_{\text{mse}} + \lambda_{\cos}\mathcal{L}_{\cos} + \lambda_{\text{nce}}\mathcal{L}_{\text{nce}}, \quad \mathcal{L}_{\text{read}} = \lambda_{\text{con}}\mathcal{L}_{\text{obj-con}} + \lambda_{\text{scene}}\mathcal{L}_{\text{scene-con}},$$

with $\mathcal{L}_{\text{bind}} = \lambda_{\text{bind}}\mathcal{L}_{\text{bind}}^{\text{raw}}$. A legacy unit-norm term was also evaluated in the training code, but does not impose the factor-table constraint previously ascribed to it; its implementation is specified below.

$$\mathcal{L}_{\text{mse}} = \frac{1}{N}\sum_{i=1}^{N} \|\hat{z}_i - z_i\|_2^2, \tag{22}$$

$$\mathcal{L}_{\cos} = \frac{1}{N}\sum_{i=1}^{N}\left(1 - \frac{\hat{z}_i^\top z_i}{\|\hat{z}_i\|_2\|z_i\|_2}\right). \tag{23}$$

Algorithm 1 Training GAMA on frozen scene embeddings

Require: Symbolic scenes $\{s_i\}_{i=1}^N$, frozen targets $\{z_i\}_{i=1}^N$, train fraction ρ
1: Build object identifiers from concept tuples and split scenes by object coverage.
2: Initialize additive concept tables, factor tables, gates, position embeddings, context branch, and global branch.
3: **for** each mini-batch $\mathcal{B}$ **do**
4: Compute additive object vectors using Eq. 5.
5: Compute multiplicative residuals using Eqs. 7–8.
6: Combine both paths into bound slots with Eqs. 9–10.
7: Compose the scene embedding with Eqs. 12–20.
8: Compute reconstruction, contrastive, concept, and binding losses.
9: Update θ by AdamW on Eq. 21.
10: **end for**
11: Select the checkpoint by validation reconstruction and angular consistency.
12: **return** Predicted embeddings $\hat{z}_i$ and intermediate object slots $b_{i,k}$.

The symmetric mini-batch contrastive term prevents many-to-one collapse:

$$\mathcal{L}_{\text{nce}} = \frac{1}{2}\left[\text{CE}\left(\frac{\bar{Z}\bar{Z}_T^\top}{\tau}, I\right) + \text{CE}\left(\frac{\bar{Z}_T\bar{Z}^\top}{\tau}, I\right)\right], \tag{24}$$

where $\bar{Z}$ and $\bar{Z}_T$ are row-normalized predicted and target embeddings in a batch. Object-level and scene-level concept supervision are

$$\mathcal{L}_{\text{obj-con}} = \frac{1}{KM}\sum_{i,k,m}\text{CE}\Big(g_m(\tilde{b}_{i,k}), c_{i,k}^{(m)}\Big), \tag{25}$$

$$\mathcal{L}_{\text{scene-con}} = \frac{1}{M}\sum_{i,m}\text{BCE}\Big(h_m(\hat{z}_i), y_i^{(m)}\Big), \tag{26}$$

with $y_i^{(m)}$ indicating which values of concept dimension m occur in scene i. For binding consistency, let $\mathcal{P}(v)$ be the set of slot occurrences with the same object-concept tuple v, let $\mathcal{V}_{>1} = \{v : |\mathcal{P}(v)| > 1\}$, and let d be the slot dimension. The implementation uses position-aware slots $\tilde{b}_{i,k}$ and computes

$$\mu_v = \frac{1}{|\mathcal{P}(v)|}\sum_{(i,k)\in\mathcal{P}(v)}\tilde{b}_{i,k}. \tag{27}$$

The raw binding-consistency term averages the mean squared deviation within each repeated-object group, then averages the groups equally:

$$\mathcal{L}_{\text{bind}}^{\text{raw}} = \frac{1}{|\mathcal{V}_{>1}|}\sum_{v\in\mathcal{V}_{>1}}\frac{1}{d|\mathcal{P}(v)|}\sum_{(i,k)\in\mathcal{P}(v)}\|\tilde{b}_{i,k} - \mu_v\|_2^2. \tag{28}$$

The term is set to zero when $\mathcal{V}_{>1}$ is empty. The training scripts also add $\lambda_{\text{unit}}\mathcal{L}_{\text{unit}}$ with $\lambda_{\text{unit}} = 0.001$, where $r_{i,k}$ denotes the multiplicative residual:

$$\mathcal{L}_{\text{unit}} = \frac{1}{NK}\sum_{i,k}\left(\|\text{normalize}(r_{i,k})\|_2 - 1\right)^2. \tag{29}$$

Because normalization precedes the norm test, this term is numerically near zero for nonzero residuals. It is not a regularizer on the factor tables P_m, and no result here establishes a benefit from such a regularizer. Equation 21 describes the effective optimization terms; the reported experiments retain their original training runs.

C Evaluation Operators and Protocol Details

The main split is by object-combination coverage rather than by random rows: training scenes contain a subset of object-concept tuples, and test scenes contain scenes formed from held-out tuples. A closed-set object-ID classifier cannot identify labels absent from probe training; we therefore report held-out object-ID behavior as a boundary diagnostic and use semantic-anchor retrieval only as a test-time calibration probe, not as evidence that the model solves open-vocabulary object recognition. Model selection balances validation reconstruction and angular consistency, while intervention and probe-suite diagnostics determine whether the representation is controllable.

Decomposition. Linear explainability is measured by

$$R^2(Y, \hat{Y}) = 1 - \frac{\sum_i \|Y_i - \hat{Y}_i\|_2^2}{\sum_i \|Y_i - \bar{Y}\|_2^2}. \tag{30}$$

We report scenes-from-objects, objects-from-concepts, and scenes-from-concepts variants.

Object replacement intervention. Let $e(a, o, p)$ be an estimated object component for attribute a, object identity o, and optional position p. A counterfactual edit is

$$\tilde{z} = z - \alpha e(a, o, p) + \alpha e(a', o, p), \tag{31}$$

where a' is the target replacement and α is the steering strength. Retrieval accuracy is

$$\text{Ret@1} = \mathbb{I}\left[\arg\max_j \cos(\tilde{z}, z_j) = j^\star\right], \tag{32}$$

where $j^\star$ is the target counterfactual scene. AVG uses position-independent object averages; AVG+POS keeps position-specific averages.

Probe-suite subtraction and permutation. The probe suite trains diagnostic probes and then evaluates modified embeddings. In subtraction tests, removing concept components should reduce concept accuracy while preserving object information, and removing object components should reduce object accuracy while preserving concept information. In permutation tests, concept or object components are shuffled across scenes. These tests ask whether information is localized in the intended branches or distributed everywhere.

Table A.1: Full probe-suite subtraction and permutation table. Conc. and Obj. denote concept and object probe accuracy.

Intervention	Model	Text Conc.	Text Obj.	Image Conc.	Image Obj.
None	CLIP-B/32	1.00	1.00	0.94	0.96
None	GAMA	1.00	1.00	1.00	0.99
None	Delta	+0.00	+0.00	+0.06	+0.03
Subtract concepts	CLIP-B/32	0.06	0.99	0.05	0.85
Subtract concepts	GAMA	0.17	1.00	0.56	0.97
Subtract concepts	Delta	+0.11	+0.01	+0.51	+0.12
Subtract objects	CLIP-B/32	0.05	0.04	0.02	0.01
Subtract objects	GAMA	0.08	0.23	0.48	0.36
Subtract objects	Delta	+0.03	+0.19	+0.46	+0.35
Permute concepts	CLIP-B/32	0.92	0.99	0.99	0.97
Permute concepts	GAMA	0.96	1.00	0.86	0.97
Permute concepts	Delta	+0.04	+0.01	-0.13	+0.00
Permute objects	CLIP-B/32	0.96	1.00	0.86	0.92
Permute objects	GAMA	1.00	1.00	0.94	0.86
Permute objects	Delta	+0.04	+0.00	+0.08	-0.06

Protocol matching. All numerical claims require the dataset, frozen encoder, split, train fraction, and diagnostic protocol to match. PUG-SPAR-33x5 is a selected subset of public PUG:SPAR images with a custom split, whereas Synthesized-Product625 is an AI-generated supplement; neither result is presented as a score under another paper's original evaluation protocol. Newer object-centric VLM papers are cited as related work unless their inputs and evaluation protocol can be adapted explicitly. For ObjEmbed, we report a fixed-split adapted baseline and mark transplant rows as oracle diagnostics rather than deployable replacement methods.

D Protocol-Matched Baselines and Experiment Scope

Table A.2: Baseline and related-method scope with publication years. Numerical rows are reported under a matched protocol in this manuscript; related rows are recent work cited for timeliness but not used as numerical baselines because their inputs, supervision, or task protocols differ.

Method/family	Year	Reference	Status	Decision in this manuscript
CLIP-B/32	2021	Radford et al. (2021)	Numerical	Main matched frozen-encoder baseline for all reported CLIP-vs-GAMA rows
DINOv2-B/14	2024	Oquab et al. (2024)	Numerical on visual datasets	Included where the protocol supports image-only frozen encoder evaluation
SigLIP-B/16	2023	Zhai et al. (2023)	Numerical supplement	Rerun on CLEVR, CLEVR-2D, and PUG-SPAR-33x5 as a stronger visual-encoder context row
Original binding protocol	2026	Uselis et al. (2026)	Numerical and procedural	Source of reconstruction, intervention, decomposition, probe-suite, and complexity diagnostics
Synthesized-Product625	2026	This work	Numerical supplement	Product-style images generated with a Codex image-generation tool and evaluated under fixed metadata and splits; not a substitute for external product-image assets
Occluded, 3-object, continuous	2026	This work	Numerical supplement	Stress tests for occlusion, three-object scalability, and continuous-factor recovery
Real-image pseudo-tuples	2026	This work	Numerical scope check	Grounding DINO boxes plus CLIP crop scoring provide noisy automatic tuples for a 2000-image / 1000-edit real-image study
CLIP binding analysis	2025	Koishigarina et al. (2026); Gurung et al. (2025)	Related work	Recent binding-failure evidence; not a direct reconstruction/intervention baseline
Object-centric contrastive binding	2025	Assouel et al. (2025)	Related work	Requires contrastive pretraining, so direct numerical comparison is not protocol-matched
ObjEmbed	2026	Fu et al. (2026)	Protocol-adapted numerical baseline	Frozen ObjEmbed box features and adapter rows are evaluated on the same fixed real-image split; oracle transplant rows are upper-bound diagnostics
Structured image-text-graph learning	2023	Singh et al. (2023)	Related work	Uses graph-structured training and different objectives; motivates explicit interaction but is not a matched baseline
Approximator-family controls	2026	Uselis et al. (2026)	Procedural reference	Source of the sum-multiplicative complexity baseline rather than a separate encoder row

Table A.2 separates protocol-matched numerical baselines from related methods whose inputs or training protocols do not match this diagnostic suite. A method is placed in the main result tables only when it uses the same dataset, frozen embedding source, train fraction, and downstream reconstruction/intervention/probe protocol; otherwise it is cited as related-work context.

E Advanced Visual Encoder Baselines

To address encoder-timeliness concerns, we rerun an additional SigLIP baseline on CLEVR, CLEVR-2D, and the PUG-SPAR-33x5 supplement. These rows are not used as replacements for the matched CLIP comparison, but they show whether the observed behavior is specific to one frozen image encoder. Table A.3 reports the same decomposition, intervention, and probe-suite diagnostics where the protocol is compatible.

Table A.3: SigLIP baseline summary. Decomp. is scenes-from-objects R^2 reported as AVG / AVG+POS. Intervention columns report retrieval and probing as AVG / AVG+POS. Base gives probe-suite concept/object accuracy; Perm. gives concepts-permutation concept accuracy and objects-permutation object accuracy.

Dataset	Decomp.	Interv. retrieval	Interv. probing	Base C/O	Perm. C/O
CLEVR	0.750 / 0.786	0.828 / 0.874	0.962 / 0.970	1.000 / 0.870	0.845 / 0.835
CLEVR-2D	0.678 / 0.690	0.879 / 0.883	0.973 / 0.973	1.000 / 0.934	0.864 / 0.847
PUG-SPAR-33x5	0.665 / 0.699	0.969 / 0.969	0.950 / 0.948	0.996 / 0.977	0.934 / 0.992

The SigLIP rows are competitive on several retrieval and probe-suite columns, which confirms that encoder choice matters. However, the main conclusion remains protocol-relative: GAMA is evaluated against the frozen target embedding it is trained to reconstruct, while SigLIP and DINOv2 provide contextual baselines for representation quality.

F SYNTHESIZED-PRODUCT625 SUPPLEMENT

We construct Synthesized-Product625 to test whether the same binding diagnostics extend from CLEVR-style rendered objects to product-style images. Synthesized-Product625 is a synthesized supplement with five product categories and five surface-pattern attributes, forming all ordered two-object combinations. Its 625 images were generated with an OpenAI image-generation tool accessed through Codex, then evaluated under fixed category-pattern metadata, train/validation/test splits, and the same replacement diagnostics used for the controlled visual experiments. This fixed image suite provides a controlled within-suite diagnostic, not a direct external-benchmark comparison or a reproduction of the unavailable Gemini-generated images of prior work. Prompt templates reconstructed from the category labels should not be mistaken for verified records of the original generation calls. Figure A.2 summarizes aggregate retrieval and decomposition effects without using the generated samples themselves as visual evidence.

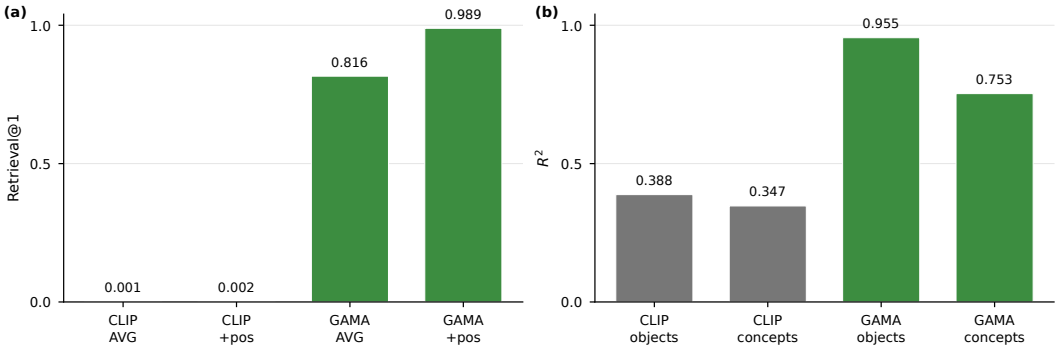

Figure A.2: Results on Synthesized-Product625. (a) Object-edit Retrieval@1 and (b) decomposition R^2 provide supplementary evidence on the synthesized product-style set; higher values are better for both metrics. Evaluation uses the fixed generated images, metadata, and splits.

Table A.4: Synthesized-Product625 diagnostics. Panel (a) reports decomposition R^2; panel (b) reports object-intervention top-1 retrieval over 1000 capped interventions per point; panel (c) reports probe-suite diagnostics.

(a) Decomposition R^2

Model / mode	Scenes from objects	Objects from concepts	Scenes from concepts
CLIP-B/32 AVG	0.360	0.934	0.337
CLIP-B/32 AVG+POS	0.388	0.895	0.347
GAMA AVG	0.921	0.782	0.719
GAMA AVG+POS	0.955	0.788	0.753

(b) Steering-strength sweep

Model	Mode	$k = 0.50$	$k = 0.75$	$k = 1.00$	$k = 1.25$	$k = 1.50$	$k = 2.00$	Best
CLIP-B/32	AVG	0.000	0.000	0.000	0.000	0.000	0.001	0.001 @ 2.00
CLIP-B/32	AVG+POS	0.000	0.000	0.000	0.000	0.000	0.002	0.002 @ 2.00
GAMA	AVG	0.002	0.660	0.805	0.813	0.816	0.752	0.816 @ 1.50
GAMA	AVG+POS	0.007	0.904	0.989	0.981	0.960	0.852	0.989 @ 1.00

(c) Probe-suite diagnostics

Model	Object fit	Concept fit	Base C/O	Perm. concept	Perm. object
CLIP-B/32	0.212	0.205	0.629 / 0.125	0.740	0.451
GAMA	0.907	0.698	1.000 / 0.987	0.941	0.845

Table A.4 shows that Synthesized-Product625 is consistent with the benefits of explicit binding under intervention and probe-suite diagnostics. The most direct intervention result is the steering sweep: CLIP remains near zero across all tested strengths, while GAMA reaches 0.816 without position information and 0.989 with position-aware object vectors. Because the suite consists of generated images evaluated under fixed metadata and splits, we report it as a controlled supplement rather than as a direct external-benchmark comparison.

G Complex Scene and Continuous-Factor Supplements

We add two controlled stress tests for scene complexity and one stress test for continuous attributes. The first uses occluded CLEVR/CLEVR-2D scenes with the same two-object symbolic structure. The second uses a synthetic CLEVR-2D three-object dataset. The third uses real-valued hue, size, and position factors, so the continuous evaluation is no longer limited to binned labels. Figure A.3 summarizes the two most direct model-effect views for these stress tests. The occlusion results show that GAMA remains competitive under corruption; the occlusion-specific sweep improves objects-permutation object accuracy from 0.769 to 0.867 on CLEVR-occ30 and from 0.821 to 0.855 on CLEVR2D-occ30. The independent-slot three-object setting diagnoses why explicit interaction is useful: the best non-graph robustness sweep narrows the probe-suite gaps to -0.008 for concept permutation and -0.029 for object permutation, and the Graph/CrossAttn variants below move the corresponding extension rows to near-matched or positive gaps. The true-continuous variant reaches about 0.965 mean R^2 on real-valued hue, size, and position attributes.

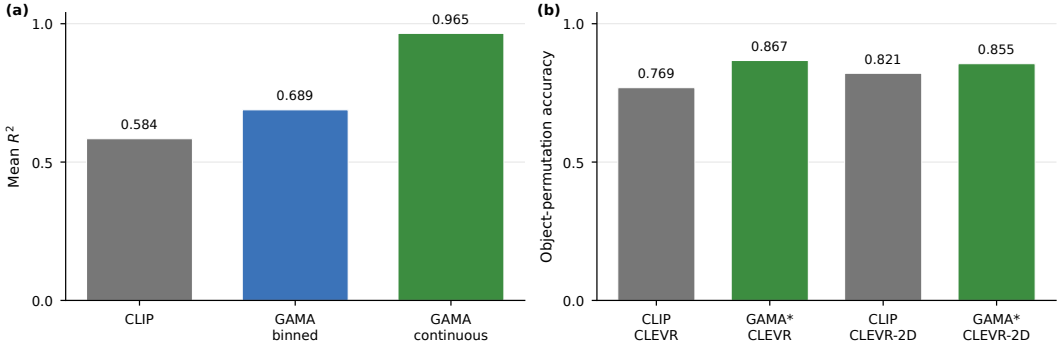

Figure A.3: Continuous and occlusion stress tests. (a) Mean R^2 for continuous-factor recovery compares frozen CLIP, binned GAMA, and true-continuous GAMA. (b) Object-permutation accuracy on CLEVR-occ30 and CLEVR2D-occ30 compares matched CLIP with GAMA from the corresponding occlusion-specific sweep. Higher values are better. The occlusion panel is a stress-specific result rather than a general model claim.

Table A.5: Occluded CLEVR supplements. Decomp. is scenes-from-objects R^2 as AVG / AVG+POS; intervention columns report retrieval and probing as AVG / AVG+POS. Obj-perm. reports object accuracy after object permutation. For GAMA, the obj-perm. value uses the best occlusion-specific sweep; the other columns use the selected reconstruction/intervention model.

Dataset	Model	Decomp.	Interv. retrieval	Interv. probing	Base C/O	Obj-perm.
CLEVR-occ30	CLIP-B/32	0.628 / 0.870	0.561 / 0.763	0.850 / 0.797	0.988 / 0.699	0.769
CLEVR-occ30	GAMA	0.964 / 0.977	0.906 / 1.000	0.991 / 0.992	1.000 / 0.962	0.867
CLEVR-occ30	Delta	+0.336 / +0.107	+0.345 / +0.237	+0.141 / +0.195	+0.012 / +0.263	+0.098
CLEVR2D-occ30	CLIP-B/32	0.703 / 0.759	0.794 / 0.827	0.962 / 0.957	1.000 / 0.948	0.821
CLEVR2D-occ30	GAMA	0.970 / 0.974	0.943 / 1.000	1.000 / 0.999	1.000 / 0.974	0.855
CLEVR2D-occ30	Delta	+0.267 / +0.215	+0.149 / +0.173	+0.038 / +0.042	+0.000 / +0.026	+0.035

G.1 Exploratory Multi-Object Interaction Architectures and Study

The following three-object and dense K=6 experiments probe the independent-slot boundary. They remain exploratory rather than central evidence: the three-object architectures differ from base GAMA in more than the interaction operator, while dense K=6 retrieval does not improve uniformly.

The controlled three-object variants initialize nodes by layer-normalizing concept-token sums plus learned positions; they do not contain the base gated low-rank binder. Dense K=6 variants instead add interaction to the base position-conditioned slots and retain its scene-composition streams. Thus the three-object comparison is not a graph-only ablation of base GAMA.

Let $G = (V, E)$ contain one node per object slot. For dense K=6 Graph-GAMA, the graph input is

$$h_{i,k}^{(0)} = \tilde{b}_{i,k}, \tag{33}$$

using Eq. 12. Dense attention variants first apply slot self-attention ($Q = K = V$), residual addition, and LayerNorm; GraphCrossAttn-GAMA passes these slots to the graph. The graph connects distinct slots,

$$E = \{(j, k) : j \neq k\}, \tag{34}$$

with no learned adjacency, edge features, or relative-position edge term. At layer ℓ, the message from j to k is

$$m_{i,j\to k}^{(\ell)} = \phi_{\text{msg}}^{(\ell)}\left([h_{i,k}^{(\ell)}; h_{i,j}^{(\ell)}]\right), \tag{35}$$

where $\phi_{\text{msg}}^{(\ell)}$ is a LayerNorm–linear–GELU–dropout–linear MLP mapping $2D \to H \to D$. Self messages are masked and, for $K > 1$,

$$\bar{m}_{i,k}^{(\ell)} = \frac{1}{K-1}\sum_{j\neq k} m_{i,j\to k}^{(\ell)}. \tag{36}$$

The residual update is

$$h_{i,k}^{(\ell+1)} = \text{LN}\left(h_{i,k}^{(\ell)} + \phi_{\text{upd}}^{(\ell)}\left([h_{i,k}^{(\ell)}; \bar{m}_{i,k}^{(\ell)}]\right)\right), \tag{37}$$

with the same MLP form for $\phi_{\text{upd}}^{(\ell)}$. Graph blocks use two layers and active attention blocks use eight heads. Dense nodes have $D = d$, whereas three-object nodes have $D = H$. Three-object variants add concept-type embeddings to value tokens. CrossAttn uses a learned slot query over that slot's tokens, residual addition, and LayerNorm; GraphCrossAttn applies this attention before Eqs. 34–37. Attention variants also use a learned scene query over all slots and concatenate it with slot sum/mean pooling. Graph-only three-object readout uses sum/mean pooling. Per-slot concept attention alone does not pass messages between objects.

Table A.6: Synthetic CLEVR-2D three-object supplement. Intervention columns are AVG / AVG+POS. Probe-suite columns report concept/object accuracy for the base, concepts-permutation, and objects-permutation conditions. The robustness-sweep row reports the best probe-suite result only, so intervention columns are left blank rather than mixed across configurations.

Model	Interv. retrieval	Interv. probing	Base C/O	Concepts-perm. C/O	Objects-perm. C/O
CLIP-B/32	0.761 / 0.877	0.290 / 0.758	0.978 / 0.993	0.903 / 0.954	0.904 / 0.921
GAMA selected	0.801 / 0.919	0.267 / 0.649	0.938 / 0.939	0.861 / 0.902	0.859 / 0.858
GAMA robustness sweep	–	–	0.965 / 0.965	0.895 / 0.933	0.893 / 0.892

Table A.7: Graph and cross-attention extensions on the CLEVR-2D three-object stress test. The original selected GAMA row is included as the independent-slot diagnostic; the graph and cross-attention variants add pairwise object message passing and/or object-concept attention. Gaps are relative to the matched CLIP probe-suite row.

Model	Architecture change	Base C/O	Concepts-perm. C/O	Objects-perm. C/O	Gap C	Gap O
CLIP-B/32	frozen visual embedding	1.000 / 1.000	0.994 / 1.000	– / 1.000	0.000	0.000
GAMA original	two-object-style binding	0.950 / 0.930	0.914 / 0.925	– / –	-0.080	-0.075
Graph-GAMA	pairwise object graph	1.000 / 0.999	0.995 / 0.999	– / 0.999	+0.002	-0.001
CrossAttn-GAMA	object queries attend concepts	1.000 / 1.000	0.999 / 1.000	– / 1.000	+0.005	+0.000
GraphCrossAttn-GAMA	graph plus cross-attention	1.000 / 1.000	0.996 / 1.000	– / 1.000	+0.002	-0.000

Table A.7 compares independently trained architectures at the multi-object permutation boundary. The graph and attention variants have near-matched or slightly positive gaps where the base row has negative gaps. Their concept-token encoders and pooled readouts replace the base binder and scene streams; training uses MSE, cosine, NCE, scene-concept, and scene-object multi-hot supervision. The three-object scripts select checkpoints every 50 epochs by reconstruction MSE plus cosine distance on `test_idx`, the strict-unseen subset whose object tuples all belong to the held-out object-tuple vocabulary. Final strict-unseen reconstruction and retrieval use that same subset; checkpoint selection and final evaluation therefore do not form a fully independent validation/test protocol. These rows are exploratory stress diagnostics rather than a controlled graph-only ablation. The composition-level held-out retrieval harmonic mean rises from 0.173 for the base GAMA row to 0.744 for Graph-GAMA and 0.727 for GraphCrossAttn-GAMA, and is reported as a complementary retrieval diagnostic rather than the definition of unseen object-name recognition.

Explicit interaction adds computational overhead relative to Base GAMA and provides a trade-off between efficiency and denser multi-object reconstruction. Forward-only latency measurements are reported in Table A.30; the encoder cost is excluded for all variants.

Table A.8: Dense real-image K=6 pseudo-tuple stress test. All rows use the same automatically extracted Grounding-DINO/CLIP tuples, fixed 700/100/200 train/validation/test split, frozen CLIP-B/32 image targets, and three random seeds. The wide-base row increases base-GAMA capacity to match GraphCrossAttn-GAMA parameters. Lower MSE/cosine distance and higher retrieval are better. Retrieval is not uniformly improved, so the table supports dense-scene reconstruction and geometry fitting rather than a broad dense real-image editing claim.

Model	Params (M)	Test MSE	Cos. dist.	Test Ret@1/5	Crowd MSE	Street MSE
Base GAMA	2.47	0.191	0.438	0.167 / 0.400	0.203	0.222
Wide Base GAMA	6.68	0.186	0.437	0.163 / 0.345	0.197	0.202
Graph-GAMA	5.63	0.164	0.395	0.162 / 0.382	0.176	0.181
CrossAttn-GAMA	3.52	0.179	0.434	0.150 / 0.375	0.191	0.198
GraphCrossAttn-GAMA	6.68	0.151	0.377	0.155 / 0.413	0.158	0.173

Table A.9: Text concept-permutation extension. Cross-attention lets object queries attend text concept tokens and uses scene-token attention pooling. In these reported rows, the previous text concept-permutation deficit moves to a small positive gap relative to the matched CLIP text baseline.

Model	Strict unseen ret.	Mixed unseen ret.	Base C/O	Concepts-perm. C	Gap C/O
CLIP-B/32 text	–	–	0.999 / 0.894	0.992	0.000 / 0.000
GAMA v13 text	–	–	0.933 / 0.855	0.898	-0.094 / -0.060
Text-CrossAttn-GAMA	0.567	0.709	1.000 / 0.939	0.997	+0.005 / +0.013
Text-GraphCrossAttn-GAMA	0.622	0.755	1.000 / 0.933	0.997	+0.004 / +0.001

Table A.10: True continuous-factor supplement on CLEVR-2D. CLIP is the frozen target embedding. The binned row uses discretized factor supervision, while True-continuous GAMA uses real-valued hue, size, x, and y inputs. Regression columns are linear probes back to the underlying continuous attributes. Higher retrieval/R^2 and lower MSE/MAE are better.

Model	MSE to CLIP	Scene Ret@1	Mean continuous R^2	Mean MAE
CLIP-B/32	0.000	1.000	0.584	0.074
GAMA bins, frac. 0.5	0.014	0.275	0.689	0.057
GAMA true-cont., frac. 0.5	0.004	0.754	0.965	0.019
GAMA true-cont., frac. 0.7	0.004	0.816	0.964	0.019

H PERMUTATION ROBUSTNESS AUDIT

Permutation probes test whether slot-local changes remain attached to the intended object. We therefore report seed sweeps, loss-weight sweeps, linear-adapter sweeps, and robustness-specific adapter trials separately from the main results. Table A.11 shows how targeted stress-specific tuning changes the diagnostic gaps.

Table A.11: Permutation robustness audit. Entries are GAMA minus CLIP for concepts-permutation concept accuracy and objects-permutation object accuracy. Positive values indicate that GAMA matches or exceeds CLIP.

Dataset	Concept perm. gap	Object perm. gap	Status
Text	-0.014	+0.001	partial
CLEVR	+0.118	+0.035	pass
CLEVR-2D	+0.072	+0.012	pass
CLEVR-2D-3obj	-0.008	-0.029	stress gap
CLEVR-occ30	+0.039	+0.098	pass
CLEVR2D-occ30	+0.000	+0.035	pass
PUG-SPAR-33x5	+0.002	+0.001	pass

The audit clarifies why the graph and cross-attention variants are useful: independent-slot models can leave permutation gaps in the most demanding settings, while stress-specific tuning and explicit interaction improve the controlled diagnostics. We therefore keep the audit as supporting evidence for the multi-object extension rather than as a standalone main result.

I PUG-SPAR-33X5 SUPPLEMENT

We form PUG-SPAR-33x5 by selecting 4356 labeled image records from the public PUG:SPAR release Bordes et al. (2023). In the archived source labels, all 4356 selected filenames and their seven original label fields match the public release; the 4356-row selection is retained in a labels file with the experiment artifacts. The subset contains 33 character categories and 5 texture values and is evaluated under fixed metadata, custom train/validation/test splits, and the same downstream protocols for all compared models. We report it as a matched within-subset rerun that follows the binding-diagnostic structure of prior work Uselis et al. (2026), not as a score under the original PUG:SPAR evaluation protocol or any PUG:SPARE extension.

Table A.12: PUG-SPAR-33x5 reconstruction, retrieval, and probing. The R^2 score is AVG / AVG+POS. Retrieval is the strict object-based reconstruction retrieval audit; probing reports object probe accuracy when the probe artifact is protocol-compatible. The DINOv2 row is reported as a reconstruction diagnostic, with pair-top1 retrieval shown for AVG+POS and probing omitted.

Dataset	Model	R^2	Retrieval	Probing
PUG-SPAR-33x5	CLIP-B/32	0.591 / 0.636	0.682	0.998
PUG-SPAR-33x5	DINOv2-B/14	0.702 / 0.738	$0.967^{\dagger}$	–
PUG-SPAR-33x5	GAMA seed2244	0.683 / 0.698	0.730	0.987
PUG-SPAR-33x5	Delta vs CLIP	+0.092 / +0.063	+0.048	-0.011
PUG-SPAR-33x5	Delta vs DINOv2	-0.018 / -0.040	–	–

Table A.13: PUG-SPAR-33x5 object replacement intervention. Entries are AVG / AVG+POS; SINGLE-OBJ is omitted because this controlled suite does not include a matched single-object embedding set.

Dataset	Model	Probing	Retrieval
PUG-SPAR-33x5	CLIP-B/32	0.800 / 0.812	0.939 / 0.939
PUG-SPAR-33x5	DINOv2-B/14	0.927 / 0.931	0.934 / 0.934
PUG-SPAR-33x5	GAMA seed2244	1.000 / 1.000	1.000 / 1.000
PUG-SPAR-33x5	Delta vs CLIP	+0.200 / +0.188	+0.061 / +0.061
PUG-SPAR-33x5	Delta vs DINOv2	+0.073 / +0.069	+0.066 / +0.066

Table A.16 clarifies the held-out object-ID probe behavior in PUG-SPAR-33x5. The held-out split asks a closed-set classifier to predict object IDs absent from probe training, so near-zero held-out object accuracy should not be read as absence of object information. When the label space is covered by random-scene splits, GAMA recovers object identity with 0.890–0.913 slot-level accuracy

Table A.14: PUG-SPAR-33x5 decomposition scores. Values are AVG / AVG+POS. Delta rows compare GAMA against CLIP and DINOv2 on the same images, splits, and feature-extraction pipeline.

Model	Scenes from objects R^2	Objects from concepts R^2	Scenes from concepts R^2
CLIP-B/32	0.591 / 0.636	0.667 / 0.767	0.261 / 0.404
DINOv2-B/14	0.702 / 0.738	0.710 / 0.802	0.430 / 0.550
GAMA seed2244	0.683 / 0.698	0.799 / 0.874	0.312 / 0.502
Delta vs CLIP	+0.092 / +0.063	+0.132 / +0.106	+0.051 / +0.099
Delta vs DINOv2	-0.018 / -0.040	+0.089 / +0.071	-0.118 / -0.048

Table A.15: PUG-SPAR-33x5 probe-suite subtraction and permutation. Conc. and Obj. denote concept and object probe accuracy.

Intervention	Model	Conc.	Obj.
None	CLIP-B/32	0.973	0.957
None	DINOv2-B/14	0.989	0.924
None	GAMA seed2244	1.000	0.973
Subtract concepts	CLIP-B/32	0.471	0.845
Subtract concepts	DINOv2-B/14	0.193	0.868
Subtract concepts	GAMA seed2244	0.501	0.992
Subtract objects	CLIP-B/32	0.287	0.023
Subtract objects	DINOv2-B/14	0.144	0.002
Subtract objects	GAMA seed2244	0.390	0.261
Permute concepts	CLIP-B/32	0.921	0.967
Permute concepts	DINOv2-B/14	0.930	0.960
Permute concepts	GAMA seed2244	0.955	0.937
Permute objects	CLIP-B/32	0.961	0.990
Permute objects	DINOv2-B/14	0.971	0.965
Permute objects	GAMA seed2244	0.981	0.994

Table A.16: PUG-SPAR-33x5 random-scene object identity probe. Unlike the held-out object-combination split, this diagnostic uses random scene splits so all object IDs appear in both train and test. Values are means over three probe seeds.

Model	Slot object acc.	Set exact	Set recall	Unseen test objects
CLIP-B/32	0.996	0.992	0.996	0
DINOv2-B/14	0.997	0.996	0.997	0
SigLIP	0.997	0.996	0.997	0
GAMA seed2162	0.905	0.966	0.983	0
GAMA seed2244	0.890	0.969	0.985	0
GAMA seed591	0.913	0.973	0.986	0

and 0.966–0.973 order-invariant set exact match. It does not exceed the frozen encoders on this diagnostic, but it shows that object identity remains recoverable from GAMA embeddings. Tables A.12–A.15 show that PUG-SPAR-33x5 is internally aligned. GAMA improves reconstruction R^2, retrieval, intervention stability, and most probe-suite controls relative to CLIP. The additional DINOv2 rerun strengthens the baseline story: DINOv2 is competitive on decomposition and reconstruction retrieval, while GAMA is best among these reported rows on controlled intervention retrieval and on most probe-suite object controls. The only small GAMA deficits are object probing in the Table A.12 reconstruction audit, permuted-concept object probing in Table A.15, and two decomposition columns versus DINOv2; we keep these visible and interpret PUG-SPAR-33x5 as its own controlled supplement rather than as a direct external-benchmark comparison. †DINOv2 Table A.12 retrieval is pair-top1 retrieval under the reconstruction diagnostic; its concept probe in this audit is near zero and the object probe artifact is not protocol-compatible, so probing is omitted from that row. A separate DINOv2 complexity diagnostic gives concept/object accuracies of 0.589/0.027 at the best learning rate; its embedding-space MSE is not compared directly against CLIP or GAMA because DINOv2 uses a different dimensionality and scale.

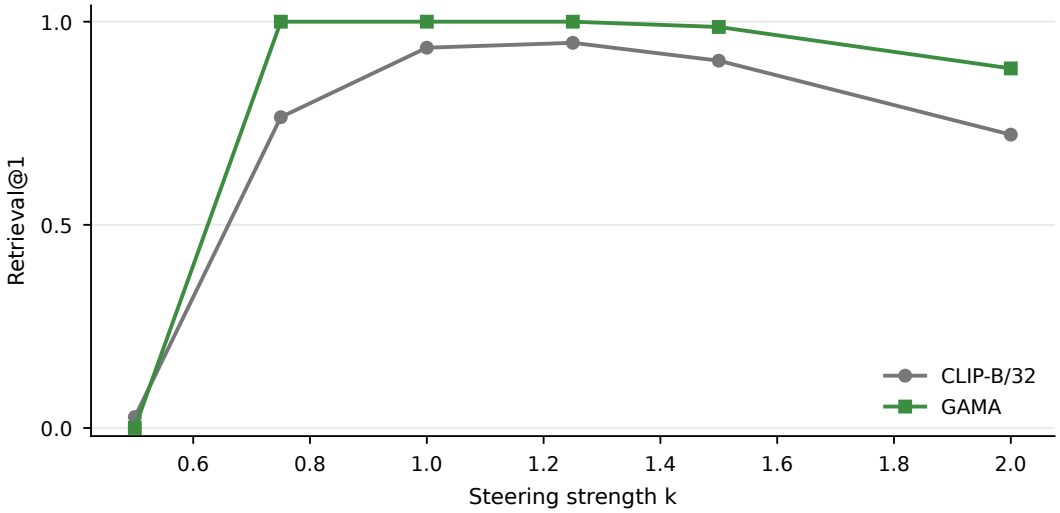

Figure A.4: Stability across the PUG-SPAR-33x5 steering-strength sweep. We vary the intervention strength k and report Retrieval@1 for matched CLIP-B/32 and GAMA; higher values are better. The verified curve shows a broad high-accuracy region for GAMA: retrieval is 1.000 for $k \in \{0.75, 1.0, 1.25\}$ and 0.987 at $k = 1.5$, while the matched CLIP-B/32 curve peaks at 0.948.

Figure A.5 provides two real intervention examples from the same PUG-SPAR-33x5 case-level evaluation. In the first case, the edit replaces the left object from Goldfish to Zebra while preserving BlackRockFish; CLIP retrieves Horse, whereas GAMA retrieves the target Zebra scene. In the second case, the edit replaces Caribou with Tarantula while preserving Crow; CLIP retrieves a scene where the preserved object changes to Chicken, while GAMA keeps Crow fixed.

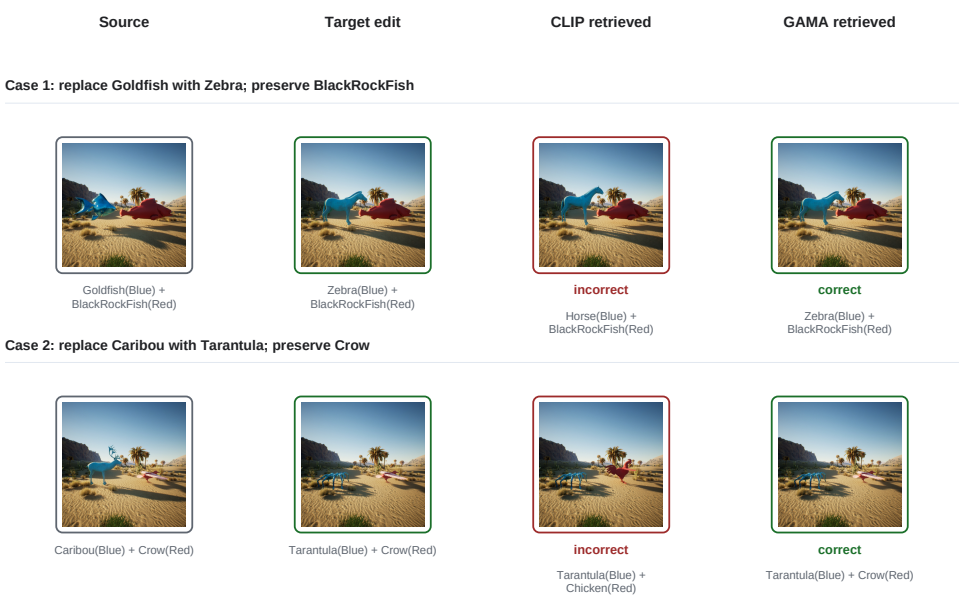

Figure A.5: Qualitative PUG-SPAR-33x5 object interventions. CLIP retrieves nearby but incorrect scenes, while GAMA retrieves the target edit and preserves the non-edited object.

J Real-Image Automatic Tuple Pipeline

To test whether symbolic tuples can be partially relaxed, we run real-image pipelines in which tuples are not provided by dataset annotations. Grounding DINO proposes object boxes Liu et al. (2024); CLIP crop scoring assigns coarse object/color concepts; and box centers are quantized into position bins. The resulting pseudo-tuples are used as GAMA inputs while full images are embedded by frozen CLIP-B/32 targets. These are automatic-input scope checks rather than general-purpose grounding benchmarks.

Table A.17: Larger real-image pseudo-tuple pipeline. Tuples are extracted automatically with Grounding DINO boxes, CLIP crop color scoring, and box-center position bins from 2000 HF COCO images. Tuple-exact requires the retrieved image to match the edited object–color–position tuple, whereas exact retrieval requires the designated image ID. The CLIP edit baseline uses text directions rather than learned object slots, so the row is a scope comparison rather than a protocol-matched object-slot replacement.

Input source	Images	Train/Test scenes	Recon. Ret@1/5	Edit cases	Model	Exact Ret@1/5	Tuple-exact@1	Edited obj.	Preserved obj.	Preserved slot
HF COCO	2000	1080 / 164	–	1000	CLIP text direction	0.000 / 0.010	0.000	0.000	1.000	1.000
HF COCO	2000	1080 / 164	0.437 / 0.536	1000	GAMA local tuple edit	0.413 / 0.557	0.619	0.810	0.815	0.660

The pipeline prepares 2000 images, then the object-coverage split uses 1080 training scenes and 164 held-out test scenes. These values are reported only as evidence that the post-hoc layer can interface with automatic tuple front ends; the controlled experiments remain the main evidence because their tuples, targets, and counterfactuals are cleaner and more tightly protocol matched. Exact image-ID retrieval is intentionally strict: it evaluates instance and full-scene recovery in addition to the object factors supplied to GAMA. We therefore report tuple-exact and local-intervention diagnostics alongside it rather than treating any one metric as a complete real-image score.

The gallery contains 1681 distinct tuples for 2000 images. For the 40% collision subset, the structured query cannot uniquely specify the designated image because instance appearance and background are not input variables. The large separation between exact and tuple-exact retrieval on this subset therefore diagnoses underdetermination in the image-ID criterion. This is not a complete explanation of real-image error: exact Retrieval@1 remains 0.585 for unique tuples, and tuple-exact Retrieval@1 over all cases remains 0.619.

Table A.18: Tuple-identifiability audit for the fixed 1000-edit HF COCO evaluation. A collision target shares its complete ordered object–color–position tuple with at least one other image in the 2000-image gallery.

Target group	Cases	Exact Ret@1	Tuple-exact@1	Edited obj.	Preserved obj.	Preserved slot
All targets	1000	0.413	0.619	0.810	0.815	0.660
Unique tuple	600	0.585	0.585	0.755	0.760	0.630
Tuple collision	400	0.155	0.670	0.893	0.898	0.705

As a separate context diagnostic, we matched 200 images to COCO instance masks and compared three-seed models on the same fixed 140/20/40 split. Replacing background with gray reduced test MSE from 0.2505 to 0.2348 and cosine distance from 0.5308 to 0.4306, but Retrieval@1 changed only from 0.158 to 0.167 and Retrieval@5 decreased from 0.475 to 0.442. Background therefore affects frozen-target geometry, but this intervention does not support attributing the exact-retrieval bottleneck to context alone.

Table A.19: Real-image GT-box frontend analysis with retraining. The second row uses COCO ground-truth boxes and category labels, CLIP crop color scoring, and box-center position bins. GAMA is retrained on the resulting tuples. The automatic run has 2000 images and the GT-box run retains 1868; each constructs its own 1000 edit cases. This is a pipeline comparison, not an inference-only or paired box-only ablation.

Setting	Recon. Ret@1	Edit Ret@1	Edited obj.	Preserved obj.
GroundingDINO tuples	0.4365	0.413	0.810	0.815
GAMA, COCO GT-box tuples (retrained)	0.4979	0.414	0.847	0.878

Table A.19 reports better reconstruction and local object matching for the GT-box pipeline, while exact full-scene edit retrieval remains nearly unchanged. Reconstruction is evaluated over each prepared image pool, including training images; edit cases are regenerated from each tuple pool rather than held fixed between rows. Localization, category assignment, sample coverage, and retraining therefore change together, preventing attribution to localization alone. Attribute extraction and scene-level variation encoded by the frozen full-image target remain possible contributors to the gap, not independently identified causes.

Table A.20: Protocol-adapted ObjEmbed baseline on the fixed real-image pseudo-tuple split. All rows use the same 1400/200/400 train/validation/test split and 140 strict test edit cases. ObjEmbed features are extracted on the same Grounding-DINO boxes. Text-direction rows are deployable embedding-direction baselines; transplant rows use target object slots and are therefore oracle diagnostics rather than directly deployable methods.

Model / operation	Recon. MSE	Recon. Ret@1/5	Exact edit Ret@1/5	Edited obj.	Preserved slot
GAMA local tuple edit	0.163	0.132 / 0.369	0.093 / 0.314	0.483	0.164
CLIP text direction	–	–	0.000 / 0.121	0.000	1.000
ObjEmbed-Adapter reconstruction	0.106	0.628 / 0.890	–	–	–
ObjEmbed-Adapter text direction	–	–	0.031 / 0.179	0.124	0.593
ObjEmbed-Adapter transplant oracle	–	–	0.198 / 0.500	0.329	0.548
ObjEmbed slot transplant oracle	–	–	0.950 / 1.000	0.950	1.000

Table A.20 should not be read as a single ranked comparison. The adapted ObjEmbed row reconstructs frozen CLIP targets more accurately than GAMA on this split, while deployable text-direction edits remain weak for exact local replacement. The strict GAMA exact edit score on this fixed split is low, so the safer interpretation is that different baselines fail in different places: text-direction edits preserve much of the original scene but often do not realize the edited object, whereas GAMA changes the edited object more often but remains sensitive to frontend and full-image entanglement. The oracle transplant rows show that object-centric features can support strong slot replacement when target slots are available, which is a useful upper-bound diagnostic rather than a replacement for the post-hoc tuple-edit protocol.

K PEFT TRADE-OFF

The PEFT supplement measures the trade-off between a strictly frozen encoder and limited encoder adaptation on CLEVR-2D. LoRA adapters are inserted into the CLIP visual encoder while keeping

the vast majority of encoder parameters frozen. The experiment asks whether limited adaptation can expose fine-grained object-concept information that a purely post-hoc layer cannot create.

Table A.21: CLEVR-2D PEFT trade-off. Encoder pct. is the fraction of CLIP encoder parameters updated by LoRA. Rank 0 updates 0% of the encoder. Small updates improve the frozen CLIP object-intervention retrieval/probe diagnostics in this CLEVR-2D setting; larger ranks show the performance-capacity trade-off.

Rank	Encoder pct.	PEFT probe C/O	CLIP obj ret./probe	GAMA obj ret./probe	GAMA pos ret./probe
0	0.0000	–	0.837 / 0.983	0.882 / 1.000	0.999 / 1.000
8	0.0049	0.996 / 0.968	0.888 / 0.996	0.855 / 0.999	1.000 / 0.999
48	0.0292	0.999 / 0.988	0.896 / 0.999	0.862 / 1.000	1.000 / 1.000
256	0.1560	1.000 / 0.997	0.939 / 1.000	0.844 / 0.996	1.000 / 1.000
512	0.3119	1.000 / 1.000	0.902 / 0.997	0.865 / 0.999	1.000 / 0.999

This limited-adaptation trade-off is not part of the base GAMA contribution. The main contribution remains post-hoc controllability over a frozen encoder.

L Composition-Level Held-Out Retrieval Boundary

This appendix section is a boundary analysis for CZSL-style recombination, not a claim that GAMA solves open-vocabulary unseen-object recognition. Held-out object splits are deliberately harsh for closed-set object classifiers because test labels are absent from the probe fitting set. We therefore report composition-level held-out retrieval on PUG-SPAR-33x5 as a complementary retrieval diagnostic, not as a substitute for object-ID analysis. Success is measured by nearest-neighbor recovery of the scene-level object composition, and the main summary reports seen accuracy, unseen accuracy, and their harmonic mean. The adapter in Table A.22 is selected using seen scenes only, does not change GAMA weights, and never uses unseen scenes for fitting.

Table A.22: PUG-SPAR-33x5 composition-level held-out retrieval diagnostic. Hmean is the harmonic mean of seen and unseen top-1 retrieval. The adapted rows use a seen-only ridge calibration and complement the closed-set held-out object probe with a scene-composition retrieval diagnostic.

Model	Base seen	Base unseen	Base hmean	Adapted seen	Adapted unseen	Adapted hmean	Adapted top5 hmean
GAMA held-out seed2	0.219	0.241	0.230	0.842	0.412	0.553	0.766
GAMA fullcover seed591	0.172	0.175	0.173	0.763	0.303	0.434	0.664

M Held-Out Object-Name Boundary Audit

We additionally evaluate the strict held-out object-ID issue directly. The split fits probe or calibration components on 23 seen object values and evaluates on 10 held-out object values, with 2240 held-out-any rows and 400 stricter held-out rows. This diagnostic separates whether editable factors are present from whether unseen object names can be attached to those factors without supervised labels. Closed-set probes cannot learn classifier weights for labels absent from the fitting split, so we compare them with zero-shot object-name semantic anchors derived from CLIP text embeddings and calibrated on seen object-image prototypes only.

Table A.23: Held-out object-ID diagnostic on PUG-SPAR-33x5, evaluated on held-out-any scenes containing at least one unseen object value; these scenes may also contain seen values. Slot top-1/top-5 reports the seen-only closed-set probe over all slots in these scenes. Object recall@K and scene all-objects@K evaluate object-name semantic-anchor retrieval with K equal to the number of object slots. Semantic anchors provide a bounded text-semantic bridge, but exact recovery of all objects in these scenes remains low.

Representation / method	Diagnostic	Held-out slot Top1/Top5	Object recall@K	Scene all-objects@K
CLIP target	Closed-set seen-only probe	0.403 / 0.410	–	–
GAMA held-out seed2	Closed-set seen-only probe	0.333 / 0.395	–	–
CLIP target + calibrated text anchors	Semantic-anchor diagnostic	–	0.350	0.039
GAMA held-out seed0 + calibrated text anchors	Semantic-anchor diagnostic	–	0.303	0.021
GAMA full-coverage seed2244 + calibrated text anchors	Semantic-anchor diagnostic	–	0.334	0.029

Table A.23 gives the intended interpretation of the held-out object-ID boundary. External object-name semantics can recover some labels that were absent from the probe-training split, but multi-object exact recovery remains difficult. GAMA should therefore be described as learning editable binding over observed structured factors, with semantic anchors as an auxiliary text-semantic bridge rather than a full open-vocabulary solution.

We also tested seen-prototype calibration, temperature scaling, and seen-class score biases for semantic-anchor retrieval. Uniform positive temperature scaling alone preserves top-K rankings; the class-specific bias can change them. Under a seen-only model-selection protocol, the selected settings did not close the full-scene held-out object recovery gap. Settings selected using held-out scores are diagnostic upper bounds and are excluded from this conclusion. We therefore treat open-vocabulary object-name transfer as a boundary of the current post-hoc binding layer rather than as a solved capability.

We further tested whether transfer across the ten held-out identities is explained by CLIP text-space proximity to the nearest seen identity. The exploratory Pearson correlation between semantic distance and per-identity recall is 0.731, but its bootstrap 95% interval $[-0.305, 0.957]$ crosses zero and its direction is opposite to the hypothesis that nearer identities transfer better. The rank correlation is also inconclusive (Spearman $\rho = 0.396$, permutation $p = 0.263$). We therefore do not use semantic proximity to explain the boundary: the supported distinction is between recombining observed concept-table entries and discovering an identity for which the base model has no learned entry.

N SEED ROBUSTNESS

Table A.24: PUG-SPAR-33x5 seed robustness for full-coverage seeds. We report the held-out reconstruction loss and probe accuracies for three train/test object splits.

Seed	Train/Test scenes	MSE	Cosine dist.	C0 acc.	C1 acc.	Concept avg.
2162	508 / 1927	0.075	0.171	0.216	0.988	0.602
2244	424 / 2041	0.070	0.158	0.192	0.999	0.596
591	466 / 1984	0.071	0.161	0.209	0.995	0.602
Mean $\pm$ std.	–	0.072 ± 0.003	0.163 ± 0.007	0.206 ± 0.012	0.994 ± 0.006	0.600 ± 0.004

The PUG-SPAR-33x5 seeds differ in held-out object coverage and scene counts, yet the reconstruction MSE and average concept probe remain stable. The low C0 probe reflects the hard held-out-character split rather than an intervention-protocol artifact: the same seed2244 embedding reaches 1.000 intervention probing in Table A.13.

Table A.25: Text seed robustness for the optimized v13 setting at train fraction 0.3. The text runs use $\lambda_{\text{con}} = 30$ and $\lambda_{\text{scene}} = 1$.

Seed	MSE	Cosine dist.	C0 acc.	C1 acc.	Concept avg.
0	0.041	0.153	1.000	1.000	1.000
1	0.040	0.149	1.000	1.000	1.000
2	0.041	0.150	1.000	1.000	1.000

The optimized text setting is stable across seeds: all three runs retain near-ceiling concept readability while keeping MSE near 0.04.

O TRAINING-FRACTION CURVES

The fraction sweep is used as a diagnostic rather than as a replacement for the matched-fraction tables in the main text.

The PUG-SPAR-33x5 curve shows the expected monotonic trend for reconstruction: increasing the fraction from 0.1 to 0.5 lowers MSE from 0.198 to 0.064. The object probe remains at zero because the split evaluates held-out object combinations, so this table is reported as a reconstruction and concept-generalization diagnostic rather than as evidence for closed-set object classification.

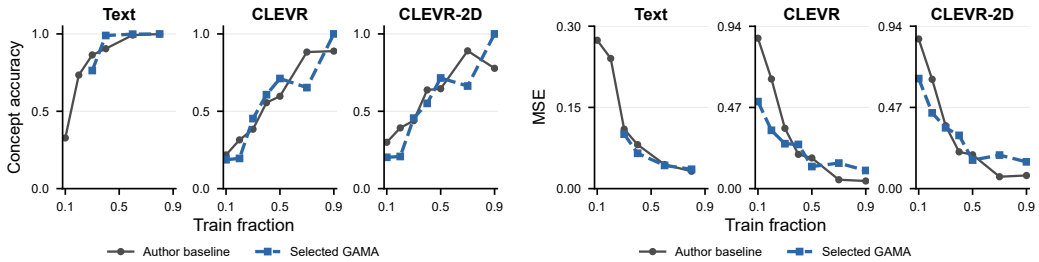

Figure A.6: Training-fraction diagnostics for the author baseline and the selected GAMA model. Left: concept accuracy across training fractions for Text, CLEVR, and CLEVR-2D. Right: MSE across training fractions; lower values indicate better reconstruction of the frozen target embedding. These curves are diagnostic rather than replacements for the protocol-matched main tables.

Table A.26: PUG-SPAR-33x5 training-fraction supplement for seed2244. Higher train fraction improves reconstruction MSE and character-concept readability, while held-out object classification remains difficult.

Train frac.	Train/Test scenes	MSE	Cosine dist.	C0 acc.	C1 acc.	Concept avg.
0.1	70 / 3403	0.198	0.230	0.023	0.699	0.361
0.2	200 / 2675	0.091	0.187	0.064	0.977	0.521
0.3	424 / 2041	0.070	0.158	0.192	0.999	0.596
0.5	1016 / 1181	0.064	0.145	0.319	1.000	0.660

P Parameter-Matched Ablation Summary

Table A.27: Parameter-matched CLEVR-2D objective ablation over three seeds. Every row uses rank 128, hidden width 512, and 2,267,926 trainable parameters. Concept avg. is the mean of the two concept probes. Lower MSE and higher concept readability are better.

Variant	Trainable parameters	Test MSE	Concept avg.
Full GAMA	2,267,926	0.301 ± 0.035	0.518 ± 0.013
No scene-concept supervision	2,267,926	0.276 ± 0.030	0.523 ± 0.006
No InfoNCE / contrastive term	2,267,926	0.266 ± 0.034	0.437 ± 0.052

The parameter-matched ablation controls total trainable capacity so that the effect of removing supervision or alignment terms is not confounded by model size. Removing the contrastive term lowers MSE in this run but also lowers concept readability; removing scene-concept supervision changes little on readability and also lowers MSE. The table therefore supports a metric trade-off rather than a claim that every objective improves every diagnostic.

Q Additional Qualitative Cases

Table A.28: Additional PUG-SPAR-33x5 qualitative intervention cases. Each row is a CLIP-wrong/GAMA-correct retrieval case under AVG object editing.

Case	Source	Target edit	CLIP retrieved	GAMA retrieved
1	Goldfish(Blue)+BlackRockFish(Red)	Zebra(Blue)+BlackRockFish(Red)	Horse(Blue)+BlackRockFish(Red)	Zebra(Blue)+BlackRockFish(Red)
2	Caribou(Blue)+Crow(Red)	Tarantula(Blue)+Crow(Red)	Tarantula(Blue)+Chicken(Red)	Tarantula(Blue)+Crow(Red)
3	Elephant(Blue)+Elephant(Red)	Elephant(Blue)+blank(Red)	Elephant(Blue)+Elephant(Red)	Elephant(Blue)+blank(Red)
4	Elephant(Blue)+Scorpion(Red)	Elephant(Blue)+Zebra(Red)	Elephant(Blue)+Cattle(Red)	Elephant(Blue)+Zebra(Red)
5	Elephant(Blue)+EarlessSeal(Red)	Orca(Blue)+EarlessSeal(Red)	Orca(Blue)+WhiteShark(Red)	Orca(Blue)+EarlessSeal(Red)
6	Camel(Blue)+Elephant(Red)	Camel(Blue)+Orca(Red)	Impala(Blue)+Orca(Red)	Camel(Blue)+Orca(Red)
7	Camel(Blue)+Camel(Red)	Camel(Blue)+Hippopotamus(Red)	Impala(Blue)+Bear(Red)	Camel(Blue)+Hippopotamus(Red)
8	Penguin(Blue)+WhiteShark(Red)	Penguin(Blue)+Chicken(Red)	Penguin(Blue)+Crow(Red)	Penguin(Blue)+Chicken(Red)

These cases broaden Figure A.5: CLIP often retrieves a scene that has the requested new object but changes the preserved object, whereas GAMA more consistently returns the exact edited pair.

Table A.29: Main method versus the original sum-multiplicative complexity baseline. Concept is higher-is-better; MSE is lower-is-better.

Dataset	Base Conc.	Base MSE	GAMA Conc.	GAMA MSE	Delta
Text	0.864	0.110	0.949	0.069	+0.085 / -0.041
CLEVR	0.384	0.349	0.453	0.260	+0.069 / -0.089
CLEVR-2D	0.439	0.364	0.456	0.352	+0.017 / -0.012
PUG-SPAR-33x5	0.404	0.304	0.596	0.070	+0.192 / -0.234

R Supplementary Capacity, Efficiency, and Failure Analyses

R.1 Original Complexity Baseline

These matched rows use the same train fraction within each dataset. Across the reported entries, GAMA improves concept accuracy and lowers reconstruction MSE relative to the original complexity baseline.

R.2 Graph/CrossAttn Efficiency

Table A.30: Forward-only inference latency for the dense K=6 interaction variants. Measurements use identical inputs on an NVIDIA RTX 4090, exclude frozen-encoder cost, and report mean $\pm$ standard deviation over 10 blocks of 200 iterations after 100 warm-up iterations. Batch-64 latency is the latency of the full batch.

Variant	Trainable parameters (M)	Batch 1 latency (ms)	Batch 64 latency (ms)
Base GAMA	2.47	2.700 ± 0.051	3.108 ± 0.040
Graph-GAMA	5.63	3.272 ± 0.018	3.700 ± 0.078
CrossAttn-GAMA	3.52	2.865 ± 0.083	3.263 ± 0.040
GraphCrossAttn-GAMA	6.68	3.338 ± 0.011	3.763 ± 0.050

Explicit interaction adds measurable overhead relative to Base GAMA. CrossAttn-GAMA is the lighter interaction variant in this implementation, whereas Graph-GAMA and GraphCrossAttn-GAMA incur larger latency and parameter costs in exchange for the dense-scene reconstruction behavior reported in Table A.8.

R.3 Real-Image Failure Mode Analysis

We manually audited a simple random sample without replacement of 100 cases from the 587 Exact Retrieval@1 failures in the fixed 1000-edit HF COCO experiment (seed 20260920). Each case received one primary label using the source, target, automatic tuple, and retrieved image.

Table A.31: Manual failure-mode audit for 100 randomly sampled Exact Retrieval@1 failures. Categories are mutually exclusive primary annotations.

Failure mode	Count	Percentage
Frontend perception error	26	26%
Background/context entanglement	43	43%
GAMA binding failure	31	31%

Frontend perception error denotes visibly inconsistent Grounding DINO localization or CLIP crop-level attribute/tuple extraction. Background/context entanglement denotes cases in which the edited semantic content is closer to the target but full-scene retrieval is dominated by background, lighting, occlusion, or broader context. GAMA binding failure denotes a visually plausible pseudo-tuple for which the intervention fails to replace or preserve the intended slot-level factor. These annotations describe the observed error composition; they are not a controlled intervention and therefore do not causally decompose pipeline error. Because the sample is conditioned on Retrieval@1 failure, the 26% frontend category cannot be subtracted from the overall error rate to obtain a detector-corrected binding success rate or a theoretical upper bound.

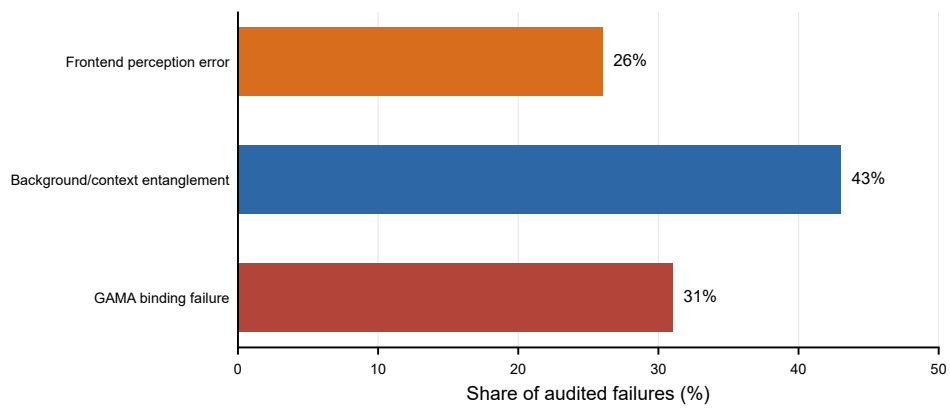

Figure A.7: Distribution of primary failure-mode annotations in the fixed 100-case manual audit. The bars reproduce the locked counts in Table A.31.

S Main-Result Visual Summary

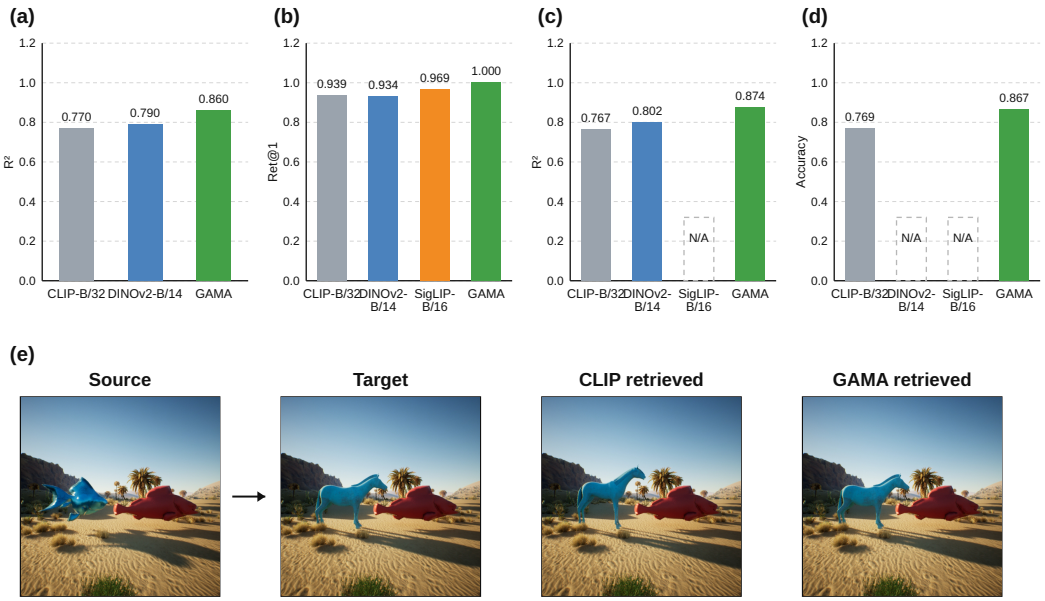

Figure A.8: Visual summary of the main and supplementary evidence. (a) Frozen-space reconstruction R^2 on CLEVR-2D under AVG+POS for protocol-matched models. (b) Object-replacement Retrieval@1 on the PUG-SPAR-33x5 supplement. (c) Objects-from-concepts decomposition R^2, with unavailable protocol-matched values marked N/A. (d) Object permutation under occlusion using the corresponding stress-test sweep. (e) A stored qualitative retrieval example. Quantitative claims are defined by the protocol-matched tables in the main paper and appendix.